NBSIR 74-545

Notes on the Fundamentals of Measurement and Measurement as a Production Process

Paul E. Pontius

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

September 1974

Final

Prepared for

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

NBSIR 74-545

NOTES ON THE FUNDAMENTALS OF MEASUREMENT AND MEASUREMENT AS A PRODUCTION PROCESS

Paul E. Pontius

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

September 1974

Final

Prepared for
Optical Physics Division
Institute for Basic Standards
National Bureau of Standards
Washington, D. C. 20234

U. S. DEPARTMENT OF COMMERCE, Frederick B. Dent, Secretary

NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, Director

INDEX

List of Symbols:

1.0 Introduction

At a very early age, most humans are aware of concepts such as size, distance, quantity, force, time, hot and cold. From the beginning, man has used these, and similar concepts, to understand and to shape the environment and society in which he lives. In so doing, more or less formal means of quantizing these concepts have evolved. The resulting procedures, called measurements, are now an accepted part of every day life. Because measurements are only made to support the accomplishment of a variety of tasks, the interest of the individual is more often in the task itself rather than in the measurement detail. It is only necessary that the measurement procedure which one uses, whatever they might be, will produce adequate results for the task at hand. As long as the criteria of the individual are satisfied, the procedural detail is of little consequence. The situation changes, however, as soon as the task is beyond the ability of a single individual to perform.

For the complex task, many value judgments are made by many different people, frequently at various stages of completion as well as on the completed work. Elements from many sources are assembled to make the whole. For functional reasons as well as an aid to communication, it is essential to establish acceptable limits for measurement error for all of the measurements necessary to accomplish the desired end. For each contemplated measurement, in addition to acceptable error limits, one must also assess the consequences of failure to meet the requirements as well as the benefits, if any, which might be obtained with additional measurement effort. This evaluation is not always easy to do. Often the dividing line between success and failure is not well defined. One may not be aware of failure until long after the completion of the task. In the end, the best measurement process is that which produces adequate measurement data with a minimum expenditure of measurement effort. As a consequence, the formulation of a measurement process starts with the establishment of realistic estimates of acceptable limits of error. Measurement process analysis provides realistic estimates of the process variability, which in turn provide the assurance that the results are adequate for the task at hand.

2.0 Measurement "Rules"

Philosophers of science consider two types of measurables called extensive magnitude quantities and intensive magnitude quantities [1, 2].* Extensive magnitude quantities are those for which there is a realizable addition operation, as is the case with mass and length. Addition permits replicas and subdivisions of the unit to be combined to construct any desired magnitude of the particular property. Three "rules" are sufficient for the measurement of such quantities. Intensive magnitude quantities are those which do not have a realizable addition operation. For these quantities, one must subdivide an interval between defined fixed states at, or near, the maximum and minimum of the range of interest. The measurement of intensive magnitude quantities is based on five "rules." In reality, however, it is not always clear which set of "rules" is applicable to a given measurement. This is particularly true in the case of extensive magnitude quantities which are normally discussed in context with the "three rule" scheme, but, without exception, are measured in accordance with the "five rule" scheme associated with intensive magnitude quantities.

The three "rules" for extensive magnitude quantities are:

1. The unit rule.
2. The additive rule.
3. The equality rule.

Measurement attempts to establish a one-to-one correspondence between a set of numbers and the magnitude of a particular property. In order to do this, some magnitude, however arbitrary, must be defined to correspond to a particular number. In most cases, the number is the "unit" corresponding to the numeral "1", however, it can be any convenient number. When many people are concerned with the same quantity, communications are considerably simplified if there is a uniform acceptance of a definition of the unit. For consistency of measurement, however, it is only necessary that there be defined relations between the various units in common use.

---

* The numbers in brackets refer to similarly numbered references at the end of this paper.

The addition rule, in essence, states the manner in which the multiples or subdivisions of the unit are constructed, as appropriate to a particular measurement. For example, accepting the length of a line interval as a length unit, with a simple set of dividers, one can step off intervals equal to the unit along a straight line, or subdivide the unit as appropriate. A line interval of any desired length relative to the unit can be constructed by making the "starting" termination of each added increment coincide with the "ending" termination of the preceding increment.

The equality rule states the circumstances under which one announces the magnitude of a property embodied in an object as being the same as the magnitude of the property embodied in the unit, or some multiple and subdivisions in summation. The equality in magnitude is expressed by assigning the same number to the magnitude of the unknown as has been assigned to the appropriate accessible embodiment, or extrapolation, of the unit.

