JOURNAL OF RESEARCH of the National Bureau of Standards
Vol. 86, No. 3, May–June 1981

# Foundations of Metrology

John A. Simpson*

National Bureau of Standards, Washington, DC 20234

January 7, 1981

The theory of measurement has attracted the attention of a number of philosphers whose works remain largely unknown to metrologists. Recent work in the development of Measurement Assurance Programs has demonstrated the power of this theory as a tool for guiding the development of measurement systems. The elements of the theory, especially that of Carnap and its applications to metrology, are developed as an aid to program planning and evaluation.

Key words: Epistemology; measurement; measurement assurance; metrology.

## 1. Introduction

Metrology is defined as the science of measurement, and if broadly construed would encompass the bulk of experimental physics. The term is usually used in a more restricted sense to mean that portion of measurement science used to provide, maintain, and disseminate a consistent set of units or to provide support for the enforcement of equity in trade by weights and measurement laws, or to provide data for quality control in manufacturing.

In this restricted sense metrology has taken on the nature of an art or craft rather than a science, and has attracted little academic interest. As a consequence its literature, although extensive, tends to be of an *ad hoc* character, is widely scattered, and appears mainly in the form of reports or internal documents. There exists no extensive systematic treatment of the subject comparable to the great texts of other disciplines. However, the subject does have an internal logical structure, one version of which has been articulated at NBS over the past two decades as the concept of Measurement Assurance Programs has been developed and applied to measurement services.

While presenting in some detail this version, our treatment does not aspire to be a definitive text but rather to provide an overview of the subject to give those responsible for managing organizations active in metrology a conceptual grasp of the subject sufficient for intelligent program planning and evaluation. Because of the background of the author the few examples given will generally be taken from the field of mechanical metrology; but the principles illustrated are task independent.

## 2. Context of Measurements

A measurement is a series of manipulations of physical objects or systems according to a defined protocol which results in a number. The number is proported to uniquely represent the magnitude (or intensity) of some quantity[1] embodied in the test object. This number is acquired to form the basis of a decision effecting some human goal or satisfying some human need the satisfaction of which depends on the properties of test object.

These needs or goals can be usefully viewed as requiring three general classes of measurements.

1. *Technical:* This class includes those measurements made to assure dimensional compatibility, conformation to design specifications necessary for proper function or, in general, all measurements made to insure fitness for intended use of some object.

2. *Legal:* This class includes those measurements made to insure compliance with a law or a regulation. This class is the concern of Weights and Measures bodies, regulators and those who must comply with those regulations. The measurements are identical in kind with those of technical metrology but are usually embedded in a much more formal structure. Legal metrology is much more prevalent in Europe than in the United States, although this is changing.

3. *Scientific:* This class includes those measurements made to validate theories of the nature of the universe or to suggest new theories. These measurements, which can be

*Center for Mechanical Engineering and Process Technology, National Bureau of Standards.

[1] For our purposes we adopt the B. Ellis [1][2] definition of a quantity. A quantity is a kind of property that admits of degrees, and which therefore is to be contrasted with those properties that have an all or nothing character (for example, being pregnant).

[2] Figures in brackets indicate the literature references at the end of this paper.

called scientific metrology, properly the domain of experimental physics, present special problems and will be dealt with only briefly at the end of this paper.

The path of reasoning between an identified goal or need and the measurement to attain that goal is a thorny one. Many valid measurements do not result in useful information. The property measured often does not adequately predict the fitness for use. Often quantities for measurements are chosen for their convenience rather than their importance. The metrologist is seldom in a position to do much about this unfortunate state of affairs. This problem is the concern of the design engineer, the regulator, the Quality Control Manager, or in brief, the decision maker. Supplying the decision makers with the most reliable numbers characterizing the properties they have designated, in the most economical manner, is all metrologists can do in their professional capacity.

This task, although limited, is a worthy one requiring all of the ingenuity, knowledge, and professionalism one can muster. It is a two-fold task: one must generate a measurement system, which in the NBS view is a *production system* whose product is numbers, and a *quality control system* to confirm the validity of those numbers [2].

The first of these tasks is an engineering hardware problem while the second is largely a software management problem. The software consists of record keeping, reporting, qualification, and similar activities often depending heavily on statistical mathematics. We will deal with each in turn.

## 3. Elements of a Measurement System

There are many ways of enumerating the elements of a measurement system since it consists of these eight elements, combined in more or less complex groupings:

Physical Concepts
Physical Laws
Instruments
Standards
Human Operators
Procedures
Environments
Computations

It has proved useful over the years to group these elements under two general headings: *Properties of the Investigated Object*[3] and *Properties of the Measurement Algorithm* where the *Investigated Object* is the subject to be measured and the *Measurement Algorithm* includes all means and procedures used to produce the desired number. This grouping is useful for identifying sources of error and for remedial action when such are discovered. The procedures for successfully accomplishing such action differ markedly depending on the grouping in which the faulty element lies. This important fact seems to have first been recognized by Volodreskii [3].

### 3.1. The Role of the Investigated Object

The investigated object (henceforth shortened to object) plays two essential roles in a measurement system: it must embody the quantity of interest and it must generate a signal to which the measurement algorithm can respond. This signal must be unambiguously related to the magnitude or intensity of that specified quantity. Knowledge of the relationship between the quantity and the signal requires a model of the object. This model is based on the laws of physics or our understanding of the universe. It is usually a software model, and equation or the like which quantitatively predicts the signal as a function of the quantity to be measured. Unfortunately, objects have complex natures and hence seldom are perfect embodiments of single quantities. For example, the Kilogram of Paris embodies a specific volume as well as the unit of mass: a standard cell is not a "pure" voltage source but rather such a source in series with a non-linear complex impedance. Moreover the magnitude of the quantity of interest in the object may itself be a function of environmental parameters not of immediate interest. The length of a material body, say a gage block, is intrinsically a function of the temperature of that block. The model must include all relevant properties of the object.[4]

The model must also predict the signal that is to be used to drive the measurement algorithm. This signal is almost invariably a quantity which differs in nature from the quantity to be measured. For example, the beam of the common balance used in mass determinations responds to a force signal generated by gravity operating on the mass of the object in the pan. Many objects generate more than one signal that could be used for measurement. A gage block as an embodiment of length, can, if the measurement algorithm is a caliper, generate a force signal as the jaws close on the block, an optical signal, if measured by an interferometer, or an electrical signal, if used in a capacitance arrangement. Any of these signals can be and are used, the choice being made on considerations of convenience or current state-of-

[3] The investigated object may in fact be a complex system with internal structure but for purposes of this discussion the word object has the advantage of compactness of expression and no generality is lost.

