aquatic toxicology and hazard assessment

TWELFTH VOLUME

cowgill/williams
EDITORS

ASTM *STP 1027*

STP 1027

Aquatic Toxicology and Hazard Assessment: 12th Volume

U. M. Cowgill and L. R. Williams, editors

ASTM
1916 Race Street
Philadelphia, PA 19103

ASTM: Publication Code Number (PCN): 04-010270-48
ISBN: 0-8031-1253-X
ISSN: 1040-306X

Peer Review Policy

Each paper published in this volume was evaluated by three peer reviewers. The authors addressed all of the reviewers' comments to the satisfaction of both the technical editor(s) and the ASTM Committee on Publications.

The quality of the papers in this publication reflects not only the obvious efforts of the authors and the technical editor(s), but also the work of these peer reviewers. The ASTM Committee on Publications acknowledges with appreciation their dedication and contribution of time and effort on behalf of ASTM.

Printed in Ann Arbor, MI
July 1989

Foreword

This publication, *Aquatic Toxicology and Hazard Assessment: 12th Volume,* contains papers presented at the 12th Symposium on Aquatic Toxicology and Hazard Assessment, which was held 24–26 April 1988 in Sparks, Nevada. The symposium was sponsored by ASTM Committee E-47 on Biological Effects and Environmental Fate and its Subcommittee E47.01 on Aquatic Toxicology. U. M. Cowgill, Dow Chemical Co., presided as symposium chairperson and Llewellyn R. Williams, EPA/EMSL, served as cochairperson. Both served as coeditors of this publication.

Contents

INTERPRETATION OF MICROCOSM DATA

QUALITY ASSURANCE IN ECOTOXICITY TESTING

SEDIMENT TOXICITY TESTING

Overview

During the past twelve years, the aquatic toxicology group (Subcommittee E47.01) of ASTM has sponsored an annual symposium for the major purpose of bringing together aquatic specialists from industry, government, and academe. The end result of these gatherings has been a debate on the merits of test development, animal and plant culture, nutrition and testing, and, last but not least, the lack of interagency harmony. The underlying intent of the 12th Symposium on Aquatic Toxicology and Hazard Assessment was to hold sessions devoted to updating all the various subfields of aquatic toxicology. Thus, the meeting began with a discussion of the benefits of interagency harmonization, which was largely devoted to the need for unity among the various regulatory bodies devoted to protecting the environment. Reviews of common modes of toxic action, target toxicant analysis, and field techniques offered a forum for discussion on advances that have occurred in these fields since the last deliberation on these topics some symposia ago. The problems associated with statistical interpretation of the results of microcosm testing occupied a full session. New approaches in sediment toxicity testing, the culturing and testing of new organisms, exclusively marine, and the never-ending association between nutrition and testing were brought up to date in several minisymposia. A session was devoted to quality assurance in ectotoxicity testing, which is represented in this symposium volume by a discussion of the New Jersey laboratory certification program. The symposium closed with a heavy attendance at a session on toxicity testing problems involving effluents.

These many minisymposia resulted in this volume, which records the events of this meeting in the form of 34 papers. An update of the diverse subfields of aquatic toxicology will serve as a summary useful to all aquatic specialists. It must be realized that the subjects these minisymposia covered are advancing quickly and, therefore, a review or update is advisable on an annual basis. This information is produced at such speed it would be difficult for one individual alone to be proficient in all areas. Thus, annual symposia of the present type serve a very useful purpose to aquatic specialists from industry, government, and academe by appraising them of the most recent advances in the many subfields of aquatic toxicology.

The symposium chairman is indebted to the minisymposia organizers and to their speakers, who made this symposium a success. The ASTM staff is gratefully acknowledged for their assistance in organizing this twelfth symposium and their efforts toward making its publication timely.

U. M. Cowgill

The Dow Chemical Co., Midland, MI 48674,
symposium chairman and coeditor

Benefits of Interagency Harmonization of Methodology

William Horwitz[1]

Harmonization of Methodology: International Perspective

REFERENCE: Horwitz, W., **"Harmonization of Methodology: International Perspective,"** *Aquatic Toxicology and Hazard Assessment: 12th Volume, ASTM STP 1027,* U. M. Cowgill and L. R. Williams, Eds., American Society for Testing and Materials, Philadelphia, 1989, pp. 5–10.