The "rules" for intensive magnitude quantities are:

1. Rule for ordering, i.e., is A greater than or less than B?

2. The "zero" rule.

3. The "unit" rule.

4. Rule for subdividing the interval between "zero" and "unit."

5. The equality rule.

Temperature is an example of an intensive magnitude quantity. For a large temperature difference, our senses can easily tell us that A is hotter than B, therefore, if temperature is related to hotness, it is logical to assume that the temperature of A is higher than the temperature of B. To establish a temperature interval, one must define a "zero" state and a "unit" state, as for example, the triple point of water and the steam point[1]. Having defined a reasonably useful

---

[1] The triple point of water, a state where ice, liquid, and water vapor exist in equilibrium is approximately 0°C. The steam point, the maximum temperature where water and water vapor exist in equilibrium, is approximately 100°C.

temperature interval, one needs a temperature proportional transducer which has an output suitable for the construction of a number scale, i.e. the change in height of a mercury column due to the differential expansion of mercury and glass in a normal thermometer[1]. With the instrument scale interval terminators marked for both the "zero" state and the "unit" state, one must then define the manner in which the instrument interval is to be subdivided[2]. Finally, for the definition of equality, the temperature of the environment surrounding the thermometer is the instrument scale reading as defined by the height of the mercury column.

One normally learns about the measurement of quantities such as mass within the context of the simple three rule scheme of measurement. Everyone is familiar with the accepted mass standards such as the kilogram and the pound. The addition "rule" is merely stacking the required number of weights on the balance pan. The major area of difficulty is associated with the equality rule. For example, one classical definition of equality of mass is: "Two masses are equal if they can be interchanged on the pans of a "perfect" balance without disturbing equilibrium." This statement has no meaning for the situation in which the two objects being compared, A and B, are such that the mass of A is not equal to the mass of B.

There are at least two possible courses of action. One can obtain the prescribed equality condition by adding small auxiliary weights as appropriate, or by altering the mass of one of the objects. For either course of action, at best, one can state only that:

$$|A - B| < \varepsilon$$

where ε accounts for the minimum size weight which can be manipulated, or the minimum amount of material which can be added or removed from one of the objects, and for the operator value judgment relative to having not disturbed "the equilibrium." The value judgment, in turn, depends upon the skill and perhaps political motivation of the operator, as well as the sensitivity of the instrument and a host of other factors. For an

---

[1] Of all the human senses, vision is the most sensitive. With adequate reference points, the eye can easily detect changes on the order of a few thousandths of an inch.

[2] A common example of different ways of doing this is the Fahrenheit and the Celsius temperature scales.

ε which is small relative to the manner in which the object is to be used, the numerical value of the mass of one object may be assigned to be the same as the number value assigned to be the magnitude of mass embodied in the other. The disturbing result is that identical numbers are assigned to objects which are not equal in mass. The whole operation can be a value judgment where one takes refuge in terms such as "exact", "accurate", "right on....", and the like.

With the addition of a small weight, C, with a known mass relative to the "standard", A, one can quickly determine:

$$(A + C) > B > A$$

If, with (A+C) on the balance pan, the instrument produces an indication O_1, and with A on the balance pan, an indication of O_2, with the "unknown" B on the balance pan, one will obtain an indication O_3 such that $O_1 > O_3 > O_2$. By subdividing the instrument indication interval $(O_1 - O_2)$, one can relate the indication O_3 directly to the mass of B. Further, for any object with mass in the interval of (A+C) and A, one can obtain directly from the instrument indication a verifiable estimate of the mass of that object. With this procedure, while the process variability remains, there are no value judgments other than that associated with reading the instrument scale. These procedures clearly follow the intensive magnitude quantity rules, as is always the case where one relies on the instrument to subdivide some increment of the accessible unit. In mass measurement, one can hardly see an object with mass of one microgram, let alone the task of manipulating such an object (about the size of the smallest visible dust particle). In length measurement, a microinch is about one-twentieth of the wavelength of the light source from a helium-neon stabilized laser. With the amplification of the measurement instrument, such small increments are clearly discernible on the instrument reading scale.

All direct reading instruments, regardless of the quantity being measured, are based on the intensive magnitude quantity rules. For the most precise measurements, the instrument may subdivide some small increment, as is the case with most precise mass measurement instruments. In using the substitution weighing procedure, one compares the object with an appropriate standard and a "sensitivity" weight,

relying on the instrument to subdivide the mass of the sensitivity weight. For practical measurements, the instrument is usually designed to subdivide the maximum capacity of the instrument. The instrument also may effectively assume the role of the "unit", or standard. In some cases, the embodiment of the unit is actually changed, as is the case for most multiple lever scales where a fixed weight and a variable lever arm replaces many summations of "units" of known value. While the use of intensive quantity magnitude "rules" provide the means for practical measurement, the problems associated with providing assurance in the adequacy of the end result become complex.

3.0 The Conceptual Measurement Process

In discussing a measurement process, one must keep constantly in mind the dual nature of human activity. With our hands, we manipulate objects, operate equipment and the like, but with our minds we manipulate conceptions of those objects, or actions. We observe what happens when we take certain actions and, as long as the results confirm what we think should happen, we are satisfied with our conceptions. When the results do not agree, we must either change what we do or change our mental conceptions.

