

NBSIR 77-1240

MEASUREMENT ASSURANCE

J. M. Cameron

Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234

April 1977

Final

Prepared for
Office of Measurement Services
Institute for Basic Standards
National Bureau of Standards
Washington, D.C. 20234



U.S. DEPARTMENT OF COMMERCE, Juanita M. Kreps, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Acting Director*

Measurement Assurance

Introduction

A single measurement can be the basis for actions taken to maintain our health, safety or the quality of our environment. It is important therefore that the errors of measurement be small enough so that the actions taken are only negligibly affected by these errors. We realize this necessity on a personal basis when we consider medical measurements, or our exposure to radioactivity. In any government regulatory action or measurement involved in legal actions it is also obvious that the shadow of doubt surrounding the measurements should be suitably small. But this is no less true for all other measurements in science and industry and even though legal action may not be involved, the validity of scientific inference, the effectiveness of process control, or the quality of production may depend on adequate measurements [2].

Allowable Limits of Measurement Error

How does one achieve this condition--that the measurements are "good enough" for their intended use? It would seem obvious that one has to start with the need--i.e., deciding upon what is "good enough". There are a number of cases where physiological restraints provide the definition such as in the allowable error in exposure to cobalt radiation in cancer treatment or in the amount of pollutant entering a lake. In nuclear materials control the allowable error is a function of the amount of material which would pose a hazard if diverted. In industrial production or commercial transactions, the error limit is determined by a balance between the cost of better measurement and the possible economic loss from poorer measurement.

By whatever path such requirements are arrived at, let us begin with the assumption that the allowable error should not be outside the interval $(-a, +b)$ relative to the quantity being measured. Our problem is one of deciding whether the uncertainty of a single measurement is wholly contained in an interval of that size. We therefore need a means of assigning an uncertainty to a single isolated measurement and, in fact, we need a perspective (i.e., physical and mathematical model) in which to view measurement so as to give operational meaning to the term "uncertainty."

Reference Base to Which Measurements Must Be Related

It is instructive to contemplate the possible "cross-examination" of a measurement if it were to become an important element in a legal controversy. Two essential features emerge. First, that the contending parties would have to agree on what (actually realizable) measurement would be mutually acceptable. The logic of this seems unassailable--if one cannot state what measurement system would be

accepted as "correct," then one would have no defensible way of developing specifications or regulations involving such measurements. Second, the scientific cross-examination by which one establishes the "shadow of doubt" relative to this acceptable value gives one the uncertainty to be attached to the measurement.

The consensus or generally accepted value can be given a particularly simple meaning in dealing with measurements of such quantities as mass, volt, resistance, temperature, etc. One may require that uncertainties be expressed relative to the standards as maintained by a local laboratory or, when appropriate, to the national standards as maintained by NBS. In other cases, nationally accepted artifacts, standard reference materials or in some cases a particular measurement process may constitute a reference base. One basic quality should not be overlooked--all are operationally realizable. The confusion engendered by introducing the term "true value" as the correct but unknowable value is thus avoided.

Properties of Measurement Processes

In discussing uncertainty, we must account for two characteristics of measurement processes. First, repeated measurements of the same quantity by the same measurement process will disagree and, second, the limiting means of measurements by two different processes will disagree. These observations lead to a perspective from which to view measurement namely that the measurement be regarded as the "output" of a process analogous to an industrial production process. In defining the process, one must state the conditions under which a "repetition" of the measurement would be made, analogous to defining the conditions of manufacture in an industrial process.

The need for this specification of the process becomes clear if one envisions the "cross-examination" process. One would begin with such questions as

Within what limits would an additional measurement by the same instrument agree when measuring some stable quantity?

Would the agreement be poorer if the time interval between repetitions were increased?

What if different instruments from the same manufacturer were used?

If two or more types (or manufacturers) were used, how much disagreement would be expected?

To these can be added questions related to the conduct of the measurement.

What effect does geometry (orientation, etc.) have on the measurement?