[4] What constitutes a relevant parameter is a problem that has attracted philosophical attention, Rudolph Carnap [4], for example. In practice, except at the highest levels of scientific metrology, enough is known about the object that identification of the relevant parameters is easy; in any event, if one parameter is overlooked the resultingly high observed uncertainty of the measurements will soon call this fact to one's attention.

the-art in measurement algorithm development. While this signal redundancy makes life simpler, the fact that most signals are generated by several quantities embodied in the object makes life at times difficult. For example, the force signal generated by the mass of an object on a balance pan is contaminated with a signal identical in kind generated by the buoyant force of the air displaced by the object's volume. This particular problem has recently given the mass community problems [5, 6]. In length metrology the fact that the distance between the apparent reflection planes (the optical length) of a gage block depends both on the length of the block and the complex dielectric constant of the material remains an unsolved problem limiting among other things work on absolute density.

The signal, besides being impure, may also be a function of environmental parameters even if the quantity itself is not. A case in point is the dependence of gravity force generated by a mass on the value of the local acceleration of gravity; here, then, the signal is a function of location while the mass is not.

More generally the nature of the object can be expressed as a matrix where the rows are all the physical quantities embodied in the object while the columns are the all of the signals generated by that object. An ideal object would be one in which the matrix was diagonal, in the sense that for every quantity there would be one and only one signal. No such object exists. The proper treatment of the off-diagonal terms is one of the central problems of metrology. We shall return to this problem in section 4.

In any event the first step in constructing a measurement system is to reduce the object to an idealized model which represents those properties or attributes believed to be germane to the intended measurement, i.e., those which satisfactorily predict the signal as a function of the magnitude or intensity of the desired quantity. For example, a gage block may be modeled for a force-based algorithm as an impenetrable rectangular parallelpiped characterized by a single length between its gaging surfaces. However, in this case the model is too simplified for most purposes, and current models include the fact that length is a function of temperature, that the block is elastic, deforming on contact, and that the faces may be non-parallel. Thus, the model may be simple or complex, where complexity and desired accuracy go hand in hand, but the model only weakly reflects the measuring system by being required to predict the signal to which the chosen measurement system responds. The converse is not true; the measurement system reflects the model strongly since it must provide all of the parameters necessary to permit the model to predict the quantity from the observed signal or signals. Hence, generally a more complex model will call for a more complex measurement system measuring a greater number of properties of the object or of the environment.

The model is never complete or perfect and the difference between the model and the real properties, including the signal expected, is called *model ambiguity.* The model ambiguity sets a lower limit on the uncertainty of the measurement since below this level the object is not in fact fully defined. In more complex objects this model ambiguity is most often the dominant uncertainty term; an example that comes to mind arises in screw thread metrology where a measured quantity, flank angle, implies a model in which the thread flanks are planes. In practice, when dealing with carefully gound thread gages, this model is useful. However, in fasteners made by use of a die or a roll, the flanks most definitely are not planes, and flank angle loses its meaning.

Model ambiguity is a particular type of systematic error which exists if the measurement algorithm is flawless. Failure to recognize this fact can lead to major wastes of resources since no improvement in the measurement algorithm can reduce this error. No amount of research on precision balances will reduce the inconsistencies of the mass scale caused by air buoyancy correction problems. Model ambiguities are the source of the vast majorities of measurement inconsistencies which can only be reduced by improvement of the model.

Given that a certain condition is satisfied there exists a strategy which can reduce model ambiguity identically to zero. This strategy uses objects called variously "prototypes," "artifact," or "gold plated" standards and, in effect, takes a particular object and defines it to be its own model. This amounts to saying that this particular object is the perfect and complete realization of the class of objects to which it belongs and hence the model ambiguity is, by definition, identically zero. The condition to be satisfied is that all objects to which the standard refers must be essentially identical to the standard both in kind and in degree. For example in mass, the only SI unit still using this strategy where the Paris Kilogram is the kilogram of mass, the only objects where mass can be unequivocally defined are one kilogram weights made of platinum. All other objects differing in size or material have masses that can only be approximated (admittedly to a high degree) by comparison with the kilogram. The strategy has the further disadvantage that if the prototype is lost, destroyed, or changed in value, all objects of its class must be recalibrated. In principle, if someone drops the Paris Kilogram, every scale in the world would be out of calibration the instant it hit the floor.

However, lower down the measurement hierarchy the strategy works well; for example, the American Petroleum Institute pipe threads, where sets of "gold plated" gages kept at NBS and other National Laboratories can be compared to almost identical working gages by algorithms much simpler than those required to compare a gage to a drawing. The problem of extensibility, i.e., using a two-inch gage to calibrate a three-inch gage never arises and the gages are so close to identical that functionally no harm is done by replacing a worn or broken gage by the last good gage and carrying on. For highly derived or complex objects, gears are another example that comes to mind. The possibility of using this ploy should always be explored since it is extremely efficient in those cases where it can be used, even though it requires a separate standard for each and every size of a class of objects.

