ORGANICS LABORATORY PERFORMANCE MANAGEMENT BY INTERLABORATORY COMPARISON

D.E. King and S. Cussion*, Environment Ontario, Laboratory Services Branch, P.O. Box 213, Resources Rd., Rexdale, Ontario, Canada, M9W 5L1

ABSTRACT

Historically, most interlaboratory studies have documented current performance and evaluated the performance against a standard in tabular form. Graphical techniques of evaluation provide incentive for participants to improve their performance. Examples demonstrate this approach.

INTRODUCTION

Interlaboratory studies have three primary objectives: documentation of current performance, evaluation of that performance against a standard, and inducement to improve both individual and group performance. In many studies the latter receives little attention. Data is treated and summarized in classical statistical fashion without identifying patterns within the data which might invalidate the analytical technique used. The data is not used to identify laboratories showing particularly good and comparable performance, so there is then little incentive on the part of borderline participants to re-examine the adequacy of their quality control program. On the other hand, graphical techniques readily reveal patterns which can be the key to resolving the more likely sources of error and bias among laboratories. Graphical approaches tend to be much more convincing to participants because of their visual impact. It is then more difficult for a given participant to justify somewhat inadequate performance, in the face of visible evidence that some of the other laboratories are significantly better. Examples from a number of studies of multi-analyte scans of volatile and extractable organics show how errors related to the accuracy of standards or calibration, can be distinguished from those due to inadequate method recovery or instrumental conditions.

BACKGROUND

Many interlaboratory comparison studies are limited to one or two samples shared among several laboratories. Occasionally several analytes may be examined at the same time, particularly when examining organics by gas chromatography or metals by ICP. But rarely is the data examined on other than a single analyte basis. The evaluation process will usually identify outlier values, and provide a data summary based on the mean or median, and standard deviation or range. Group bias is assessed by comparison of the mean against the expected or reference value.

Since these tasks are computerized and the data distribution is not visually examined, this approach often fails to recognize data patterns which may invalidate the conclusions. At best, the summary provides an estimate of how the participants performed as a group. If enough marginal laboratories are present, then the performance reflects the 'lowest common denominator'. Excellent analysts are not identified, and mediocre analysts escape attention. At worst, the small number of precise and accurate participants may be inundated by the larger number of those with method recovery problems, inaccurate standards, or inadequate control of blanks or contamination.

When the purpose of a study is to determine a concensus value for a potential new reference material, or to evaluate the performance of a new method, this failure to examine the data for evidence of poor performance can only lead to poor conclusions about the reference material or method. If the study is intended to identify the best laboratory for contract purposes, failure to recognize unacceptable patterns can lead to the selection of a less acceptable laboratory.

SETTING A STANDARD FOR PERFORMANCE

Performance criteria should derive from the repeatability of a typical single analyst. This defines the expected range for random deviation from the mean of a series of within batch replicate analyses. Replicate tests over several runs do show increased variability, but this is primarily caused by daily fluctuation in calibration. Inadequate control of calibration is the major source of bias, both within and among laboratories. Based on the *f*-test, with adequate degrees of freedom, a ratio of greater than 1.5 for between versus within run variability (standard deviation) suggests inadequate calibration control.

Errors in analysis may be classified as:

a) acceptable deviation based on method repeatability,
b) inexplicably erratic due to indeterminate causes,
c) biased by a systematic effect due to inadequate correction for method blank or baseline/zero conditions,
d) biased by a systematic effect due to inadequate calibration control, or inaccurate standards,
e) biased by variable or erratic control of method recovery,
f) biased due to matrix or other sample related factors,
g) mistakes.

Youden's two-sample approach for demonstrating the presence of systematic error among analysts is well known and accepted (1). When a larger number of samples is used, the data reported by each analyst can be plotted to show the relationship between the reported values and the expected (or median, or mean) values for all samples. Ideally all values for all analysts should fall along a straight line of slope 1.00 and intercept zero, within a band related to the analytical repeatability. In actual fact the following situations may arise in various combinations. Some analysts show:

a) significantly better fit to the expected line,
b) a difference in slope,
c) a difference in intercept,
d) good fit with one or two 'erratic' points,
e) generally poorer fit to a line.

These plots evaluate performance in terms of both precision and bias. They can be used to identify the more comparable laboratories as a basis for setting performance criteria for evaluating individual performance. This is particularly true when the test samples include unknowns. Inclusion of suspect data must be avoided if a reliable estimate of sample concentration is desired. A pattern of bias or imprecision provides justification for excluding an analyst's data during an iterative criteria setting process.

