

INTERLABORATORY TESTING: A QUALITY ASSURANCE TOOL FOR THE REAL WORLD

CHRISTOPHER J. CZYRYCA

THE PAPER DISCUSSES THE VARIOUS MODES OF INTERLABORATORY TESTING AND THE DIFFERING PURPOSES OF EACH TYPE AND SHOWS THE ROLE OF INTERLABORATORY TESTING WITHIN A SYSTEMATIC LABORATORY QUALITY ASSURANCE PLAN. LABORATORY MANAGERS CAN USE THE RESULTING METHOD OF OVERSIGHT WHEN PARTICIPATING IN COMMERCIALY AVAILABLE INTERLABORATORY TESTING PROGRAMS.

CONSIDERING THE MODERN LABORATORY SYSTEM with calibration standards, check samples, and automated testing, how does interlaboratory testing fit into the overall quality assurance plan? This paper will focus on the scope and purpose of interlaboratory testing and propose a plan for implementation and interpretation of results in four parts:

- The scope and purpose of several modes of interlaboratory testing
- The integration into the quality assurance plan
- Recommendations for use
- The continual interpretation of results.

SCOPE AND PURPOSE OF INTERLABORATORY TESTING

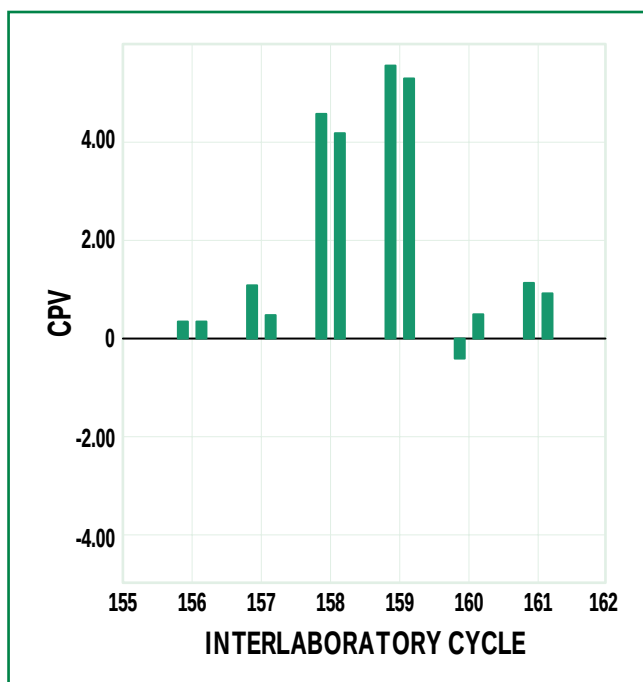
Interlaboratory testing has a broad range of application. The term covers any testing completed in multiple laboratories. Classifying interlaboratory testing by one of four models is often possible: round robins, interlaboratory surveys, proficiency testing, and split or check sample testing. The purpose and sample type help to differentiate these models of testing. The same classifications also appear in a different format in ISO Guide 43 (1).

Traditional round robin testing involves the same sample or specimen passing to all participating laboratories. It usually ends with the first laboratory determining if the sample has undergone excessive changes. This is called a measurement comparison scheme in ISO Guide 43. The purpose of the process is often to ensure sample

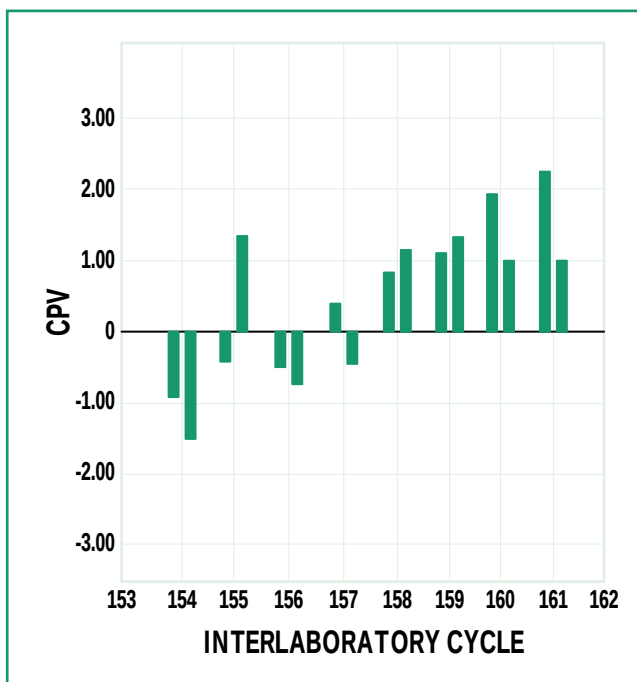
neutrality with highly variable solid materials, trace element examinations, or measurement artifacts such as gauges. Applications of this interlaboratory model would include a small group of participants, unique or highly specialized samples, and materials that the environment or procedure will not affect. The limitations are that labs do not receive timely feedback and sample damage may occur due to procedures, environment, or transportation. Laboratories participating in this type of testing must balance the advantage of all participants testing the same sample against both the time intensive process of each laboratory passing on the sample and the limited field of application.