What about environmental conditions--temperature, moisture, etc.?

Is the result dependent on the procedure used?

Do different operators show persistent differences in values?

Are there instrumental biases or differences due to reference standards or calibrations?

The questions serve to define the measurement process--the process whose "output" we seek to characterize.

The current understanding of a scientific or industrial process or of a measurement process is embodied in a physical model which explains the interactions of various factors, corrections for environmental or other effects, and the probability models necessary to account for the fact that repetitions of the same event give rise to nonidentical answers. For example, in noise level measurement one is involved with assumptions regarding frequency response, weighing networks, influence of procedures and geometry, and an accepted theory for making corrections for temperature and other environmental factors. In mass the properties of the comparator (balance) the environmental effects, and the procedure used all enter into the description of the method.

One thus begins with the specification of a measurement method--the detailed description of apparatus, procedures and conditions by which one will measure some quantity. Once the apparatus is assembled and checked out, one has a measurement process whose output can be studied to see if it conforms to the requirement for which it was created.

In industrial production one tries to produce identical items but usually a measurement process is set up to measure a variety of quantities and ordinarily one does not measure the same quantity over and over. One thus has the problem of sampling the output of the measuring process so as to be able to make statements about the health of the process relative to the needs. The needed redundancy can sometimes be achieved by remeasuring some of the items, or by measuring a reference artifact periodically. It is essential that the repetitions be done under the same diversity of conditions as the regular measurements, and that the items being measured be typical of the regular workload.

As an example, a sequence of measurements was made using two sound level meters to measure a sound of nominally 90 dB re 20 μ Pa. The sound was generated by a loudspeaker fed broadband noise. On 16

different days measurements were made outdoors and over grass with the loudspeaker in the same orientation and location relative to a building 2 m behind the loudspeaker. The sound level meter was always the same distance (10 m) from the loudspeaker and on a line perpendicular to the face of the loudspeaker. Other than the grass, the person holding the sound level meter, and the building to the rear of the loudspeaker, there were no other reflecting surfaces or obstacles within 50 m. No measurements were made in the rain or in winds exceeding a few km/hr. The results from these 16 repetitions are shown in Figure 1. Typically, had duplicate measurements been made on the same day they would have given results as shown in Figure 2.

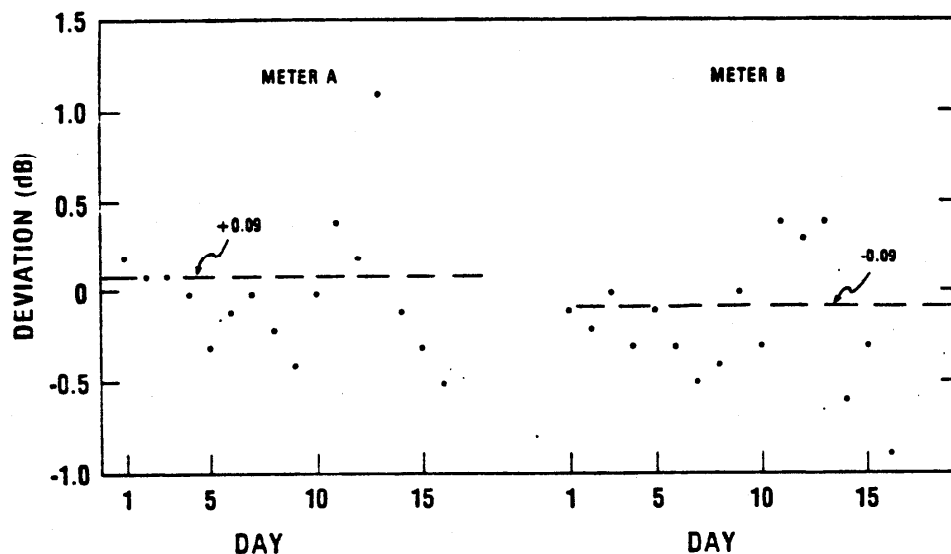


FIGURE 1: DAY-TO-DAY VARIATION IN METER READINGS.