In the search for model ambiguities it is often possible to use a measurement algorithm known from another context to be error free (at some level of accuracy) to measure the object. In this case the total uncertainty can be ascribed to the model ambiguity. The use of high precision optical interferometry to read caliper jaw movement when checking for the correction due to the elastic deformation of the object under the force of the caliper jaws is an example. Optical interferometry can, for this purpose, be considered error free.

## 4. The Measurement Algorithm

In our classification, the measurement algorithm includes everything involved in the measurement except the object and the final number. It includes the instruments used, the protocols followed, the characteristics of the human operator, the calculations made and the environment in which they operate. In brief it is the "factory" whose raw materials are objects and whose product is numbers. Just as in the case of the object we must have a model of this "factory" which predicts how it treats the signal from the object, processes it, and generates the number. To be fully satisfactory the model must account for the effects of environment on this process and, most importantly, predict how it "loads" the object signal source and hence affects the relationship between the quantity and the signal of the object. Neglect of this factor, typified by using a low impedance voltmeter on a high impedance source or using a high force caliper to measure the diameter of an egg, leads to gross errors.

The process, if it is to be useful, must generate numbers which have certain properties. These properties arise out of our expectations concerning them. We would like to use them as surrogates of measurement, i.e., once they are obtained for a stable object we would like to use them to avoid measuring the object at a future time or a different place. Prepackaged food is a clear example where the scale in the manufacturing plant virtually eliminates weighing in every store. However, to accomplish this goal we must be assured that, within some predetermined uncertainty, every competent metrologist with suitable equipment at any different point in the space/time continuum would assign to the same object the same numbers representing the same quantity that we did. When we have accomplished this often difficult feat we say we have a *proper measurement algorithm* and our numbers are *proper measurements.*

The concept of properness is a generalization of the concept of precision or reproducibility often used by writers on measurement. We prefer the more general term since it is often not clear whether the authors refer to the spread between successive repeated measurements, between runs, between measuring stations, or between laboratories. With properness you are assured you are working with worse-case figures.

Before we discuss the details of the measurement algorithm and how we accomplish a proper measurement, we must examine some general principles which allows us to define in broad generalities what constitutes a "competent metrologist with suitable equipment." Our guidance in this case comes not from the theory of physics but from philosophy. Rudolph Carnap and similar thinkers [1, 4] have articulated the requirements of any measurement algorithm which is to yield proper measurements. They also simultaneously determine the minimum *a priori* information that a metrologist must have in order to be competent to duplicate another's measurement.

Although based on Carnap, what follows is a modification in detail of his exposition and somewhat an extension. It may be called the NBS school. The differences will not be explained in detail, only noted in passing. The position starts from an operational point of view and considers that every measurement algorithm defines the quantity measured.[5]

Our position is that, for example, interferometry defines the optical length of a gage block and a caliper defines its mechanical length. These lengths are separate and distinct properties of the block and logically unrelated. One arbitrarily chooses which length to measure on grounds of intended use or convenience. In this view, optical length and mechanical length are not imperfect measures of "true" length but independent quantities of full stature each in its own right. The question of "true" length is considered moot since it cannot be decided upon by a "true" measurement algorithm. Obviously such a posture gives rise to problems in the relationship of measurement to experimental physics, some of which will be touched on later;[6] however, in technical or legal metrology, since all the different lengths are in fact within less than a micrometer of being the same, it is perfectly practical to adopt this non-judgmental point of view.

For any such length or other quantity, a competent metrologist with suitable equipment is one who has a realization of four axioms.

[5] This position is closer to P. W. Bridgeman [7] than to Carnap who takes a more expansive view, including an operational definition as only one element. For a contradictory position, see H. C. Byerly and V. A. Lazara [8].

[6] For treatments of these difficulties, see Byerly and Lazara, op. cit. or Ellis, op. cit.

1. Within the domain of objects accessible one object must be the unit.

2. Within the domain of objects accessible one object must be zero.[7]

3. There must be a realizable operation to order the objects as to magnitude or intensity of the measured quantity.

4. There must be an algorithm to generate a scale between zero and the unit.[8]

To make the above clear, consider the following system:

> The quantity of interest, temperature;
> the object, a human body;
> the unit, the boiling point of water 100°;
> the zero, the triple point of water 0°;
> the ordering operator, the height of a mercury column in a uniform bore tubing connected to a reservoir capable of being placed in a body cavity;
> the scale, shall be a linear subdivision between the heights of the column when in equilibrium with boiling water and water at the triple point.

This is a well-known system the properness of which is the basis of medical diagnosis.[9]

Once a measurement has been made the test for competency stiffens and all other metrologists to be considered competent must have the identical realizations of the axioms.

The essence of designing a measurement algorithm capable of making proper measurements is choosing realizations of these axiometric operators which are capable of independent replication by the universe of metrologists interested in the measurements.

For certain parts of the task one has a great deal of help. The Treaty of the Meter sets up an international structure of various organizations, including the International Bureau of Weights and Measures, charged with defining units for additive quantities and both units and zeros for nonadditive quantities. The International Bureau also disseminates such units by "prototype standards" (Kilograms Numbers 4 and 20 for the U.S.) or prescriptions such as those for the realization of the $Kr^{86}$ standard of length or the Cs second. Many other standards groups do the same for highly derived standards such as pipe gages from the American Petroleum Institute.

The zero units for extensive quantities are usually commonly agreed upon as null objects such as no mass on the balance or a short circuit (beware of thermal voltages) on a voltmeter.

The scale generation usually does not become a matter of controversy in the practical world although many have suggested it plays a central role in scientific metrology.[10] Whether to adopt 212 or 100 degrees between fixed temperature points or to divide an inch into thousandths or 25.4 mm can usually be worked out between metrologists.