In a conceptual measurement process, repeated measurements of the same thing, or the difference between two things, agree exactly or within predictable limits. This concept applies to processes which are used to order, or sort, similar things relative to a specified magnitude of a particular property as well as to processes which assign numbers to represent the magnitude of the embodied property. The prediction limits relate to the details of the process. In order to achieve the result, the conceptual process utilizes: a model concept associated with the object or property to be measured; an algorithm concept which includes all of the instrumentation, manipulative procedures, computations, and the like, necessary to make the measurement; and a unit concept which relates to the way in which the unit is introduced into a particular measurement[1]. For most measurement processes, the results are to be passed on to other persons. In some cases, the measured object and the assigned value become the accessible unit for another measurement process. In other cases, results from measurements on selected items, or material samples, determine the disposition of large numbers of similar items, or quantities of material. In all cases, the area of doubt, or uncertainty associated with the individual measurement result is the basis for a judgment concerning the adequacy of the measurement effort.

The area of doubt, or uncertainty, associated with the result reflects the disparity between the various concepts and the performance of the measurement process in the real world. Realistic uncertainties are based

---

[1] The unit-algorithm-model concept suggested by Volodarski, Rozenberg, and Rubichev [20] is, in essence, a regrouping of the elements of a measurement method and process discussed by Eisenhart in reference [3], and more extensively in reference [4].

on the variability of the results from repeated measurements. Two sources of such variability are model ambiguities and algorithm errors. Model ambiguity, where the conceptual model differs from the actual object, will cause "perfect" measurements to disagree because the object is not behaving like the conceptual model. In contrast, algorithm errors cause measurements on "perfect" objects to disagree because the algorithm is not proper. Model ambiguity and algorithm error can be reduced to insignificance by either of two methods: one can refine the conceptual model and adjust the algorithm accordingly, or the object can be refined to fit the existing conceptual model.

To illustrate, a conceptual model of mass might be simply the property of an object. The algorithm could be simply the act of achieving an "exact balance" on a suitable instrument, the "exact balance" condition implying equality of mass. Repeated measurements of the difference between two similar objects, i.e. mass increment required for "exact balance", would produce a sequence of numbers with a characteristic variability. If one of the objects is changed, for example, a stainless steel kilogram is compared with an aluminum kilogram, the variability of the collection of repeated measurements increases drastically relative to the comparison of two stainless steel kilograms. This behavior is a clear indication that either the model concept, or the algorithm, is not correct.

There are two courses of action. One can restrict the range of density of the material to be compared, thus "outlawing" aluminum. With such action, the variability of the collection of repeated measurements would always be well behaved. In this case, the object has been refined to fit the conceptual model. However, in order to obtain meaningful mass measurements over a variety of materials, one must change the concept of the model to either the property of a "point" or the property of a body in the vacuum of space. In either case, for each object, the displacement volume, and perhaps temperature and coefficient of volumetric expansion must be known. The algorithm must be modified to account for the buoyant force of the environment in which the measurement is made. With these actions, the variability of repeated differences between stainless steel and aluminum will also return to normal.

For all measurable quantities, there must be some defined unit. This unit, however arbitrary, is accepted as having zero error by definition. The unit, in some form or other, must be introduced into each measurement. With few exceptions, the accessible unit is the output of some previous measurement process, and thereby defined by a model-algorithm combination. If alternative unit model-algorithm combinations do not produce compatible results, in addition to accepting the defined unit, all must also accept a unit model-algorithm combination[1]. The unit error, that is, the disagreement between the unit as realized and as expressed by the assigned number, is frequently beyond local control. In processes where the "unknown" is compared with an artifact reference standard, the unit error is the uncertainty from the measurement process which was used to establish the number assigned to the standard. In the case of direct-reading measurements, the unit error may include additional components relating to the instrument design.

---

[1] If for example, mass measurements based on the conservation of momentum, such as used in atomic physics, would produce results which were not consistent with the results from traditional mass measurement processes, one would have to accept one method or the other to define the measurement system. As an alternative, one might define the domains within which each of the two methods are to be used.

4.0 Variability - Two Approaches [5]

As a point of departure, all measurement processes are, directly or indirectly, comparative operations. Even the most simple concept of such a measurement contains certain implicit assumptions:

(a) a constancy in the basis for the ordering or comparing; and

(b) a stability in the equipment, procedures, operator and the like which are used to make the easurement;

(c) a stability in the object, effect or property being observed.

Quantitative ordering implies an invariant basis for the ordering, thus a long term constancy in a standard unit and a stability in the realization of a standard unit, is necessary. In a similar manner, the property to be measured must also be stable. If a measurement process detects a difference between two things, it is expected that repeated measures of that difference should agree reasonably well. In the absence of severe external influence, one does not expect things to change rapidly.

There is a difference between stability and constancy in context with the above. Repeated measurements over time can exhibit a random like variability about a constant value, or about a time dependent value. In either case, if the results are not erratic (with no unexpected changes), the process is considered to be stable. The objects being compared may have constant values, or may be changing at a uniform rate, or may be changing at different rates. For continuity, time dependent terms must be included in quantitative descriptors for both objects being compared. Stable changes with time can be extrapolated in the same manner that one "extrapolates" a constant value over time. The extrapolations can be verified whenever desired by making additional measurements. Constancy, then, merely means that the coefficients of time dependent terms are essentially zero. This is not to say that features such as constancy are not desirable for certain usage, but only that such features are not necessary restrictions on the ability to make good and useful measurements.