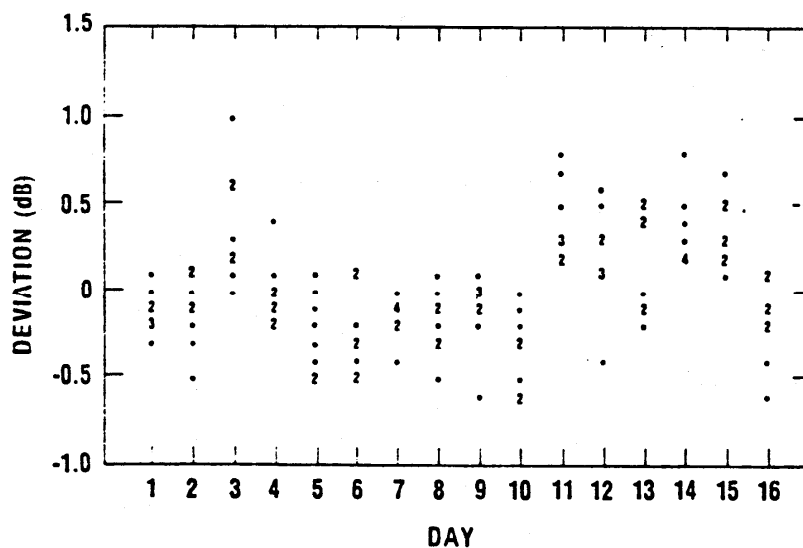


FIGURE 2: DAY-TO-DAY VARIATION IN METER READINGS WITH MULTIPLE VALUES PER DAY. (COINCIDENT POINTS INDICATED BY NUMBERS.)

One now faces the question of how to describe the variation that exists. Obviously there will be a different level of agreement expected between pairs on the same day, but this variation in no way predicts that encountered from day-to-day. The issue is not so much the statistical procedures to be used--these will follow after one defines the set of repetitions over which his conclusions must apply. For measuring the short term change in noise level, the difference between duplicates would apply; for any regulatory action, the day-to-day variation would have to be considered.

The crucial step in assessing the effects of random error is that of defining the set of repetitions over which the measurement is to apply. In the context of legal proceedings, one arrives at the degree of credibility of evidence by questions designed to find out how far the statement could be in error. In measurement, the uncertainty is arrived at by determining the amount of disagreement expected in the set of repetitions that would be appropriate in the context of the intended use of the measurement.

The Concept of a Repetition of a Measurement

Every measurement has a set of conditions in which it is presumed to be valid. At a very minimum, it is the set of repeated measurements with the same instrument-operator-procedure-configuration. (This is the type of repetition one would envision in some process control operations.) If the measurement is to be interchangeable with one made at another location, the repetition would involve different instrument-operator-procedure-environment configurations. (This type of repetition is involved in producing items to satisfy a specification and of manufacturing generally.) When the measurement is to be used for conformance to a health, safety, or environmental regulation even different methods may be involved in a "repetition."

To evaluate a measurement process some redundancy needs to be built into the system to determine the process parameters. This redundancy should be representative of the set of repetitions with which the uncertainty statement is to apply. In NBS' measurements of mass, a check standard is measured in parallel with the unknowns submitted for calibration. One thus generates a sequence of measurements of the same object covering an extended time period. From these results one can answer questions relating to the agreement expected in a recalibration and the operating characteristics of the measurement process. In this simple case the check standard is treated exactly the same way as the unknowns so that the properties of the process related to it are transferrable to the unknown.

The essential characteristic in establishing the validity of measurement is predictability that the variability remains at the same level and that the process has not drifted or shifted abruptly from its established values. One must build in redundancy in the form of a

control--the measurement of a reference quantity of known value--or by remeasuring some values by a reference method (or by an instrument with considerably smaller uncertainty). In cases where the phenomenon can be repeated, one can learn about random errors by remeasuring at a later time sufficiently far removed to guarantee independence.