Thus, as a rule of thumb, a difference of more than 5% in slope, or two standard deviations in intercept, is worth comment as a possible bias. The criteria for precision of fit can be based on a factor of about 1.5 times the median value. The factor allows for tolerable between-analyst, between-run, between-laboratory variability. Median values are preferred because they are relatively insensitive to distribution or outliers.

In a sense we start with the best possible estimate of performance and work outwards until a particular participant's performance does not meet the tolerance factors derived from the data set at hand.

The more common practice of using the data set for each sample to identify outliers is not recommended when the objective is to improve performance. Mediocre performance on the part of some tends to protect those with poor data from detection by the ordinary statistical techniques. This approach should be used only to describe the current average performance of participants. It does tend to keep everyone happy whether they deserve it or not.

EXAMPLES

37 Sewage Treatment Plant Pre-Contract Study

Prior to tendering a contract for analysis of volatile and extractable organics including pesticides, a split sample study was carried out between the Environment Ontario Laboratory Services Branch and each of three commercial labs. If successful, one laboratory would analyse volatiles, one would analyse base/neutral and acid extractables, and one would analyse pesticides and herbicides. Samples were prepared by spiking at three levels into tap water, STP effluent and raw sewage matrices. Samples were submitted to the laboratory in duplicate. Figure 1 shows a typical plot of the reported values versus the expected value. The characteristics observed were:

a) in the range 1 to 50 ug/L (tap water and effluent spiked samples) all reported values fit a straight line within 1 to 2 ug/L, for all of the 64 analytes evaluated,

b) the slopes of all lines (i.e., average % recovery) for all analytes within a scan were generally within a range of ± 5 to 10% of a common value for a particular laboratory, and intercepts were all essentially zero,

c) the overall average % recovery for different laboratories differed very significantly, and ranged from 20% to 140% of the expected values.

d) extremely few erratic points were observed.

The conclusions drawn were that the sample spiking was precise, the data within a laboratory was extremely repeatable, the calibration of individual analytes within a scan was consistent, but calibration between laboratories was extremely biased. The average accuracy of the mixed standard solutions used by the laboratories was obviously different. Although the original hypothesis was that recoveries would vary greatly from sample to sample and analyte to analyte this was not borne out by the findings. We concluded that the major problem was inaccurate standards or poor calibration control (2).

Resin and Fatty Acids (RFA)

A series of studies were carried out among six laboratories as part of collaborative method study. A primary issue centered on the availability of reliable standards. The usual sources of RFA are somewhat impure and often unstable. Therefore the preferred analytical procedure, based on extraction at pH 9 and methylation, required standardization using dehydroabietic acid (DHA), on the assumption that all the analytes of concern would have essentially the same response factor after methylation. Previous work by one of the participants had demonstrated this to be reasonable (3).

Each laboratory provided its own source of DHA for calibration. Four ampouled RFA concentrates were distributed. Two, prepared in methanol, were to be spiked into reagent water and analysed by the total procedure. The other two, prepared in methyl-t-butyl ether (MTBE), were to be analysed by direct methylation and injection.

The data reported in the first study showed a great deal of variability among the analytes and the laboratories (4). Subsequent studies attempted to resolve the source of the problems (5). Hypotheses centered on the source and age of the standards, the solvents used to prepare the mixed ampouled standards used in the study, the proper pH for sample storage and extraction, proper spiking of the unknowns, etc. These later studies did not help significantly in resolving the interpretation of data from the first study.

The data from the first study was then reexamined. By ratioing the observed % recovery for each analyte versus the % recovery reported for the DHA, it was found that:

a) the recovery of DHA differed significantly among analysts (Figure 2),
b) the recovery relative to DHA was quite constant (4), for six of the ten RFA's tested,
c) the least stable RFA's demonstrated the greatest variability relative to DHA (4).

The variation in average relative recovery among analytes, and its deviation from 100% in this study, may reflect the difficulty of obtaining a known high purity stable reference material for each of the RFA's. The consistency of recovery relative to each laboratory's DHA supports the decision to calibrate versus the response of DHA. But comparability of data among laboratories will require the implementation of a 'reference' DHA standard for validating each laboratory's working standards.

Volatiles and Base/Neutral Extractables

A small interlaboratory study was initiated to evaluate the comparability of data from a small number of commercial and private laboratories in Ontario (6). The methods used are similar in principle but probably differ in detail. Several analytes were investigated simultaneously. Figure 3 shows the reported percent recovery for each analyte. The data across the scan, as reported by a specified laboratory, is sorted approximately in order of elution.