ABSTRACT: Not long ago, most experimentation dealing with analytical methodology in the physical and biological sciences was conducted within a single laboratory. Method validation by other laboratories was not only considered unnecessary but also detrimental because, in the words of one commentator, "the results are too variable." Within the last two decades, however, largely as a result of the requirements of international environmental and food standards programs, it has become increasingly apparent that a collaborative interlaboratory study is the only way to estimate the systematic and random error characteristics of methods of analysis as they will be performed by the population of laboratories typical of those who will be using the method. To obtain a common basis for measuring the statistical characteristics of analytical methods, representatives of almost two dozen international organizations meeting in Geneva in May 1987 approved by consensus a protocol that will be useful for the design and interpretation of collaborative studies of chemical methods of analysis. Much of this protocol will also be useful for the study of biological measurement methods.

KEY WORDS: methods of analysis, accuracy, precision, interlaboratory studies, proficiency studies, collaborative studies, systematic error, random error, performance parameters

Biologists, chemists, toxicologists, and other scientists are constantly making and interpreting measurements without giving a second thought to the mechanics of the process. When investigating a phenomenon, whether it be the wavelength of light, the amount of copper in liver, or the growth of organisms in an aquatic environment, scientists have been taught to isolate the activity, maintain everything else constant, and then vary one factor at a time to see what happens to the response. If many things can be varied with little effect on the response, the system is said to be rugged, and a direct cause and effect relationship can be deduced when a variable does influence the system and the result produced. Sometimes several factors can be varied simultaneously, with mathematical statistics used to sort out the individual effects. This is the way things are done in the laboratory. As soon as we move away from the well-controlled environment of the laboratory into the unrestrained setting provided by nature, direct relationships seem to disintegrate and extraneous factors begin to perturb our observations. Sometimes we can identify the source of the disturbance and control that factor. But when numerous uncontrollable variables exist that have the ability to influence the results, the net effect is high variability in the observations.

About a century ago, when chemical analysis began to be used in the United States for the control of industrial processes, and particularly for the control of agricultural commodities such as fertilizers, animal feeds, and human foods, most of the substantial variability observed between results from different laboratories was ascribed to the methods of analysis. Steps were

[1]Scientific advisor, Food and Drug Administration HFF-7, Washington, DC 20204.

taken by the organizations with responsibility for the products to control this source of variability by specifying the method of analysis to be used. To this end, the Association of Official Analytical Chemists (AOAC) was established in 1884 to approve methods of chemical analysis to control products regulated by state and federal agencies, and the American Society for Testing and Materials (ASTM) was founded in 1898 to formulate standards for commercial commodities. On an international scale, the International Organization for Standardization (ISO) was organized in 1946.

Methods of analysis, the directions for conducting chemical measurements, are the most visible and consequently the most adjustable aspect of biological and chemical experimental systems. When dealing with the macrochemistry of analytes in the 1 to 100% concentration range, the effect of major external variables on the observed system was easily demonstrated because the inherent random and systematic variabilities of the measurements utilized in classical analytical chemistry were relatively small, of the order of 1 to 2% of the analytical value. Therefore, a perturbation from external sources resulting in a difference of several percent in the final value of an analyte was easily visible against a background of 1% or so of analytical noise. The analytes dealt with today, however, are present in the parts per million (ppm, 10^{-6}) and parts per billion (ppb, 10^{-9}) range, where analytical variabilities are of the order of magnitude of 10 to 20% of the analytical value, so the relative influence of variables affecting the system must be considerably greater than this magnitude for their effect to be noticeable above the analytical noise.

Errors

A method of analysis is basically a set of directions on how to conduct a specific biological, chemical, or physical measurement, which ultimately involves the determination of the amount or concentration of an analyte or property. The adequacy of the result is considered in terms of the difference (called systematic error) between the result found and the result known or assumed to be true as well as the degree of dispersion (called random error) of the multiple results among themselves. In discussing experimental measurements, the term "error" is not used in the ordinary sense of a mistake or blunder. Rather it is used in the statistical sense as a numerical measure of the difference between an experimental result and a true value. These experimental variabilities and errors of measurement can be classified into three major categories according to their source: sampling, within laboratory, and between laboratory.