Interlaboratory surveys measure the performance of a method or procedure in a selection of laboratories. In this type of testing, a central facility distributes homogeneous samples and collects the results. The purpose of this method of interlaboratory testing is the appraisal of a procedure rather than of individual laboratories. In interlaboratory surveys, there is little prior information about expected performance regarding material to method interactions. Despite this level of uncertainty, this type of testing does provide some information on laboratory performance.

Proficiency testing works similarly to an interlaboratory survey. A central facility originates the samples and collects the data. The sample has a "correct" value. This correct value may come from the issuing laboratory or a referee laboratory or laboratories, or it may derive later from participant results. The closeness to this correct value indicates laboratory performance. The limitation in



1. Graph of interlaboratory testing results showing CPV values for Parker Print Roughness



2. Graph of interlaboratory testing results showing CPV values for tearing strength

this test design is the need for many specimens—at least one per laboratory. Since these samples may not be uniform, the test model must compensate for variation in the samples supplied to each laboratory.

The check sample, known value, or split sample is the mode of interlaboratory testing that has the most control over the test specimen. In check sample testing, a sample is assigned a true value from an appropriate national or international level laboratory. This sample goes to only one laboratory. In split sample testing, the specimen is physically split between the assigning laboratory and the laboratory undergoing the test. The obvious limitation for this model of interlaboratory test is that the few labs qualified to assign “true” values can only prepare a small number of samples. Split sample schemes also find use in comparisons between manufacturers and customers.

Each testing model therefore differs by either scope or purpose. Except the interlaboratory survey, the common purpose is to gauge the accuracy of the results of a laboratory.

INTERLABORATORY TESTING WITHIN THE QUALITY ASSURANCE PLAN

The need to validate a laboratory’s competency through external comparison links to the purpose and function of the laboratory. A laboratory in a paper mill can adhere to the same structural model as any other business enterprise. A laboratory uses personnel to employ its resources to produce a product (measurements) for its customers (the mill, customers, and researchers). Interlaboratory testing addresses the quality of the laboratory

measurements and therefore offers a gauge of the performance of the laboratory personnel and resources.

In this way, interlaboratory testing is the final step of the quality assurance plan. The test is an external comparison of the product that provides for this analysis in this time an evaluation of the effectiveness of the quality assurance program.

Because there is no generally accepted format for establishing and maintaining a laboratory quality assurance program, laboratory managers will incorporate interlaboratory testing in different ways. Since this step of competency verification is critical to laboratory performance, quality plans usually include interlaboratory testing. An accepted model of a quality assurance plan is in “Typical Contents of a Quality Manual for Testing Laboratories” (2). Interlaboratory testing figures prominently in the suggested content including the use of reference samples, interlaboratory testing for accreditation, and proficiency testing.

In summary, laboratories must incorporate outside verification of competency into a comprehensive quality assurance plan to evaluate the measurement process as a whole. Other means verify the components of the process. For example, calibration verifies equipment. Confidence in the result can only come from external comparison.

RECOMMENDATIONS FOR IMPLEMENTATION OF INTERLABORATORY TESTING

Garfield lists five recommendations for interlaboratory testing summarized as follows (3):

1. Require participation by all analysts.
2. Use the results to improve quality control, initiate training, and revise procedures.
3. Participate in tests appropriate to the laboratory's interest.
4. Move quickly to determine the cause of unsatisfactory results.
5. Document results and incorporate these into the quality assurance plan.