It is important to note the pattern of increasing recovery for laboratory 2003, the consistent good recovery for laboratory 2002, and the pattern of decreasing recovery for laboratory 2001. These laboratories have set their own instrumental conditions based on internal method development work to 'optimize' their system. It would seem that these conditions differ sufficiently to require reassessment by each laboratory, or to require tighter specification of conditions within the method.

Figure 4 demonstrates the use of a Youden-type two sample plot to evaluate systematic interanalyte effects for a given laboratory. Ideally the pattern of points is circularly distributed about the expected value of 100% recovery for all analytes on both samples. By joining the points in order of elution one may discern trends or localized patterns of under or over recovery. By comparing such diagrams among analysts it may become possible to set criteria which promote better control of performance within the scan.

CONCLUSIONS

Graphical interpretation of data can reveal patterns which cannot be detected by tabular or statistical summaries. Performance improvement requires a cultural change in attitude. If some one else can achieve a better pattern, one should feel obliged to reexamine the issue. Pattern recognition initiates a fresh perspective and provides insight to relationships which are otherwise ignored.

REFERENCES

(1) Youden, W.J. and Steiner, E.H.; Statistical Manual of the Association of Official Analytical Chemists; Association of Official Analytical Chemists; ISBN 0-935584-15-3; 1975.

(2) Rutter, A.; QA/QC Report on Spiked Effluent and Sewage Samples from the 40 STP Toxic Survey Project; 1988 (Draft).

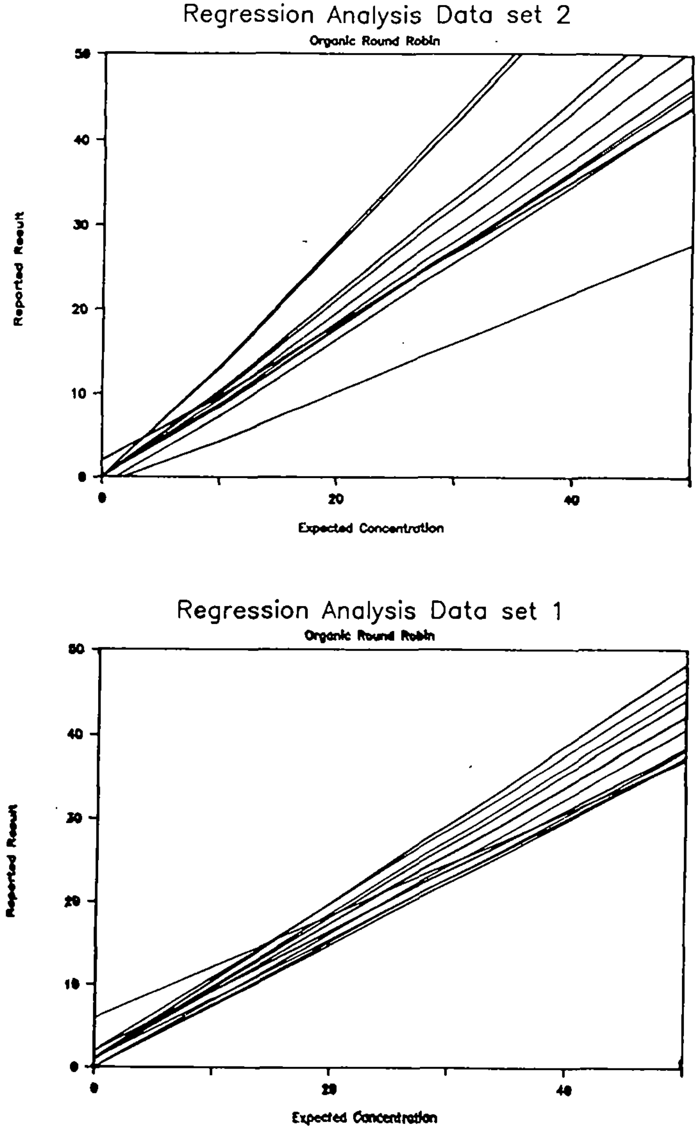
(3) Method for Resin and Fatty Acids; OFIA/MOE/EC Analytical Working Group; 1988.

(4) Interlaboratory Study 88-2A; Validation of a Method for Resin and Fatty Acids; Ampoules for Spiking Reagent Water and Direct Methylation; Environment Ontario, February 1990; ISBN 0-7729-6749-0.