The crux of most cases of improper, or allegedly improper, measurements, lies in the ordering operator. There are very few operators which have the authority of an international or national standards body. ISO has defined the ordering operator for gage blocks but not, for example, for ring gages. A cylinder, if it is used as a gear or thread wire, has a defined ordering operator but has none if it is a plug gage or the piston of a deadweight pressure generator. Decades of controversy surround the ordering operator, actually the much simpler equality operator (a special case of an ordering operator), for threaded fasteners.

The philosophers of science give little guidance on the process by which the metrologist makes the choice between all possible measurement algorithms which can be developed to satisfy the measurement axioms. Carnap sums up the total guidance as follows:[11]

> "We thus have a genuine choice to make here. It is not a choice between a right and wrong measuring procedure, but a choice based on simplicity. We find that if we choose the pendulum as our basis of time the resulting system of physical laws will be enormously simpler than if we choose my pulse beat. —This simplicity would disappear if we based our system of time measurement on a process which did not belong to a very large class of mutually equivalent processes."

Leaving aside the question of what is a simple law and how one establishes mutual equivalency without a preconceived measurement process, this advice is not particularly helpful to the practicing metrologist. What, if any, effect on physical laws a particular definition of, say, flank angle on a screw thread or specific rotation of a sugar solution for customs regulation, can be expected to have is at best obscure. There is even less help available as to how the model of the algorithm chosen is reduced to practice, i.e., hardware and protocols. To usefully attack this problem it is necessary to introduce the concept of *limited properness*.

[7] The criteria outlined here are suitable for all physical quantities and are the most general ones. For those quantities for which an addition operation can be defined a simpler set of three axioms is possible. However, measurement systems built on additivity are awkward in practice and even in mass where a particularly simple addition operator is available the least significant digits are obtained by use of a four axiom system. Gage blocks were an attempt to utilize an addition operation system but modern practice calibrates them by four axiom methods although the addition operator is important in their practical use.

[8] These operators must satisfy certain formal requirements as to symmetry, transitiveness, etc. For details see Carnap, op. cit., Chap. 6.

[9] There are some problems relating it to fundamental concepts, see Ellis, op. cit., Chap. VI.

[10] See Ellis, Byerly, previously cited works.

[11] R. Carnap, op. cit., chapter 8

All measurements taking place in the real world exhibit an intrinsic limited properness in that equality between reproduction will always have an uncertainty whose lower bound is either Johnson (*kT*) noise or that set by the Heisenberg uncertainty principle. Few measurements in technical or legal metrology approach either limit. The more important limits of properness are under the control of the metrologist and are those introduced by adopting a model for the object or the algorithm which is known to be imperfect. The governing principle is to pick the measurement system where the total measurement uncertainty can be demonstrated to be less than, but not wastefully less than, the uncertainty needed by the decision maker. It makes little sense to measure the dimensions of ceiling tile by laser interferometry when the decision to be made is whether or not the joints will appear straight to the naked eye.

There are several distinct manners in which the economies inherent in limited properness can be realized. The most frequently used manner is to restrict the class of objects "suitable" for measurement. An excellent example is the detailed design specifications applied to weights used in trade. By restricting the material, and hence the density, it is possible for legal metrology purposes to simplify the measurement algorithm and, for instance, eliminate an explicit air buoyancy correction. Such a procedure appears a direct violation of the Carnap formal requirement that all operators satisfy a connectedness property, that in the domain of the quantity *M* any object a or b which has *M* is either equal in *M* or one has less *M* than the other. What has been done in introducing the concept of "suitable" is to redefine what we mean by *any object.* Little damage to the logical structure results from such a choice.

A second choice is to limit the environmental conditions under which the measurement can be implemented. The length metrologists' insistence on 20° C working environment is an example of this way of simplifying an algorithm, or perhaps more succinctly, of restricting the universe.

The third strategy which can be used is to limit the range of magnitude to be covered.[12] This is popular in the temperature field where more than two fixed points have been defined and different interpolation algorithms are defined between pairs of them. All such strategies should be explored before a choice of measurement systems is finalized.

After a preliminary choice is made, it is useful to analyze the system for sources of uncertainty. This analysis is useful only if the object's model ambiguity has first been determined. Uncertainties or errors can enter at any of the realizations of the axioms.

There may be a unit error, a scale error, or a comparison error. Each of these errors can arise either because the realization is imperfect or, more frequently, because the realization has not been described in sufficient detail that the concerned "competent metrologists" have been able to effectively replicate it. The first of these causes can be attacked by high quality engineering, making use of all that is known of instrument design [9], properties of materials and precision workmanship and by a generalized form of experimental design.

There are three basic strategies to accomplish error control by design of the measurement which have developed over the years and which can often provide a useful conceptual framework within which to attack a given problem. The strategies deal directly with the basic problem that both the models of the object and of the candidate algorithm have to contend with mixed (non-single quantity) signals and responses, or in our matrix formalism, off-diagonal terms. For concreteness, let us consider the measurement of the *x*, *y* coordinates of *n* points in a plane by use of a two-axis measuring machine. The ideal situation would give a set of idealized observational equations as follows:

$$\begin{aligned} x_i &= k\,\hat{x}_i \\ y_i &= k'\hat{y}_i \end{aligned} \qquad (1)$$

where $x_i$ and $y_i$ are the true coordinates, i.e. in a coordinate frame attached to the workpiece, $\hat{x}$ and $\hat{y}$ are the *x* and *y* axis scale readings of the machine and *k* and *k'* are the invariant scale constants of the machine which implicitly contain the unit.

Unfortunately, machines are not geometrically perfect and if the *x* and *y* axes are not orthogonal the observational equations will develop off-diagonal, or coupling, terms, i.e.

$$\begin{aligned} x_i &= k\,\hat{x}_i + \alpha\hat{y}_i \\ y_i &= k'\hat{y}_i + \alpha'\hat{x}_i. \end{aligned} \qquad (2)$$

If the axes are curved or Abbe angle errors exist, the equations become still more complex

$$\begin{aligned} x_i &= k\,\hat{x}_i + \alpha\hat{y}_i + \beta\hat{y}_i^2 + .... + \gamma\hat{x}^2 + .... \\ y_i &= k'\hat{y}_i + \alpha'\hat{x}_i + \beta'\hat{x}_i^2 + .... + \gamma' y^2 + .... \end{aligned} \qquad (3)$$

where the $\gamma$-like terms reflect scale nonuniformities. In a measuring machine with a screw for reference, they might well be written in sine form to characterize periodic errors.