In measuring an "unknown" one gets a single value, but one still is faced with the need to make a statement that allows for the scatter of the results. If we had a sufficiently long record of measurements, we could set limits within which we were fairly certain that the next measurement would lie. Such a statement should be based on a collection of independent determinations, each one similar in character to the new observation, that is to say, so that each observation of the collection and also the new observation can be considered as random drawings from the same probability distribution. These conditions will be satisfied if the collection of points is from a sufficiently broad set of environmental and operating conditions to allow all the random effects to which the process is subject to have a chance to exert their influence on the variability. Suitable collections of data can be obtained by incorporating an appropriate reference measurement into routine measurement procedures, provided they are representative of the same variability to which the "unknown" is subject. The statistical procedures for expressing the results will depend on the structure of the data but they cannot overcome deficiencies in the representativeness of the values being used.

The results from the reference item provide the basis for determining the parameters of the measurement process and the properties are transferable. One is saying, in effect, if we could have measured the "unknown" again and again, a sequence of values such as those for the reference item would have been obtained. Whether our single value is above or below the mean we cannot say, but we are fairly certain it would not differ by more than the bounds to the scatter of the values on the reference item.

The bound $\pm R$, to be used for the possible effect of random errors may be as simple as ± 3 (standard deviation) or may involve the combination of many components of variance. Once the set of repetitions over which one's conclusions must apply is defined, the structure of the random error bound can be determined.

Possible Offset of the Process

Once one has established that his measurement process is "in control" from the point of view of random variation, there remains the question of the possible offset of the process relative to other processes. It is not helpful to speak of the offset from a "true value" which exists only in the mathematical or physical model of the process. The usefulness of considering measurement in the context of legal proceedings helps clear away some of the classical confusion about

errors of measurement. In a legal or regulatory setting, one is forced to state what would be accepted as correct such as comparison (by a prescribed process) with national standards or with the results from a designated laboratory or consensus of many laboratories.

The idea of defining uncertainty as the extent to which a measurement is in doubt relative to a standard or process defined as correct finds expression in the recent Nuclear Regulatory Commission statement [12]:

70.57(a) "Traceability" means the ability to relate *individual measurement results* to national standards or nationally accepted measurement systems ... (italics added)

One could measure the offset of his process relative to the accepted process, and make suitable corrections to eliminate the offset. However, for most processes, one is content with setting bounds to the possible offset due to factors such as:

Errors in the starting standards

Departures from sought-after instrumentation (e.g., geometrical discrepancies)

Errors in procedures, environment, etc.

and other effects which are persistent. From properly designed experiments one can arrive at a limit to the possible extent of errors from these sources in answer to the question, "If the process were set up ab initio, how large a difference in their limiting means would be reasonable?"

A bound to a number of factors can be determined as part of regular measurement. For example, the effect of elevation on sound level measurements could be evaluated by occasionally duplicating a measurement at a different height and taking an appropriate fraction of the observed difference as the limit to the possible offset due to any error in setting elevation. Figure 3 shows some results from sound level meters at two heights with the source at a constant height.