(5) Interlaboratory Study 88-2B; Validation of a Method for Resin and Fatty Acids; Ampouled Standards in Two Different Solvents for Direct Methylation and Instrumental Injection; Environment Ontario, February 1990; ISBN 0-7729-6750-4.

(6) Interlaboratory Study 88-1; Organic Parameters in Reagent Water and Effluents; Environment Ontario, July 1989, Reprinted February 1990; ISBN 0-7729-5756-8.

FIGURE 1



Regression Analysis Data set 2

Regression Analysis Data set 1

FIGURE 2

FIGURE 6: INTERLABORATORY STUDY 88-2A

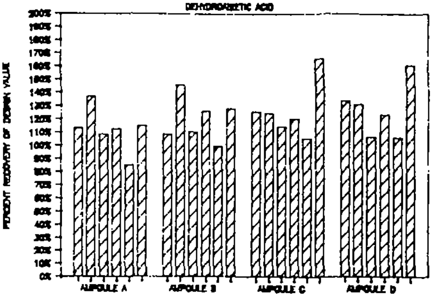


FIGURE 3: INTERLABORATORY STUDY 88-2A

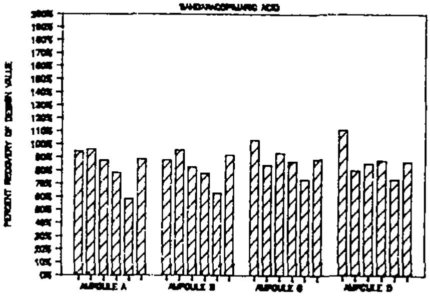


FIGURE 3A: INTERLABORATORY STUDY 88-2A

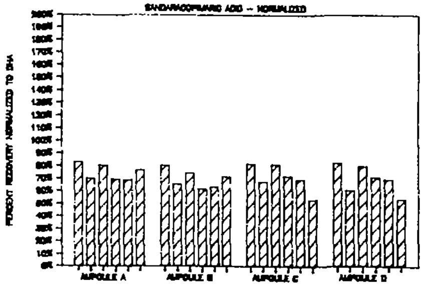


FIGURE 3

INTERLABORATORY STUDY 88-1:VOLATILES

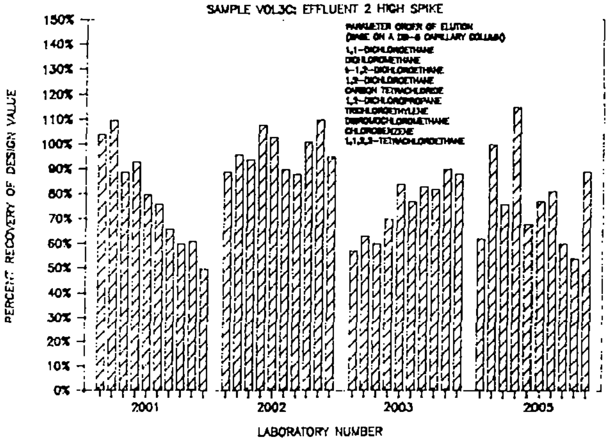
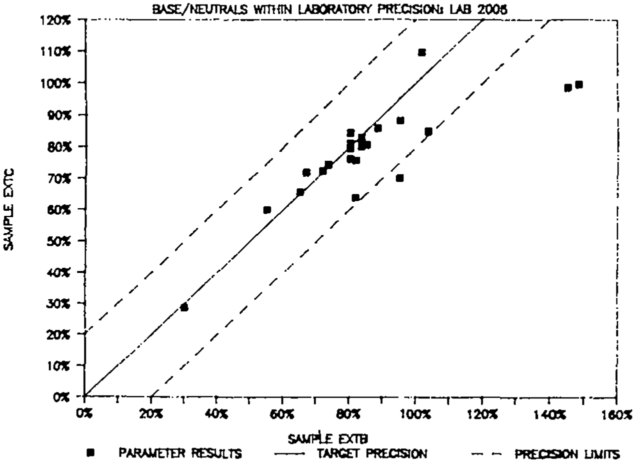


FIGURE 4

INTERLABORATORY STUDY 88-1


BASE/NEUTRALS WITHIN LABORATORY PRECISION: LAB 2005
SAMPLE EXTC
SAMPLE EXTB
■ PARAMETER RESULTS
—— TARGET PRECISION
– – PRECISION LIMITS