In general, all of the coefficients are functions of temperature and hence, if the temperature changes, become functions of time with various time delays. The problem of

[12] Support for this strategy is implied by Carnap, op. cit., chapter 10

algorithm design is to find optimum ways of dealing with the *n* equation array of eq (3).

Three general strategies have developed. The first, called by J. Bryan, "brute force," is to develop auxiliary or test algorithms which measure the off-diagonal terms and then to reduce them by "reworking" the machine (algorithm) until they become insignificant. The machine is then treated as "perfect," and equations of type 1 are used for the measurement. Since the implementation of the auxiliary algorithms is, of necessity, spread out over a long time compared to the time for a single measurement, temporal stability is of critical importance and, hence, leads a legitimate emphasis on temperature control.

A second strategy, which might be called "correction," is to measure in auxiliary or test algorithms the off-diagonal terms and then either by analog devices on the machine, i.e., compensator bars, or by computation, render them harmless. Note that, for example, eqs (2) become linear and much easier to deal with if $\alpha$ is a constant and not a variable unknown. If the off-diagonal terms are allowed to remain large, the stress put on the temporal stability is even more severe than in the "brute force" technique where the coefficients are forced to be negligible in size. Failure to provide this temporal stability by adequate temperature control probably accounts for the historical failure of this technique. Another difficulty with this approach is that it is difficult to derive auxiliary algorithms which measure the desired coefficients directly and these coefficients tend to be complex combinations of the auxiliary scale readings. The strategy moves the problem to the auxiliary system where it may or may not be easier to solve. For example, on the three-axis measuring machine, $\alpha$ is a combination of axes nonorthogonality, $y$ roll and the $y$ axis arm length. In three dimensions $\alpha$ becomes, moreover, a function of $z$. Multi-parameter factors are difficult to deal with in the analysis. A major advantage of the "brute force" technique is that any combination of negligible quantities is negligible and, hence, the detailed dependence of the coefficients on the auxiliary quantities need not be worked out.

The third strategy and the one currently being explored at NBS is a conceptually straightforward attempt to solve eqs (3) in all their complexity. It has been called a Redundant Algorithm because the coefficients ($k$, $\alpha$, etc.) as well as the variables ($x_i y_i$) are treated as unknowns. There must be many more observational equations, and hence, measurements, than the "*n*" variables the algorithm sets out to measure. Looked at another way, all of the measurements which are auxiliary in the other schemes are lumped together with the desired measurement into a single procedure. The measurement need not be redundant in the statistical sense.

The greatest advantage of this attack is that the "calibration" of the machine occurs simultaneously with the measurement instead of days, weeks, or months apart, and the question of loss of calibration by misadventure cannot occur. For instance, the "four points in a plane" algorithm we have tested takes about one hour to perform. It consists of measuring the $x$, $y$ position of four points on a plane repeated with the plane rotated approximately 90°. It accomplishes the auxiliary measurements, an "absolute" calibration of a rotary table at 90° increments, and a measurement of the orthogonality of the machine axes. There are, in fact, sufficient measurements taken to determine in principle 24 coefficients. This telescoping in time greatly reduces the demands on the temporal stability of the machine, especially since that portion of the drift in each coefficient which is linear with time can be eliminated relatively simply by the introduction of more explicitly time-dependent coefficients. This particular ploy has been used successfully in our gage block laboratory for a number of years.

The comparison of methods cannot be complete, however, without a discussion of the different manner in which the second part of the reported number, the error estimate, is obtained. The error estimates reflect a considerable difference in philosophy, although both the "brute force" and "correction" strategies divide the error into two parts, a "random" and a "systematic." The random component is obtained by repeating the measurements, both prime and auxiliary, a number of times, and inferring the variance around the mean by the rules of elementary statistics. The systematic component is in the "brute force" method bounded by a "worse case" calculation based on the residual values of the off-diagonal terms after the machine has been refined to its current state. This results in a conservative estimate of the error in most cases. It is essentially this calculation which defines the term "insignificant" as a goal for machine correction. An insignificant fault is one whose maximum possible effect on the measurement is less than some limit set by end-use of the measured object.

There remains a danger, which may be remote, but to which most metrologists have fallen victim. This danger involves what one may call hidden coefficients; these are variables which affect the measurement but which are not modeled by the observational equations. For example, suppose one neglects temperature change in a gage block measurement. The protection against such an oversight is redundancy by repeating the measurement, averaging, and observing the deviation from the mean which will reflect this temperature drift if it is significant alert the metrologist. As has been so often pointed out, the "random" variations of metrology are seldom random in the statistical sense, but reflect a wider spectrum of ills, the change of an unmodeled parameter being foremost among them. This protection achieved by averaging is far from absolute since in the limited period of the measurement the critical variable may not change, as it might at some future time, but the protection is nonetheless helpful. To obtain this limited help, the redundance must "span" the measurement in question. Where there are gaps, as in the time between the "calibration" and "use" of an instrument, this assurance is missing. Any super-redundancy, i.e., statistical redundancy beyond that needed to characterize all parameters of the model, introduced into a redundant algorithm spans the complete process and avoids this trap. Note also that in this scheme, there is no separation of "random" from "systematic" and the indices of fit derived from the massive least squares solution of the now overdetermined equations are the "metrologically random" errors of the complete measurement process. They reflect not "worse case" but the case that existed over the range of parameters used in that particular measurement.