Sampling Errors

Although sampling errors are often the greatest source of variability in environmental measurements, these errors are really beyond the control of the laboratory, unless the laboratory is responsible for the sampling operation. Regardless of who collects the samples, this source of variability must be considered individually for each lot or population to be examined. The magnitude of this source of variability is assessed by taking a number of random samples from the population and examining them individually. With experience, the amount of variability to be expected from similar sites or populations can be determined and used as a first approximation for calculating sample size in terms of number of units and their size. Such a protocol is called a statistical sampling plan and is intended to control the variability encountered when taking additional samples from the same population at different times.

Because each sampling situation is unique, the variability introduced by sampling will not be considered here. Sampling variabilities should always be kept in mind, however, when analytical variability appears to be excessive. Sometimes laboratories have been unfairly accused of producing unreliable results when, actually, the variability lies in the samples submitted to them. As pointed out by Youden [*1*], if the sampling error is greater than twice the analytical

error, reduction in analytical error is of little avail in improving estimates of composition or properties. Emphasis must then be placed on the sampling aspects of the situation, not on the analytical aspects.

Analytical Errors

Analytical errors are inherent in any measurement system. They result from unavoidable differences which arise during the conduct of the measurement operations. There are numerous sources of these deviations, many of which occur without the knowledge of the operator. From the point of view of a single laboratory, it is assumed that the average of a series of measurements of the same analyte or property is the best estimate of that item. The variability of the individual within-laboratory measurements that compose an average provides an indication of the reliability of the estimate of that average. The within-laboratory variability can be broken down into smaller components such as the variability between analysts in the same laboratory, between runs and within runs, between days and within days, between calibration curves and within the same calibration curve, and between and within any perceived source of variability. However, many of these sources of variability overlap and some encompass others (nested) to form a complex error structure that can be very difficult to unravel. Many of these sources of variability are so small that they can be classified as "microprecisions." Sometimes a statistical analysis will indicate that a variable is significant, and at other times, with a different set of data, that it is not significant. Ascribing significance to such unstable factors is the source of many of the controversies involving the interpretation of biological data. All of these microprecisions can be encompassed in the single estimator "within-laboratory" standard deviation.

The situation changes dramatically when more laboratories enter the picture. Each laboratory produces a similar but individual within-laboratory variability and its own long-term average. A new source of variability, *between* laboratories, now emerges which did not exist previously. The *inter*laboratory variability presents a more complex problem than the *intra*laboratory variability because the between-laboratory variability includes the within-laboratory component. In fact, the between-laboratory component of the total variability is usually the largest fraction of the total analytical error. In biological work, this component can be so large that toxicologists, biologists, and microbiologists rarely perform interlaboratory experiments because, as one microbiologist put it, "the results are too variable." If this is the case, whose results are correct—those from my laboratory or from your laboratory? Actually, both results may be "correct" when there is no control of systematic differences in analytical methods between laboratories.

Organizations that provide standard methods of analysis, such as ASTM, AOAC, and ISO, have developed their own procedures, called the collaborative study, to predict how methods of analysis will operate in actual practice. Such a study consists of analyses by a number of laboratories of identical materials over the concentration and commodity range of interest. The results are analyzed statistically to extract the within-laboratory standard deviation, s_r, and the "pure" between-laboratory standard deviation, s_L, which when added together as variances give the overall between-laboratory standard deviation, $s_R = (s_r^2 + s_L^2)^{1/2}$. These parameters, even better expressed as relative standard deviations, RSD_r and RSD_R, are useful as summary statistics of previously performed collaborative studies. With further statistical calculations, the results are then used to predict the performance of methods when used in the future for the same analyte in the same matrix by similar laboratories. The maximum tolerable difference (with 95% confidence) between two individual readings from the same laboratory (r) or from two different laboratories (R) are the repeatability value and the reproducibility value, respectively:

$$\text{Repeatability value} = 2 \times 2^{1/2} \times s_r$$

$$\text{Reproducibility value} = 2 \times 2^{1/2} \times s_R$$

The same statements in terms of relative standard deviations or variances are obtained by substituting the appropriate RSD or s^2 for s.

In addition to the organizations which cover a number of commodity areas, many organizations cover specialized areas, such as water and wastewater, textiles, petroleum products, wines, cereal products, and oils and fats. Many of these organizations have overlapping interests with potential duplication of effort.