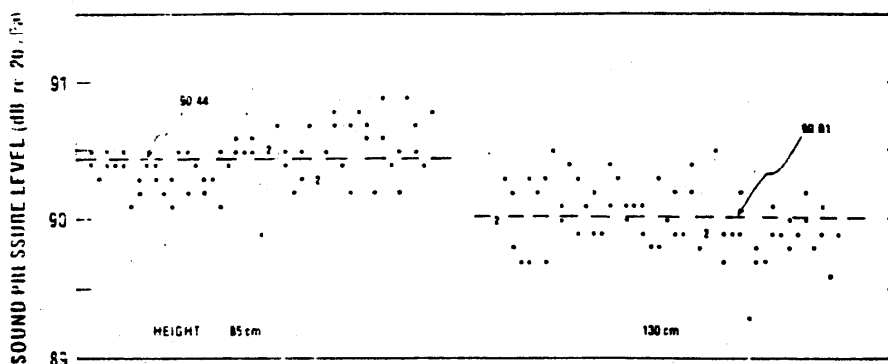


FIGURE 3: DIFFERENCE BETWEEN METER VALUES WITH CHANGE IN HEIGHT

Even if one has a functional relation, $y = f(h)$, expressing the dependence of the result, y , on height, h , one still has to carry out these measurements. The usual propagation of error approach involving partial derivatives, etc., implies that all instruments are equally dependent on the parameter under study, that there are no effects related to the factor except that contained in the formula. This can be verified for a particular instrument by actually measuring its response.

A similar comparison was made for a different orientation of the instrument with respect to this signal source and is shown in Figure 4. The effect of orientation is negligible and one would not be justified in adding an allowance for possible systematic error from this source based on a theoretical calculation.

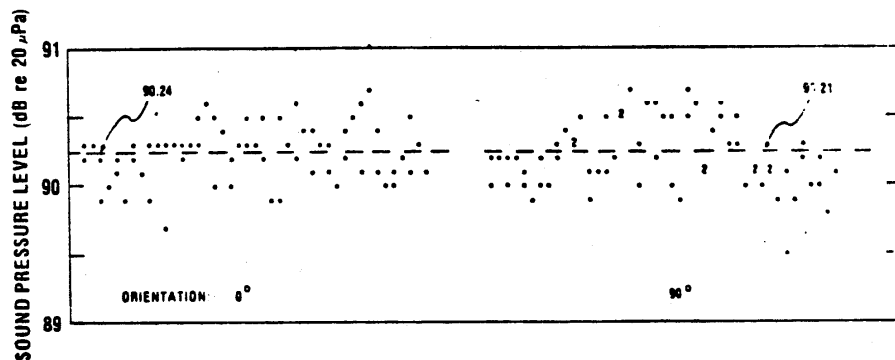


FIGURE 4: DIFFERENCE BETWEEN METER VALUES WITH A CHANGE IN ORIENTATION

From these measurements, one will have a set of bounds $E_1, E_2, E_3, \dots$ to the possible offset or systematic error from the various factors. The question as to how to combine these to a single bound to the possible offset depends on knowledge of the joint effects of two or more factors and on the physical model assumed for the process. For example, if the bounds E_i and E_j arise from independent random error

bounds, then it would be appropriate to combine them in quadrature, i.e., $\sqrt{E_i^2 + E_j^2}$. An error in the model e.g., assumed linearity even when nonlinearity exists) would act as an additive error. The properties of any combination rule can be evaluated and a selection made of the most appropriate. The result will be an overall value, E, for the possible offset for the limiting mean of the process from that of the nationally accepted process.

Uncertainty

What can one say about the uncertainty of a measurement made by a process that may be offset from the nationally accepted process by some amount +E, and is subject to random errors bounded by +R? How should these values be combined? To begin with, one could raise the question, "If the random error could be made negligible, what uncertainty would one attach to a value from the process?" Clearly the answer is +E. The next question, "If, in addition, a random error of size R is possible, what do we now say about the uncertainty?" The answer seems obvious--E and R are added to give an uncertainty of $\pm[E + R]$.

But what if E were itself the result of only random errors? The answer depends on what one calls a repetition. By the way E is defined, it is the bound for the systematic offset of the process and although it may be arrived at from consideration of random errors, the factor involved keeps the same (unknown) value throughout. Our ignorance does not make it a random variable.

Consider the case of a mass standard. NBS' certificate states that the uncertainty is based entirely on random variation, the effects from systematic errors being negligible. But unless one recalibrates, the error due to calibration remains fixed in all measurements by the user.

The uncertainty of a measurement--the width of its "shadow of doubt" in a legal proceeding--must therefore be the sum of the random error and systematic error limits.

Measurement Process Control

The essential feature for the validity of the uncertainty statement is that the process remain in a state of statistical control. Once an out-of-control condition occurs, one has lost predictability and the previous uncertainty statements are no longer valid.