The use of a single massive least squares adjustment has another advantage which arises from the peculiar nature of the coefficient $k$. This coefficient of the principal diagonal term introduces the unit into the measurement and, hence, has a special importance. The unit while vital to the measurement cannot be checked by the usual forms of redundancy since the laws of the universe operate independently of the units in which they are expressed. The reverence in which "standards" are held reflects their critical nature. In a super-redundant algorithm the unit may be introduced at several points in an independent, i.e., truly redundant, manner. For example, in our gage block algorithm it is entered in the comparator constant $k$ and in the difference $x_1-x_2$ of two masters. This provides a check of $x_1$, $x_2$, and $k$ which is difficult to obtain in any other way and provides further protection against mistakes.

The "correction" strategy is similar in principle to the "brute force" in its treatment of errors except, in this case, it is possible in theory to calculate the "actual" rather than the "worst case" systematics.

At this point we can begin to see the relative advantages of the different strategies and types of measurement programs to which they are most adapted.

The "brute force" technique requires a large initial "capital" investment in characterizing the machine over its entire working volume on a "does not exceed" basis. Also required is the establishment of an environment, both physical and procedural, that assures the maintenance of this level of performance over a "longish" time span. Depending on the requirements on accuracy, an investment in machine upgrading may also be required. However, once these conditions are met, production runs are rapid, simple to perform, and any workpiece within the machine's capacity can be characterized on a valid "maximum deviation from nominal" (tolerance) basis. It is obviously advantageous where the piece part tolerance is significantly larger than machine tolerance or where the part has a model ambiguity which is large, i.e., the piece is complex or only moderately well characterized. I would expect that the products of high precision industry lie in this class of objects.

The super-redundant strategy on the other hand requires little or no investment in machine characterization. It does, however, require a considerable investment in computer programming which is applicable to only one narrow class of objects. Moreover, the "production" runs will inevitably be more time consuming since calibration time is not spread over more than a single measurement. It does, however, offer the promise of higher ultimate accuracy since the machine needs only short-term repeatability. It also offers rigorous characterization of precision and accuracy of the values obtained.

It is advantageous in those instances where comparatively few types of workpieces are measured but where the measurements are required to be the absolute "best" in terms of accuracy and confidence in that accuracy. This requirement, of course, implies that the workpieces are simple enough in form and high enough in workmanship that the model ambiguity warrants such measurements. This workload is characteristic of the master gages with which NBS deals.

The "correction" strategy requires an inordinate capital investment in complete functional machine characterization and extensive computation on outlay which would only be justified if the brute force method was insufficiently accurate while simultaneously the workload was too heavy or diverse to make the super-redundant approach feasible. I know of no such circumstances other than when the scale of the workpieces becomes so large that measuring machine accuracy is extremely difficult to achieve, as in the aircraft industry, shipbuilding, or heavy ordnance.

Regardless of the strategy adopted there remains the problem of transferring the measurement algorithm to all interested metrologists. This communication problem is attacked largely through the voluntary standards community where measurement algorithms can be institutionalized and disseminated widely as test methods or recommended practices. The process of adoption of standards can be painfully slow.

## 5. Measurement Quality Control

Measurement quality control starts as soon as the first measurement is made. The principal tool for the metrologist is redundancy. One repeats the measurement on a stable object and compares each successive measurement with the set of all measurements. It is typical of all real systems that there will be a spread in values. Statistics tell how to derive indices of precision (reproducibility) for each system. The goal is to produce a series of values which demonstrate a stable[13] mean and a random distribution around that mean. When this situation is achieved within a single run, within-group statistical control is said to be achieved. But this within-group control is not enough; the same statistical control must be achieved between runs, between *measurement stations* within the initiating laboratory, and finally between all competent metrologists. Only after this task is accomplished can one have assurance of a proper measurement system turning out proper measurements.

Over the years a number of institutionalized aids have developed for the process of obtaining properness of measurements and the maintaining of this quality over time. The first of these institutions to develop addressed the detection and elimination of the unit error, one which seems to have dominated in earlier times. The institution is still dominant in weights and measures, and in legal metrology in general.

This institution is a calibration network where stable artifacts, often realizations of units, are sent to a different location where a "higher" artifact or unit is maintained. Working standards are "calibrated," i.e., compared with a similar object, often a secondary standard which in turn had been compared with a national primary standard and so on up the line. Since the artifacts are almost identical, model ambiguities are low and when returned to its original site the artifact provides both a source of the unit and often a one-point check of the total system. For simple, stable objects, the system has some virtue, especially if two or more calibrated objects can be used to provide an independent, even although only one point, system check.

This calibration scheme gave rise to the concept of measurement traceability [10] which wrote into contracts the requirement that such a chain be established. The system has some shortcomings:

1. it requires a stable non-ephemeral artifact;
2. it requires a measurement robust against environment;
3. it is expensive if artifact is large, fragile or complex;
4. it provides at best only a one-point check of the system; and,
5. it focuses on the quality of means of measurement rather than the measurements themselves.

To deal with cases where no stable artifact exists or where it is ephemeral in the sense that the accepted measurement algorithm is destructive, the concept of a Standard Reference Material was developed. Because chemical analysis tends to be destructive, the first SRM's were pure chemicals, solutions, or mixtures which were carefully prepared, characterized by a standards institution and made available to users to check their measurement algorithm. The system was later refined to reduce the model ambiguity by making the relevant properties of the SRM as close to those of the object of interest as possible, giving rise to such SRM as urban dust.

A newer version of the strategy is the Standard Reference Artifact; this was initially used at NBS for neutral photographic density. These artifacts are only marginally non-ephemeral, and it was introduced to alleviate the problem. In another case, linewidth standards, it is an attempt to realize some economies of scale and provide quicker response than calibration. The problem of model ambiguities must be considered carefully since the SRA's can sometimes be of higher quality than the working standards for which they substitute. This factor is one which has always limited the effectiveness of "round robins" which depend on artifacts similar in nature to SRA's.