Two quantitative descriptors are used to describe the process variability, and ultimately, to establish the bounds for the limit of error. A given measurement process is continually affected by perturbations from a variety of sources. The random like variability of the collection of repeated measurements is a result of these perturbations. One descriptor, designated *random error*, includes effects from both cyclic perturbations such as might be associated with the environment and variability associated with operating procedures. The random variability, expressed as a standard deviation, will for a process in control (see [3]) imply a low probability that the range of variability in the collection will exceed certain bounds. The second descriptor, designated *systematic error*, S.E., includes the use of constants which are in error as well as discrepancies from certain operational techniques. The S.E., expressed as a single number, is an estimate of the offset of the measurement result from some defined process average. These two descriptors, called the process performance parameters, are factors in assessing the worth of a result relative to a particular requirement.

The random error estimate reflects the effects of cyclic perturbations which are constantly changing whether the process is being used or not. These effects can be grouped into two categories; short term effects which vary through one or more cycles in the course of a single measurement or measurements made over a short time interval, and long term effects in which the period of the effect is at least as long as the time required for a given sequence of measurements.

A second category of short term effects are those which are instantaneous, or step-like, in nature. In many cases, "shocks" on the instrument, or variations in manner in which various objects are introduced to the instrument, cause changes in the instrument configuration which affect the instrument indication. The effects appear as minute, and sometimes not so minute, instrument reading scale shifts[1]. For example, the manner in which a large weight is placed on a

---

[1] In this case one is not concerned with transient changes as for example when a meter needle "jumps" and immediately returns to the original reading. The operator can be instructed to ignore such changes. Neither is one concerned with slow "drifts" which occur in the course of a single measurement. In most cases, these can be accounted for in the algorithm.

platform scale may effectively shift the scale several readable units. Off center loading on either a balance or a large weighing instrument may cause a change in lever ratio which has the same effect on the reading scale. The variability from these sources may be random in both magnitude and direction.

In terms of measurement process performance, the *within-group variability* expressed as a standard deviation[1], σ_W, reflects the combined short term effects both cyclic and step. In many cases, σ_W represents an optimum process performance. The within-group variability of the measurement process is the most familiar process parameter as it is easily demonstrated in a repeated sequence of measurements of the same thing in a very short time interval. Practically all important measurements are repeated several times. The magnitude of the within-group variability is generally established by the degree to which certain types of perturbations are controlled and by factors such as the operator skills, quality of the instrument, and attention to detail procedure. In most cases one cannot identify sources of perturbations which contribute to within-group variability. Process improvement in terms of reducing σ_W is obtained perhaps more frequently by trial and error than by design. The adequacy of a given process relative to a particular requirement is often judged on the basis of the within-group variability. Such a judgment, however, may be erroneous.

The total variability is the variability of a long sequence of data which reflects the effects of all possible perturbations. Repeating a given measurement over a time interval sufficiently long to reflect the influence of all possible perturbations establishes a total process standard deviation, σ_T, which reflects both the short term and the long term random variability[2].

---

[1] Standard deviation is used only as an index of variability, with no restriction on the distribution intended.

[2] The *total process variability*, σ_T, can be thought of as the sum of the variabilities of all of the perturbations that affect the process, that is, $\sigma_T^2=\sigma_1^2+\sigma_2^2+.....\sigma_n^2$. For one class of perturbation with variabilities σ_1 to σ_m, which are those with very short periods and with nearly equal amplitudes, it may not be possible to identify the individual perturbations. The variability from these perturbations combine to form a threshold variability σ_w. Other perturbations, with variabilities σ_{m+1} to σ_n, may be identifiable if the magnitudes are sufficiently large. These effects combine to form a *between time component of variability* σ_β. The total variability is then $\sigma_T^2=\sigma_W^2+\sigma_\beta^2$.

With a sufficiently long sequence of data, one should be able to identify the sources of the largest perturbation through supplemental measurements and correlation studies. Having identified the source of the largest perturbation, the magnitude of its effect on the measurement can be minimized, with a consequent reduction in the magnitude of σ_T. Frequently one is tempted to idealize the process in order to reduce the total variability, that is, to establish a carefully controlled environment and use only selected artifacts. Such actions are self-defeating in terms of understanding the measurement process. A more appropriate action, provided one has sufficient motivation and resources, is to modify the process to account for the variability associated with all the perturbations that can be identified. Since a large perturbation will "offset" the single value from the process average, these effects, if uncorrected, are frequently called Systematic Errors.[1]

There are several different classes of Systematic Errors. Perhaps the most familiar class of S.E. is associated with instrument reading scale offset. Such S.E.'s are not present in comparative measurements provided that the instrument indication can be related to the measurement unit, and provided that the instrument response is reasonably linear over the range of difference which must be measured. A second class of S.E.'s is associated with supplemental data such as barometric pressure, temperature and relative humidity measurements which are in turn combined to determine air density, index of refraction and the like. Each of the supplemental measurements is, in essence, a separate distinct measurement process with both random variability and systematic effects. The random variability of the supplemental measurements is, of course, reflected in the total process variability. The S.E.'s associated with supplemental data must be carefully considered.