Cycle	Sample	Laboratory mean	Grand mean	CPV
156	GJ71	2.802	2.764	0.27
	GJ72	2.986	2.935	0.31
157	GJ73	1.215	1.077	1.03
	GJ74	1.150	1.091	0.46
158	GJ75	1.554	1.0687	4.51
	GJ76	1.592	1.0846	4.13
159	GJ77	3.020	4.027	5.50
	GJ78	2.746	3.658	5.24
160	GJ79	1.809	1.0923	-0.38
	GJ80	1.628	1.1090	0.49
161	GJ81	1.496	1.0676	1.14
	GJ82	1.543	1.2443	0.95

1. Interlaboratory results for Parker Print Roughness, μm

These five principles are sound recommendations and deserve scrutiny in more detail. Recommendations 1 and 3 indicate the preferred scope of interlaboratory testing. Laboratory managers should promote interlaboratory testing for all the analysts. Limiting participation only to those deemed most competent or those needing most monitoring will not give an accurate picture of laboratory performance. Analysts are often hesitant concerning such evaluation. The explanation that analyst performance is only one component under evaluation may calm fears.

Laboratory managers must therefore decide whether blind or double blind testing is appropriate for the laboratory. For a blind sample, the laboratory management knows the value or composition of the sample, but the analyst does not. For a double blind sample, the analyst knows neither the value nor that it is a proficiency testing sample. Laboratories also should participate in tests most similar to areas of actual work. This should include sample materials, test procedures, and test conditions. Consideration also must exist for the end-use property that the analysis purports to measure. For example, consider a laboratory wishing to monitor the resistance to a heated food oil of a coated food board intended for microwave containers. Participation in a commercially offered air resistance interlaboratory test would not evaluate performance on the method here. For this example, periodic round robins between manufacturer, converter, and end user may be more appropriate.

Recommendation 2 shows interlaboratory testing being a gauge for the entire measurement process. Deficiencies in any area of the measurement process from specimen preparation to calculation of results may affect performance. Borrowing from the laboratory as business metaphor introduced previously, interlaboratory work—while literally focusing on the quality of the product—will ultimately reveal the quality of the resources of the laboratory. Training, procedure revisions, equipment changes, and environment controls are resources where interlaboratory testing may indicate deficiencies.

Recommendation 4 describes the time-dependent nature of interlaboratory testing. This principle also should include the interpretation of satisfactory results. These are as important as failing results. Most interlaboratory work defines performance of a laboratory at a particular time under a particular set of conditions. It is a “snapshot” of the measurement process. Usually a laboratory will participate in external comparison more infrequently than internal comparisons. Considering the laboratory's results too long after the completion of the analysis may give inaccurate conclusions because knowledge of the testing conditions and procedure are lost.

The final recommendation links interlaboratory testing to a continuing effort of quality assurance. Although the results of interlaboratory testing present a snapshot, evolving trends from sufficiently frequent participation may come from external comparisons viewed over time. Interlaboratory results that do not indicate problems viewed individually may reveal trends when viewed sequentially that indicate deficiencies in training, methodology, specimen handling, etc.

INTERLABORATORY RESULTS FROM A MANAGEMENT PERSPECTIVE

When a laboratory implements a quality assurance plan, the laboratory manager's role most often concerns the interpretation of results. Because of the variety of interlaboratory tests, offering specific models for interpretation of results is hard. Grouping test results into three categories can cover most cases: inconsistent failures, systematic failures, and nonfailures. The following sections describe each situation with general interpretation criteria and resolution actions.

Inconsistent failures are those results deemed unsatisfactory by the employed statistical model and contrary to what the historical performance of the laboratory or performance on other samples would have predicted. Failures of this nature usually indicate a procedural failure or an interaction between the material and the envi-

Cycle	Sample	Laboratory mean	Grand mean	CPV
154	SC67	55.63	59.28	(-0.91)
	SC68	56.54	62.33	(-1.52)
155	SC69	47.17	48.72	(-0.42)
	SC70	51.88	47.16	1.34
156	SC71	54.91	56.47	(-0.51)
	SC72	41.10	42.93	(-0.74)
157	SC73	64.92	63.55	0.38
	SC74	74.15	76.11	(-0.46)
158	SC75	49.86	47.42	0.82
	SC76	58.60	55.07	1.14
159	SC77	80.58	76.70	1.09
	SC78	68.57	62.90	1.31
160	SC79	39.89	36.28	1.94
	SC80	38.14	36.17	1.01
161	SC81	66.30	60.07	2.24
	SC82	64.59	61.31	0.95

II. Results of interlaboratory testing for tearing strength of printing papers in g

ronment or equipment. Examples of such procedural breaks would include failure to calibrate, improper sample preparation, impurities in a reagent, etc. Inconsistent failures are often hard to trace because the errors are not part of the testing procedure. Corrective action for this type of interlaboratory result should include a review of test procedure, a check of calibration, and a retest of the sample, if possible.