The most highly developed OC mechanisms for measurements are Measurement Assurance Programs. These programs, based on the pioneering collaborative work of P. Pontius and J. M. Cameron [11] at NBS in the 1960's, have become central to the Bureau's measurement services. It is difficult to explain in a few words what constitutes a MAP. A MAP is basically a measurement system which explicitly, self consciously, deliberately builds in and documents at every step tests for the properness of the measurements. MAP's are characterized by carefully thought-out redundancy and often use "self calibrating" measurement algorithms; they tend to make use of modern statistical methods and focus on the measurement results rather than on "standards" or calibrations. Hence they are software rather than hardware oriented. Since they were first applied to systems where either the quality requirements are very stringent or where the degree of assurance needed is very high, as in the accounting for Special Nuclear Materials, they are often thought of as complex and expensive to implement. This is a misconception; for a given quality or degree of assurance, they have proven to be the most efficient institution yet designed. If they have a disadvantage it is that they increase the burden on the standards organization responsible for them. This fact arises because properness must be assured on a continuing basis, and there must be a constant interchange of data and periodic exchange of artifacts between the standards lab and the operating partners. Depending on the stability of the test objects exchanged, the frequency of such exchanges may approach or equal calibration intervals. In some cases this burden can be reduced by using test objects which are more stable or rugged than the accepted standard. Voltage is such a case, where banks of fragile standard cells are compared by measurements on zener diodes.

[13] Stable is used in the expanded sense that the mean is constant during a "within group" interval; slow, well-behaved, constant changes as in the case of the slow phase change of the steel in a gage block or the slow decay of intensity of a incandescent lamp present no problems in that the mean is predictable, if not absolutely constant.

The problem of standard laboratory workload can be completely circumvented by making the standards laboratory a participant rather than an organizer. This pattern is already working for electrical quantities among the aerospace industry of the West Coast, and the Bureau has plans to formally institutionalize these regional MAP's in the future.

## 6. Scientific Metrology

When the measurements with which one is concerned are undertaken with a view towards understanding the physical universe, a series of issues is raised which do not exist in either technical or legal metrology.

The most important of these new issues is that the freedom to choose a measurement system on the grounds of convenience or economy is no longer legitimate. Of all the possible measurable quantities called mass or length which, in the Carnap view, are defined by the measurement system and which are logically independent and wholly arbitrary, only one can reflect the *concept* of mass which satisfies both the laws of Newton and of Einstein. The laws of physics appear to require what the wise metrologist avoids: a "true," "absolute," or *proper* quantity. A proper quantity cannot be defined by an international body, and indeed the BIPM committee on units has been careful to avoid this pitfall. A proper quantity is somehow defined by the underlying logic of the physical universe as reflected and (imperfectly) translated by the laws of physics.

The problem is to determine which quantity defined by a measurement system is the proper one. This problem has been addressed by a number of authors since it is implicit in the fundamentals of the philosophy of physics.[14] The problem could be attacked experimentally by use of a redundant set of fundamental constants.

There would have to be generated several sets of fundamental constants derived from experiments differing only in manner by which one quantity, say mass, was operationally defined by a set of measurement axioms. The self consistency of each set would then be a measure of the properness of the corresponding quantity called mass. This procedure would have to be repeated for each of the SI base quantities. It would be a monumental task, and so far has not been attempted.

There are other differences between the viewpoints of scientific and technical metrology. Experimental physics is concerned with differences between the model and object, i.e., the "name of the game" is find the model ambiguity. Except in so-called "absolute" measurements which are few and far between, any unit errors are ignored. Since the laws of physics are invariant under such errors this is understandable and explains, incidentally, why absolute determinations are so much more difficult. The physicist's attitude toward algorithm error is complicated. In an ideal experiment the experimental design would be such that the algorithm error is reduced to insignificance by including it in the model being tested. In practice, this ideal is approached in $4\pi$ radiation counting and in such experiments as Faller's [12] falling reflector "g" experiment where two SI definitions are combined in a conceptually simple measurement algorithm. Far more frequently, with more or less intellectual arrogance, it is assumed that the algorithm, or a vital part of it, is so complete that the response of the instrument can be calculated beyond the desired precision with negligible risk. This is the assumption that is not shared by the technical metrologist. Why can physicists usually get away with it? I believe there are three major reasons.

First and foremost, the assumption is often justified; as a rule, a great deal more study and design effort goes into a physics experiment than into the design of the usual metrological instrument and the physicist is unencumbered by questions of cost, manufacturing ease, reliability, and difficulty of operation. Under these conditions almost perfect algorithms can be conjured up. For example, an electron spectrograph is a horribly complex algorithm for measuring the kinetic energy of an electron, yet in an electrostatic machine the energy loss scale depends rigorously on the fact that electron energy depends on a potential function, and hence is set by its end points independent of the path connecting them. Second, generally speaking, the models (of atomic and molecular systems, for example) are very crude and the model ambiguities large, tending to "swamp" reasonable algorithm errors by their magnitude. It is notable that in the study of atomic properties "absolute" measurements and the use of internal calibration standards are much more popular than in solid state experiments where the models are even more crude. Third, for most experiments only a very few measurement algorithms exist, and only a few (usually one or two) stations exist, usually very similar. For example, the situation in electron scattering is typical where inelastic cross sections are the almost exclusive preserve of two or three university teams. Hence, almost by definition we remove the algorithm error by making the algorithm the standard in the metrological sense. Furthermore, in physics the algorithm errors tend to be of the unit variety and hence not vital to the questions of concern.