One action, which is rarely practical, would be to "randomize" the S.E. by using different instruments, operators, environmental or other factors., in which case the variation from these sources becomes part of

---

[1] If the large perturbation is truly cyclic, corrective action will frequently reduce the magnitude of σ_T with only minor change in the process average.

the random error.[1] A more practical procedure is to evaluate the S.E. associated with an instrument (or other factor) by direct experiment. When the change in response, such as for example that introduced by a temperature error of 0.1 degree, is a small fraction of the standard deviation of the process, a rather large number of measurements is required to establish the effect with a reasonable degree of assurance. Bearing in mind that an average of n measurements has a standard deviation of $1/\sqrt{n}$ times that of the original measurements, in order to determine an effect of size one standard deviation with an uncertainty (3 standard deviations) of half of its size one would need about 36 measurements. (If one relaxes the uncertainty requirement for the average to a value equal to the standard deviation of the process, then 9 measurements would be required.)

With evidence that the individual supplementary measurements are satisfactory, the next concern is the manner in which the supplementary data are combined and used to adjust the observed data. For example, having adjusted the data for thermal expansion, one would not expect a collection of values over time to correlate with the temperature measurements for each individual value in the collection. A collection of values from repeated measurements should be tested for dependence on each of the supplementary measurements, and their various combinations, as appropriate. If dependence is indicated, either the supplementary measurement is not being made at the appropriate location, or the manner in which the supplementary measurements are combined does not describe the effect that is actually occurring. Corrective action is necessary. No dependence does not necessarily indicate that there are no S.E.'s present, but only that for the supplementary measurements which have been made, the magnitude of the combined S.E.'s is not large relative to the total standard deviation of the process.

---

[1] For example, in the simple case of measuring the width of a piece of paper with a rule, a practice which permits setting the "0" of the rule to one edge of the paper will introduce the bias of the operator in setting "0" as well as the bias associated with the location of the printed scale on the rule. By placing the rule at random on the paper, both terminators of the interval are estimated, thus eliminating both sources of bias.

There may be long term systematic error effects from sources not associated with the current supplemental measurements. It is relatively easy to demonstrate the presence or absence of such effects, but it may be difficult to reduce their magnitudes. If one has available a collection of values over a long time span, one can compare the standard deviation as computed for small numbers of sequential values over short time spans with the standard deviation of the total collection.[1] While reasonable agreement is expected, frequently such is not the case. If the magnitude of the effect is sufficiently large, the collection of values may indicate grouping, with the group means appearing as random variability about the process average. If the distribution of the collection of values appears to be bi-modal, one should look for a large long term cyclic effect. Until the source of such variability is identified, and appropriate action taken to modify the process, the total standard deviation must be used as the descriptor of the random variability of the process.

The reason for making measurements is to assign numbers representing the properties of interest in such a way that the numbers will be useful to others. The reason for characterizing the measurement process is to assign meaningful error bounds, or uncertainties, to the numbers representing the properties. The magnitude of the uncertainty is established by the error bounds of the local measurement process and the error of the accessible unit. In most mass and length measurements, access to the unit is through an artifact which has been assigned a length, or mass, value by another measurement process. In the case of mass, for example, the international prototype kilogram is defined to have zero unit error. With a process operating in a state

---

[1] The use of comparison designs, such as described in reference [6], facilitates this type of analysis. The within group variability, σ_W, is computed for the prescribed sequence of measurements. Each measurement sequence includes in effect a "check standard" which is measured over and over again with similar measurement. The total standard deviation is computed for the collection of values for the "check standard." The inequality $\sigma_T > K\sigma_W$ is taken as evidence of the existence of a long term systematic effect, perhaps as yet unidentified. The term K in this relation accounts for the fact that the "reported" value of the "check standard" from the observations required by the design is not the result of a "single measurement" but, in effect, is the weighted average of "n" measurements in the design sequence, while σ_w is the standard deviation of a "single measurement."

of control, that is, with no known systematic effects unaccounted for, and with the international prototype kilogram to introduce the unit, the uncertainty is only a function of the process standard deviation, either σ_W or σ_T.[1]

The reported value may be the result of a single measurement, or the average of n independent measurements. Both results are only estimates of an expected long term process average, and, as a consequence, the "reported" result is always offset from the true process average by some amount. This offset, as determined by the process standard deviation, can be either plus or minus. If an object, as measured above, and its assigned value are used to provide an accessible unit to another process, this offset is a systematic error associated with the unit. That is, the results from the following process, which uses the object as the accessible unit, may be biased by the offset or systematic error of the first process. The uncertainty of the result from the second process must include this systematic error associated with the unit in combination with the random variability of the second process. Whenever a fixed value is assigned to the accessible unit, a S.E. component of magnitude equal to the uncertainty of the assigned value is introduced. For all well characterized measurement processes operating in a state of control, the S.E. associated with the accessible unit should be the only S.E. component in the uncertainty for the result, all other identifiable S.E.'s having been accounted for in the process.