Systematic failures are those results that are unsatisfactory and are consistent with results on previous interlaboratory testing or on other samples. Systematic failures are the result of a shift in the average. The cause of this shift may be a modification to the test procedure, improper calibration, or the condition of test equipment. Systematic failures may be easier to define and correct than inconsistent failures because recreating the conditions of the failure is possible. A possible method of corrective action is a sequential review of the test procedure to isolate the source of error. For example, if the equipment calibrates properly and check samples test normally, one should check the specimen preparation and handling (4).

Results that are nonfailures are acceptable within the guidelines of the statistical analysis. The laboratory manager should still interpret acceptable results to gain information on performance. Several types of acceptable results may warrant action. Results that approach but do not exceed critical limits may warrant investigation. Determining whether results warrant caution and investigation the cause may prevent subsequent failures. One should investigate results that are consistently offset from the correct value—high or low. Although the results may be valid and require no action, one should know the cause of the offset. Differences in calibration standards or in instrumentation may gauge these acceptable differences.

Laboratories should also track testing performance over time by maintaining a record of results. There are numerous benefits to this practice ranging from documenting laboratory competence to gaging efficacy of training. The model in **Table I** is an example for the historical examination of quantitative data from an interlaboratory test offered at regular, frequent intervals. These examples use data from the Collaborative Reference Program for Paper and Paperboard conducted by Collaborative Testing Services, Inc. Each historical record lists the laboratory mean, grand mean, and comparative performance value (CPV) or “h statistic” in chronological order and presents the CPV values graphically.

The following equations define the terms in Table I:

$$\text{Grand mean} = \frac{\sum \text{Included laboratory means}}{\text{Number of included laboratories}} \quad (1)$$

$$\text{CPV} = h = \frac{(\text{Laboratory mean} - \text{Grand mean})}{\text{Between laboratory standard deviation}} \quad (2)$$

The results listed and plotted show how historical records can be used to enhance the benefits from interlaboratory testing (5). In **Fig. 1** and Table I, a consistency failure is quite evident. Besides this information, note that the level of measurement of interlaboratory testing returned to the same point as before the break. If the testing had indicated a return to a different level after the corrective action had occurred, a new measurement baseline would have been necessary. In **Table II** and **Fig. 2**, individual results do not indicate any problems except Cycle 155 rejected for inconsistency between the two samples. The drift of the values toward the positive side of the graph is readily apparent. In this instance, the laboratory manager and technician should evaluate the procedure and equipment for the cause of the trend.

This is only an example of a historical record of interlaboratory results. Using many other equally informative formats is possible. To avoid possible confusion, the use of a bar chart is preferable in this model to a line chart to differentiate it from a control chart. One must interpret this historical record of interlaboratory testing differently. It should be easily distinguishable from a control chart.

This model differs from the recommendation in ISO Guide 43. For example, if the data is regularly patterned above or below the midline—x-bar for a control chart or

CPV = 0 for a historical performance chart, this would be cause for action in a control chart, it would be perfectly acceptable for interlaboratory testing. Results arrayed moderately both above and below the baseline are acceptable in a control chart, but would indicate inconsistency for interlaboratory testing. Historical records must be tailored to the type and scope of the interlaboratory testing. At a minimum, enough information is necessary to judge trends and consistency breaks.

SUMMARY

Interlaboratory testing of an appropriate type is an essential component of a quality assurance plan. It provides an outside gauge of results and of the measuring process as a whole. To use this tool, laboratory managers must interpret results both at the time of testing and as part of a continual monitoring program. **TJ**

Czyryca is paper and paperboard program manager at Collaborative Testing Services, Inc., 340 Herndon Pkwy, Herndon, VA 20170.

LITERATURE CITED

1. Anon., *Proficiency Testing by Interlaboratory Comparisons, Guide 43, Version 2*, ISO/IEC, Geneva, 1996, pp. 10-14.
2. Anon., *Typical Contents of a Quality Manual for Testing Laboratories*, ILAC, London, 1984, pp. 10-14.
3. Garfield, F. M., *Quality Assurance Principles for Analytical Laboratories*, 1st edn., AOAC International, Arlington, 1991, p. 93.
4. Mandel, J., *The Statistical Analysis of Experimental Data*, 1st edn., Wiley, New York, 1964, pp. 7-14 and 360-361.
5. Mandel, J. and Nanni, L. F., *Measurement Evaluation*, 1st edn., U.S. Department of Commerce, Washington, 1986, pp.44-45.