The conduct of a collaborative study involving numerous laboratories analyzing a number of identical blind test samples is an expensive undertaking. To avoid duplication of studies among the interested organizations, a working group of the International Union of Pure and Applied Chemistry (IUPAC) was established to harmonize the protocols of the various organizations for performing collaborative studies. The working group sponsored several international symposia over the past decade, culminating in a Workshop on Harmonization of Collaborative Analytical Studies, held in Geneva, Switzerland, at ISO headquarters in May of 1987. Representatives from about 20 international and national organizations interested in the subject agreed by consensus on minimum criteria for the design, conduct, and interpretation of collaborative studies. Adherence to these criteria will permit a collaborative study conducted by one group to serve for all organizations.

Minimum Criteria for Collaborative Studies

The representatives participating in the IUPAC harmonization workshop agreed on a definition of a collaborative study as a study to determine the performance characteristics of a method of analysis as distinct from a proficiency study to determine the performance of laboratories or analysts in conducting an analysis, or from a certification study to obtain a "true" or reference value for an analyte in, or property of, a material. Many individuals and organizations had used these terms, as well as such undefined terms as *round-robin, ring test,* and *intercalibration scheme,* interchangeably for any interlaboratory experiment involving more than one laboratory. A collaborative study, as agreed by the workshop, requires the analysis by at least eight laboratories of at least five materials utilizing either blind (that is, unknown to the analyst) replicates or a split level design. A split level consists of two materials of slightly different analyte concentration but sufficiently close together that they can be considered as having the same variance. When it is impossible to meet the above conditions, the study may be smaller, for example, five laboratories or three materials, but at the sacrifice of confidence in the reliability of the estimated parameters.

An important recommendation, dealing with aspects preliminary to a collaborative study, is to conduct interlaboratory trials only on methods that have received thorough testing within a single laboratory. A method that does not perform satisfactorily within a laboratory will perform less well when subjected to the abuse of several laboratories. Single-laboratory testing of a method includes analyte recoveries under various conditions, interference studies, applicability to expected variants of the analyte in anticipated matrices, comparison with previously available methods, calibration procedures, sources of standards, and developing a clear description of the actual procedure. In addition, specifications for equipment, reagents, and adsorbants must be established.

The report also gives the following advice with respect to significant figures: round standard deviations (and relative standard deviations) to two significant figures and round the mean to accommodate that parameter, that is, to the same number of places before or after the decimal point.

Until recently, few realized that permitting statisticians to analyze data any way they wished introduced as much disagreement in the interpretation of results as permitting chemists to conduct analyses any way they wished. A particular point requiring agreement was that collaborative assays must be analyzed by one-way analysis of variance, material by material. This is very important since the design of collaborative studies makes them appear to be excellent candi-

dates for a two-way analysis of variance procedure. Without the one-way analysis of variance specification there might be disagreements among statisticians as to where to place an interaction term, if this term turns out to be significant.

Another essential point of agreement is how to handle outliers. It is a shocking testimony to laboratory performance that fully 25% of collaborative assays published by the AOAC during the past 75 years contain suspect data, and approximately 3% of the reported analytical data are gross outliers. Outliers are values that are far outside the bulk of the data. The IUPAC harmonization workshop standardized on a specific procedure for the evaluation and elimination of suspect data. First the results must be screened to remove invalid data. Invalid data are results from laboratories that did not follow instructions, reported malfunctioning instruments, did not achieve the expected separations, or observed phenomena not covered by the instructions. Invalid data must be distinguished from outliers found by statistical analysis. Invalid data can be traced to a specific cause; outliers are obtained when the cause cannot be determined and a computation shows that it is very improbable that the value belongs with the bulk of the data.

The data are first tested by the Cochran extreme variance test that discards results from laboratories that cannot check themselves as compared with the bulk of the laboratories. Then data are removed from laboratories that show extreme values by the single value Grubbs test. Since an extreme value may not trigger the Grubbs test when a second extreme value is present (a phenomenon known as "masking"), a "paired Grubbs test" is applied to check for the presence of two extreme values on the same side of the distribution and for one extreme value at each end of the distribution. All flagged outliers are removed if they exceed the critical value at the 1% probability (P) level. This is an extremely conservative outlier removal protocol and is expected to make a mistake (remove a value which is really part of a normal distribution) only 1 time in 100.