A measurement process is said to be operating in a state of control when: (1) sequences of independent measurements support a single valued limiting mean; (2) the collection of values is free from obvious trends or grouping; and (3) each new measurement verifies the validity of prediction limits based on historical

---

[1] It is important to note that only the assigned value is assumed exact. Model ambiguities associated with the accessible realization of unit are reflected in σ_T. In the case of the International Prototype Kilogram, variability associated with stability in mass is not well known. Equipment with sufficient precision to evaluate this stability has only recently been developed [7]. For length, the stability of the Iodine stabilized laser is on the order of 1 part in 10^{12}, thus in all practical measurements, components of variability from this source are essentially zero [8].

performance data.[1] Most processes can be made to operate in a state of control. A process is said to be well characterized when the process performance is independent of the objects which have to be measured and the environments in which the measurements must be made. While this is a desirable goal, particularly with respect to understanding the process, it is not always achievable. To obtain such performance, the unit-model-algorithm combination must be refined and generalized so as to be applicable to each set of conditions. Sources of S.E. which can easily be identified can be taken into account. However, the condition that the within-group variability, σ_W, be identical to the total variability, σ_T, for all conditions of measurement is seldom achievable. As a practical matter, for the well characterized process one must be satisfied with realistic estimates of σ_T. Refinements to reduce the magnitude of σ_T are both costly and time consuming.

Fortunately, most measurement processes for a given property are similar so that the characterization and documentation of a typical process over the range of objects and environments in which the measurements are usually made substantially shorten the time required for characterizing other processes. As a practical matter, few can afford the time and effort to identify all perturbations related to the between-group variability of S.E. components as previously discussed. For each practical measurement it is only necessary that the uncertainty of the result be adequate for its intended usage. In the end, the uncertainty associated with a sequence of operations defined to be a

---

[1] The restrictions for control listed here apply to a variety of measurement situations where a "repetition" is a repeated measurement after a relatively long time interval. These measurement situations usually involve the properties (physical, optical, electrical, etc.) of objects or systems. A more general condition for being "in control" is that the process behaves as the output of a probabilistic model. Situations involving correlated measurements may be regarded as "in control" if the output is predictable in the sense that it can be considered as producing random variables from the assumed mathematical model. In the latter case, the measurements are usually concerned with characterizing a time dependent phenomenon such as the output of an oscillator. A detailed discussion of such a situation is given in reference [22].

measurement is determined in part by the larger of σ_W and σ_T and by the S.E. components associated with the unit. The uncertainty statement must also include the S.E.'s which are not accounted for in the measurement process for reasons of convenience.[1]

For some measurements, the difficulties associated with making the measurement, the time involved, and the cost, make it impossible to consider sequences of repeated measurements under varying conditions. In these cases one usually relies on an "error budget" to establish an estimate of the uncertainty of the reported results. The typical error budget assumes the algorithm-model concept of the measurement to be exact. A term by term analysis establishes the effect of the variability of each term with respect to the announced result. For each term, a listing is made of all known sources of error, with an estimate (usually based on theoretical analysis and "engineering judgment") of the magnitude of the expected variability. The estimates are usually combined in some appropriate manner to obtain the error bounds for the final result. While the error budget analysis is helpful in many kinds of measurement, it is not unusual to find measurements of the same thing which disagree in excess of the expected limits of error based on extensive error budget analyses. In such cases, the disagreement is strong evidence that the algorithm-model concept used in one or both of the measurements does not reflect the real life situation.

The choice between the two approaches to measurement variability depends upon the detail of the algorithm. In both cases, one is trying to establish a realistic error limit for the process result. The reliance on an error budget does not negate the need for experimental verification of the appropriateness of the error limit.

The instrument included in the algorithm is, in essence, an amplifier, with input signal being proportional to the property of interest, and including "noise" from perturbations which affect the process

---

[1] In many cases acceptable limits relative to a particular usage are large with respect to measurement process capabilities. In the interest of conserving measurement effort, detailed corrections for S.E.'s are frequently ignored. When such is the case, the effect of the ignored S.E.'s must be included in the uncertainty statement.

performance. As the "gain" is increased in order to detect smaller and smaller changes in the property of interest, the "noise" is also amplified. For the most precise processes, the gain is adjusted so that the noise is clearly observable. For these processes, sequences of repeated measurements produce sequences of non-identical numbers which reflect the process variability. In the case of many practical measurement processes, the gain is adjusted so that noise is not observable. For these processes, sequences of repeated measurements always produce the same number, and, as a consequence, the error limit must be established by some other means. In most cases, one can purposely introduce changes of known magnitude into the measurement process to verify the minimum incremental change which the process can detect.

5.0 Measurement as a Production Process

The concept of a measurement system requires that values assigned to represent certain characteristics of objects be reasonably unique and repeatable over time and changes in location. It is expectd that sequences of measurements of the same thing made at various times and at different locations show evidence of convergence to the same limiting mean. Uncertainty statements are, in essence, predictors of the degree to which such agreement can be attained.[1] Failure to agree within uncertainty limits is an indication that the two processes are fundamentally different, or that the uncertainty statement does not adequately describe the error bounds. For a practical measurement, the measurement algorithm, or the mathematical model of the measurement process, cannot possibly reflect all of the sources of variability. The instrument or comparator cannot differentiate between a real change and all of the perturbations which change the indication in the same manner as a change in the object. None the less, it is important to know the bounds for the variability which occurs in the course of making measurements. Redundancy, either by repeated measurements or incorporated in a particular measurement process, provides a means for assessing this variability.