To monitor the process some redundancy has to be built into the system. A variety of techniques can be used to give assurance of continued control. For example, one could periodically measure the same reference item or artifact or one could make duplicate measurements on some production items with enough delay to guarantee

independence. The American National Standards Institute Standard N15.18 for mass measurement [10] is an example where this approach is worked out in detail. But one has to verify more than just those parameters related to random variations. One needs to build in tests of the adequacy of the physical model by a variety of tests on the process (e.g., by repeating measurements under different conditions to verify the adequacy of the corrections for such changes) as well as periodic redetermination of the bounds for systematic error. One thus tests that the assumed model is still acceptable and that the parameters assigned to that model have not changed.

An excellent example of the efficacy of this approach is given by the recent announcement [6] of discrepancies of 1 mg in the assignment of mass to aluminum kilogram standards. The mass measurement system has long been shown to be nearly perfect for the usual standards. To check up on the performance of the system at densities nearer to that of most objects involved in practical measurement, an aluminum kilogram was sent to laboratories including several at high elevations. It turns out that the difference between the mass of a stainless steel and an aluminum kilogram is significantly different at different elevations. This unsuspected property of the real measurement system is now the subject of considerable study.

All measurements have some form of measurement assurance program associated with them although, as with quality control, we usually reserve the term for a formal program. In a formal program one treats the whole process--beginning with a study of the need, the development of a measuring process and a procedure for determining and monitoring its performance, and an evaluation of the effectiveness of the whole effort. One needs a criterion of success to be able to determine whether more of one's current measurement activity or perhaps some alternative would contribute most to the overall program, and this is not necessarily provided by the smallness of the uncertainty for a measurement.

For example, when the requirement is for matched sets (e.g., ball bearings) or mated assembly parts, then it is usually cheaper and more accurate to sort into finely divided classes and match for correctness of fit rather than perform direct measurement of each part.

When the measurement requirements are stated in terms of the needs of the system, (number of correctly matching parts, number of correctly measured dosimeters, etc.) one can measure success of the measurement effort in terms of closeness to meeting those goals. Measurement efficiency is thus judged in terms of the output of the organization rather than by the count of the number of significant digits. Also, one needs this measure of performance of the measurement effort to be able to identify those areas which need improvement.

Examples of Measurement Assurance Programs in NBS Measurements

Two easily described measurement assurance programs are those in mass and length. In routine calibration, a check standard is included with each set of weighings and process control is maintained by monitoring the value obtained for the check standard and of the random error from the least squares analysis [8, 9]. Control charts have been maintained since 1963. In the calibration of gage blocks, similar process control has been maintained since 1972 on both the interferometric process by which the assignment of length to the NBS master gage blocks is done and on the comparator process by which length values are transferred to customer gage blocks. [1, 7]

Similar programs are in effect in all divisions, but not all quantities involved in calibration have a formal program worthy of the name, measurement assurance.

Examples of Measurement Assurance Programs At Other Laboratories

Only two examples of measurement assurance programs at other laboratories have ever been reported. One at Autonetics [3] in length and one at Mounds Laboratory in mass. Once the mass measurement system for UF_6 is underway as part of the Safeguards program, NBS will be able to document the efficacy of the approach in practical measurement.

The NBS Measurement Assurance Programs Offered As A Part Of Our Calibration Service

Measurement Assurance Programs are listed as a calibration service in mass, volt, resistance, capacitance, voltage ratio, watt-hour meters, platinum resistance thermometry, and laser power. These are designed to measure the offset of measurement processes for the calibration of standards by other standards laboratories. These are applicable only to those laboratories who maintain and calibrate standards in the same manner as NBS. [See 11, 5, 13.]

These procedures enable a laboratory to determine the offset between its process of calibrating standards and that of NBS.