A program on the Food and Drug Administration computer is available to handle the calculations and to flag outliers automatically. However, for our use we do not permit the program to actually remove the outliers. Removing outliers requires a deliberate decision on the part of the examiner. We look at the statistical parameters before outlier removal. If these parameters are "acceptable" with no values removed, we do not remove values flagged at even a 1% probability value. This inaction is based on our observation that precision parameters are not meaningfully different unless they differ by a factor of approximately two for the typical collaborative study of eight laboratories. We call this protocol the "historical outlier removal treatment." On the other hand, if we wish the statistical parameters as developed by the harmonized procedure just outlined, we instruct the computer to calculate the precision parameters according to the "IUPAC-1987 Harmonized Procedure," which does remove all values flagged at the $P = 0.01$ level.

The impetus for the development of the harmonized protocol was the fact that many of the organizations participating in the workshop were also interested in supplying their approved methods of analysis to the Codex Alimentarius Commission (CAC) of the Food and Agricultural Organization of the United Nations. In fact, so many methods of analysis were submitted that it was almost impossible to make a choice among them. Consequently, a set of principles was developed which stated, among other things, that methods with reliability parameters that had been established through collaborative studies would be preferred over those without such a pedigree. Since considerable prestige was attached to having a method approved by the CAC, organizations began to perform collaborative studies. Each organization soon realized that selecting its own method and performing its own collaborative study was very inefficient, and that a single collaborative study, when there was general agreement on the design and interpretation, should be sufficient. Therefore, organizations began to look for ways to cooperate with each other. Organizationally, this led to the development of the working party of representatives of a number of organizations working under the auspices of the highly respected umbrella chemical organization—IUPAC. To date, only AOAC has approved the harmonized protocol.

The delay in other organizations is probably related to the time of annual meetings and the mechanics of approval within individual organizations.

Quality Control

The high level of outliers that appear in collaborative studies is very disturbing. Collaborative studies are conducted under such conditions that analysts know that their work will be scrutinized by other analysts as well as by their supervisors. Therefore they can be expected to perform their best work in this situation. We can only guess what the outlier rate would be under less stringent conditions, but certainly it would be no better.

The outlier problem has probably developed because in the course of a single generation of scientists, analytical chemistry has changed precipitously from a science of macroanalysis to one of trace analysis. In macrochemistry, gross errors were relatively easy to detect by the principle of consistency. A misplaced decimal point in concentration usually meant that the physical, chemical, biological, or sensory properties of a substance were changed sufficiently to arouse suspicions as to the accuracy of the analysis. A misplaced decimal point at the parts per million level can be discovered only by a completely independent analysis. Duplicate analyses in this situation are of little help because simultaneous operations usually result in repetition of the same mistakes. This is understandable when the cause of a mistake is an incorrect standard solution, a fluctuating line voltage, or an impurity in a reagent. Arithmetic errors are also likely to be perpetuated. Analysts must take very seriously the necessity for incorporating positive controls into their work to provide some assurance that no gross blunder is being made. Such controls can be certified reference materials, when available, but more frequently they must be specially formulated materials or simply a "house" standard that is analyzed repeatedly with each batch of "unknowns."

Methods of analysis must be written with quality control in mind to provide analysts with guideposts that indicate that the analysis is proceeding satisfactorily. Such a signal is easily incorporated into spectrophotometric methods by specifying an absorbance that should be obtained from a working standard solution. In chromatographic methods, an internal standard providing a signal of approximately the same intensity as a typical analyte concentration usually should be preferred over an external standard. In inductively coupled plasma and other types of emission spectroscopy, the ratios of intensities of peaks of various elements [2] are sufficiently constant to be useful as quality control indicators. Developers of methods must be educated to the necessity of inserting such guideposts into methods. Analytical chemists become so familiar with their technique that they think others are just as proficient. But even the most proficient of analysts occasionally produces outliers, and any assistance in avoiding potential damage before the report leaves the laboratory should be welcome.

Conclusions

International organizations are considering the use of a common protocol for the determination of the operating characteristics of chemical methods of analysis. These characteristics establish the expected limitations of each method, but the frequent occurrence of outliers even during the most careful analytical work suggests that more attention should be given to the incorporation of quality control guidelines into the operating directions. These guides are needed to indicate that the operations are proceeding satisfactorily or to warn the operator that an unobserved deviation in the system has occurred.

References

[1] Youden, W. J., *Journal of the Association of Official Analytical Chemists,* Vol. 50, 1967, pp. 1007–1013.

[2] Botto, R. L., *Spectrochimica Acta,* Vol. 39B, 1984, pp. 95–113.