If there are realizations of the unit and the scale by an object whose model is so good that it is almost a prototype, and these realizations are easy to reproduce, it is sometimes possible to relieve the conditions on the algorithm so that temporal stability need not be proved (stability need not

[14] All previously quoted authors have addressed this problem; the best summary is Byerly and Lazara, op. cit.

even exist). A widespread example of this occurs in spectroscopy where the wavelength scale is provided by an iron arc spectrum obtained simultaneously with that of the object material. Note the very high practical virtue of such a system: one need only a satisfactory model (theory) of the object which serves as unit and scale. An object is usually the simpler of the object-instrument pair. With such an object in hand, any measurement algorithm (subject only to the restriction of being in statistical control) is valid. Any number of ordering operators can be employed and none of them need be studied in great detail nor completely understood.

Cases arise where the instrument algorithm is simpler or more convenient than the model. In this case, the unit is often attached to the algorithm and the roles of the instrument and object are inverted. Ionizing radiation measured in Roentgens is one well-known example where international comparisons involve the exchange of free air ionization chambers, i.e., detectors, not sources. Photometry is moving in this direction.

An interesting case study on the interface of scientific and technical metrology is in the measurement of luminous intensity.

The unit in this case was fixed by the International Organization as the candela, defined as the luminous intensity of 1 sq cm of a perfect radiator at the freezing point of platinum. Because of the rather peculiar nature of the unit, no choice exists (at this time) but to take as the model a perfect *hohlraum* at the freezing point of platinum and attempt to produce a test object as near to this as possible. This was done at NBS by adopting the 1932 platinum "black body" and then making a second stage model in the form of a computer code (based on earlier Canadian work). This second stage model contains all the parameters in the "black body" which are known to effect its deviation from a perfect *hohlraum*. The *hohlraum* theory can properly be considered sufficiently complete for the purpose at hand, but neither the computer model nor the knowledge of the material properties of the "black body" can so be considered. The output of the computer then contains the deviation from *hohlraum* and an idea of the ambiguities introduced by uncertainties in the material properties. No information is available as to the ambiguity introduced by approximation in the computer model, round off errors, etc.

The total ambiguity hence must be measured. Since the object in this case is the unit standard, it is impossible to determine the ambiguity since no hierarchically higher standard exists. Thus the model was used to calculate the temperature differential along the walls for which a tested algorithm exists and the deviation of this measured temperature deferential from that calculated is used as a measure of the model ambiguity. There remains the problem of quantifying the process, i.e., deriving the ambiguity in the value of luminous flux from the measurement of the ambiguity in wall temperatures.

Once the degree of model ambiguity is determined, attention must turn to assuring that the algorithm error was less than the uncertainty engendered by the model ambiguity. In this particular case the algorithm error was not known and a program was initiated to determine it. When it is determined to be less than the model ambiguity, the measurement will be made.

Note that, in this case, where the unit is frozen into the politics of the SI system, at some point one must either accept the total uncertainty implied by the irreducible model ambiguity or attack the political problem of disengaging the unit from this model and attaching it to a model of lower inherent ambiguity. In fact such a political solution has been achieved.

## 7. Summary

There exists a reasonably complete and coherent body of theory concerning the fundamentals of metrology. It is considerably more complex than has been expounded here, but a thoughtful application of the principles dealt with will avoid many of the problems which arise in on-going measurement systems in the field of technical or legal metrology.

This document had its origin in lecture notes for the course given by George Washington University Department of Continuing Education on Calibration Laboratory Management which I have taught for several years, and a series of seminars given over the past decade at the National Bureau of Standards. Many of the principles expounded have been developed or articulated in collaboration with Paul Pontius, Joe Cameron, Churchill Eisenhart, Chester Page, and others at NBS whose contributions to my education I can never sufficiently acknowledge.

## 8. References

[1] Ellis, Brian. *Basic concepts of measurement.* Cambridge: Cambridge University Press; 1966. 219 p.

[2] Eisenhart, Churchhill. Realistic evaluation of precision and accuracy of instrument calibration systems. J. Res. Nat. Bur. Stand. (U.S.). 67C(2); 161-187; 1963 April-June. This and other fundamental papers can be found in Precision Measurement and Calibration, NBS Spec. Publ. 300, Vol. 1 (1969 February).

[3] Volodreskii, Effect on precision of measurements of the discrepancy between the investigated object and its model, (translated from Izmeritel 'naya Technika, No. 7, pp. 18-20, July 1969) Measurement Techniques No. 7, p. 907 (1969).

[4] Carnap, Rudolph. *Philosophical foundations of physics.* New York: Basic Books; 1956. 300 p.

[5] Pontius, Paul, Mass measurement: A study of anomalies science. **190** (4212): 379-380; 1974 October.

[6] Schoonover, R. M.; Davis, R. S.; Driver, R. G.; Bower, V. E. A practical test of the air density equation in standards laboratories at differing altitudes. J. Res. Nat. Bur. Stand. (U.S.) **85** (1): 27-38, 1980 January-February.

[7] Bridgeman, Percy W. *Logic of modern physics.* New York: Macmillan Co.; (1927) 228 p.

[8] Byerly, H. C.; Lazara, V. A. Realistic foundations of measurement. philosophy of science. **40** (1): 10-28; 1973 March.

[9] Whitehead, T. N. *The design and use of instruments and accurate mechanism.* New York: Macmillan Co.; 1934. 283 p.

[10] Belanger, Brian C. Traceability: An evolving concept. ASTM Standardization News. **8** (1): 22-28; 1980 January.

[11] Pontius, P. Measurement philosophy of the pilot program for mass calibration. Nat. Bur. Stand. (U.S.) Tech Note 288; 1966 May. 39 p.

Pontius, P. E.; Cameron, J. M. Realistic uncertainties and the mass measurement process. Nat. Bur. Stand. (U.S.) Monogr. 103; 1967 August. 20 p. (Republished with new figure placement in NBS Spec. Publ. 300, Vol. **1**, 1969 February).

[12] Faller, J. E. Precision measurement of the acceleration of gravity. science. **158** (3797): 60-67; 1967 October 6.