In order to illustrate the nature of a measurement process, consider first the collection of simulated

---

[1] The word "closure is used in the following sense: One does not expect the results for the same measurement by each of several processes to be identical. On the other hand, if the property being measured is stable, one would expect that collections of measurements by each process would support the same limiting mean and, for each process, one would expect the uncertainty limits centered on the result to encompass the limiting mean. While one may not know the limiting mean value, under these conditions the results from several processes can be compared by centering the appropriate uncertainty limits on the respective results. If the areas so defined close, or "overlap", the results are considered to be in agreement within the capabilities of the processes involved.

measurement in figure 1.[1] The data shown reflects the effects of variability from both cyclic and step sources over time. For the 300 measurements shown, the data has the appearance of coming from a reasonably well behaved measurement process. The measurements tend to cluster around the central line--the process average or limiting mean. Confidence that the process has a single limiting mean is strengthened as the length of the record is increased. With process performance as shown, a predictive statement concerning the next, but as yet un-made measurement, can be considered.

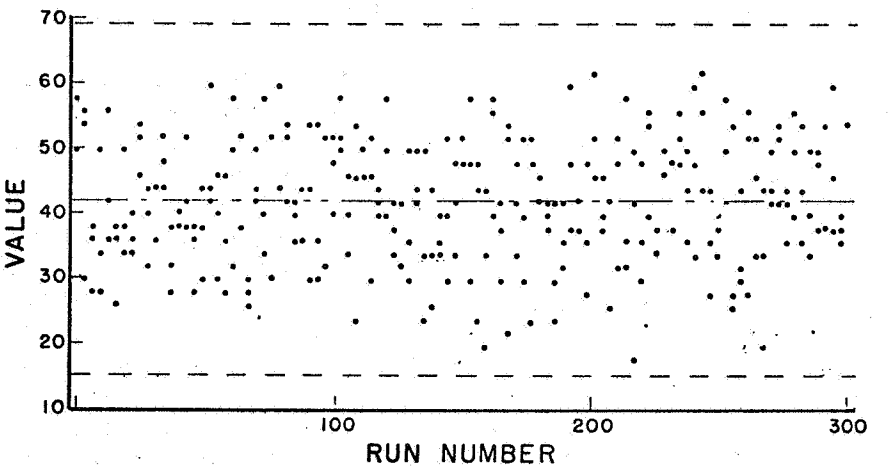


FIGURE 1

It seems clear that the predictive statement cannot be exact but will have to allow for the scatter of the results. The goal is a statement with respect to a new measurement, a measurement that is independent of all those that have gone before. Such a statement should

---

[1] For the purpose of illustration, the data shown is from a simulated measurement process. The "process" includes a simulation of both cyclic and step variability. Four sinusoidal functions, with amplitudes a1, a2, a3 and a4, and with periods differing by a factor of 10, simulate the cyclic variability. A random choice between +b, 0 and -b simulates the step variability. The value shown is the sum of four cyclic functions, sampled at random times, and the step function, with appropriate "scale shift" adjustment. For the "improved" process, as shown in figure 3, all the amplitudes of the cyclic terms and the step function are equal. For the "unimproved" process, shown in figure 1, the amplitude of one of the cyclic functions has been doubled.

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 can be considered as random drawings from the same probability distribution. These conditions will be satisfied if the collection of points is independent, that is, free of patterns, trends and so forth; and provided it 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.

From a study of a sequence of such independent measurements, control chart techniques can be used to set up limits within which the next value should lie. For an extremely long sequence, limits can be marked off on either side of the mean so that some suitable fraction, say 99 percent, of the observations are within the interval defined. The probability is also 99 percent that the limiting mean will be within the interval established by centering these same limits on any observation chosen at random. This will be true of the next observation as well, provided it is an independent measurement from the same process. The probability statement attaches to the sequence of such statements. For each individual new observation, the statement is either true or false but in the long run 99 percent of such statements will be true.

Assuming that the limits are based on large numbers of observations, for a process operating in a state of control, very nearly the intended percentage of all such limit bands, centered on the observed values, would in fact overlap the mean. This will not be true for points in the area outside of the control limits. This is expected in only 1 percent of the cases. More frequent occurrence is a clear indication of either loss of control or that the limits were not properly set.