Need For Measurement Assurance Program For Practical Measurement

The UF_6 cylinder program for Safeguards [10] is an example of NBS' service in providing a direct method for measuring the offset of practical measurement processes from that accepted as correct, namely mass measurement by NBS. Investigation of the need and possible mechanisms or artifacts for monitoring the offset of practical measurements in quantities such as voltage, resistance, length, radioactivity is underway. (For examples of the application of these principles to sound level meters, see [5].)

In personnel dosimetry procedures are being worked out [14] to monitor the output of firms providing such services. In this case, a table of allowable limits of uncertainty are based on physiological considerations. Process parameters are to be determined by an initial study. Routine monitoring will be used to confirm that the process is "in control" at those levels, otherwise the parameters are redetermined *ab initio*. These "consistency" or "in control" criteria replace the usual one-time round robin approach. The amount of effort needed to establish this predictability is a function of the risk and costs of wrong decisions.

In industrial measurement we could ask

If some critical measurements on the production line were repeated would the two measurements agree?

How much bad material is passed, or good material rejected because of errors in measurement?

To those who have not properly answered these questions, dollar savings and improved product quality are possible without redesign or changes in production procedures.

Is our faith in instruments justified? Implicit faith in the correctness of instruments means that product variability (as determined by these instruments) is attributed to variability in components, raw materials or even poor design. One wonders how many times this has led to expensive changes in production procedures without apparent improvement because the variability actually arose in the measurements themselves.

How often has the installation and methods of use degraded the output of an instrument capable of much more accuracy than is required when handled properly? Without some surveillance of the actual measurements, one would never know.

One wonders how often a product is redesigned because measurement error has led to the decision that the product does not conform to specifications.

The result of this look at measurement is measurement assurance--the quality control of measurement. If adequate control exists, then one can look elsewhere for improvements in the product line. If it does not, then one has the possibility of savings without changing production procedures.

Some form of redundancy must be built into the process to answer these questions.

References

- [1] John S. Beers, "A Gage Block Measurement Process Using Single Wavelength Interferometry," NBS Monograph 152, December 1975
- [2] J. M. Cameron, "Measurement Assurance," Review of Standards and Specifications Section of Journal of Quality Technology, Vol. 8, No. 1, pp. 53-55, January 1976
- [3] J. A. Hall, "Dimensional Measurement Assurance Programs," Autonetics Report No. X73-1146/031, presentation at 1973 Conference of the National Conference of Standards Laboratories at the National Bureau of Standards, Gaithersburg, Md., November 1973
- [4] Edward B. Magrab, "Environmental Effects on Microphones and Type II Sound Level Meters," NBS Technical Note 931, October 1976
- [5] N. Michael Oldham, "A Measurement Assurance Program for Electric Energy," NBS Technical Note 930, September 1976
- [6] P. E. Pontius, "Mass Measurement: A Study of Anomalies," Science 190, pp. 379-380, October 1975
- [7] P. E. Pontius, "Measurement Assurance Program--A Case Study: Length Measurements. Part I. Long Gage Blocks (5 in. to 20 in.)," NBS Monograph 149, November 1975
- [8] P. E. Pontius, "Measurement Philosophy of the Pilot Program for Mass Calibration," NBS Technical Note 288, May 1966
- [9] P. E. Pontius, "Notes on the Fundamentals of Measurement and Measurement as a Production Process," NBSIR 74-545, September 1974
- [10] "Mass Calibration Techniques for Nuclear Material Control," N15.18-1975, American National Standards Institute, New York, New York
- [11] "Measurement Analysis Program in Mass," NBS Technical News Bulletin, September 1970
- [12] "Measurement Control Program for Special Nuclear Materials Control and Accounting (10 CFR 70.57)," Federal Register, Vol. 40, No. 155, pp. 33651-33653, August 1975
- [13] "New Calibration Service for the Volt," NBS Technical News Bulletin, pp. 46-47, February 1971
- [14] "Proposed Standard--Criteria for Testing Personnel-Dosimetry Performance," Health-Physics Society, HPSSC WG/15, October 1976