If, over the sequence of the 300 measurements shown, the variability of the collection reflects the maximum excursions of each parameter, the s.d. of the collection is the total s.d. of the process, σ_T, computed in the usual manner:

$$\sigma_T = \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{(n-1)}}$$

where

y_i = individual result

$\bar{y} = \frac{\sum_1^n y_i}{n}$

n = no. of y's in the collection

The $3\sigma_T$ limits shown in figure 1 appear to be adequate bounds for the "process" variability.[1] One would surely expect the next "single" measurement result to be within these prescribed limits. Further, if the next measurement was defined to be the average of n independent "single" measurements from this process, one would expect this average to agree with the average of the collection within $(3\sigma_T/\sqrt{n})$. Without knowledge of independent parameters which are known to be proportional to the magnitude of each source of variability, there is no way to further analyse this data, and the random component of the uncertainty of the result, however defined, would be a function of σ_T. If this performance of the process, according to this particular algorithm, is adequate for the intended use, there would be no reason to change the algorithm.

[1] Unpublished material furnished by Eisenhart, indicates that 3 s.d. limits are appropriate for .99 probability for all distributions. For $\beta_2 > 5.6$ the probability is slightly less than 0.99; for $\beta_2 < 5.6$, a bit more than 0.99. For distributions with $\beta_2 < 3$, the 3 s.d. limits are somewhat pessimistic, however, the simplicity of using 3 s.d. limits far outweighs the complexity of determining the appropriate distribution. For the simulation shown in figure 1, $\beta_2 \approx 2.58$, and for figure 3, $\beta_2 \approx 2.36$. With sufficient measurement effort, it may be possible to achieve an algorithm for a particular measurement process which will produce a collection of values with distribution approaching $\beta_2 \approx 3$. For a collection of differences between two nichrome kilograms in excess of 300 and over approximately a 10-year period, $\beta_2 \approx 3.02$. In this case, considerable effort was made to assure that the algorithm accounted for the effects of all known sources of variability. (β_2 is the usual kurtosis parameter, having value 3.0 for a normal distribution.)

In this process simulation, there are identifiable parameters which are proportional to the effects of sources contributing to the process variability.[1] Recording the parameter values along with each measurement result permits the use of correlation studies to further evaluate the process. In figure 2, the parameter for each source of variability is plotted

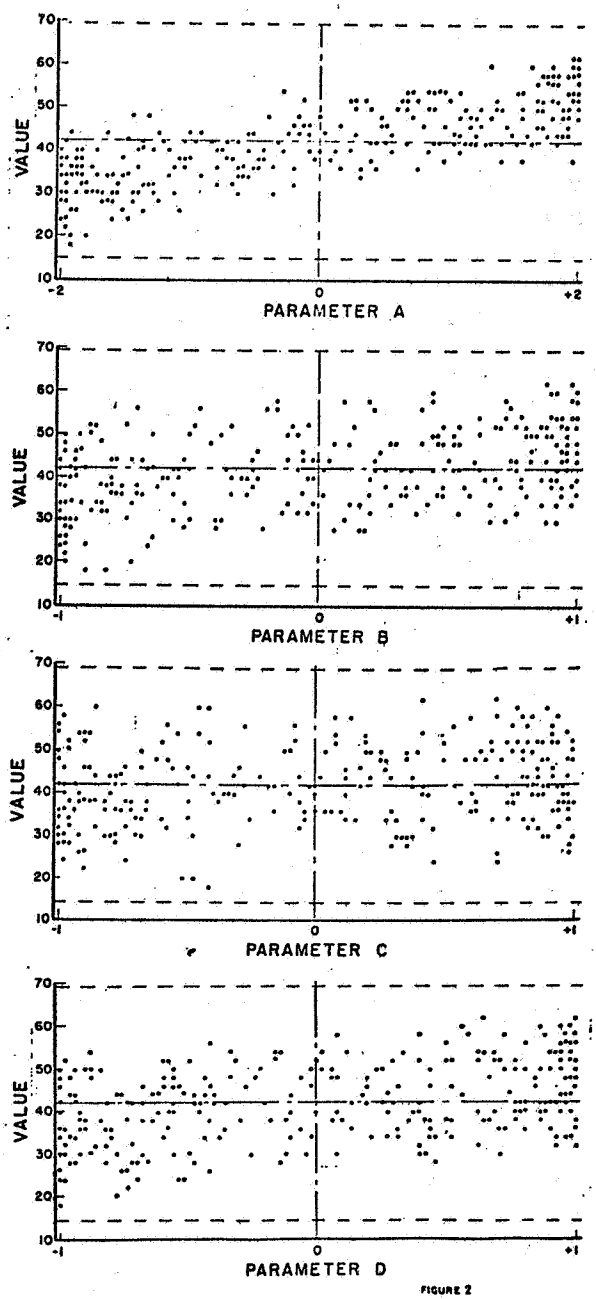


FIGURE 2

---

[1] In the simulated process, the parameters are the individual values of the sinusoidal functions which, in summation, define in part the resulting "measurement".

against the appropriate measurement result. For parameters B, C and D, there is little evidence of correlation. While the variability of these parameters contributes to the process variability, one cannot differentiate between their respective contributions. Clearly, there is a correlation with parameter A. This correlation indicates that a "between time" variability associated with parameter A is influencing the measurement results. The effect is systematic, that is, the result is high when the parameter value is high, and vice versa. It should be noted, however, that in spite of the existence of the systematic effect, the initial $3\sigma_T$ limit is still an appropriate bound for the process variability. On the other hand, having identified the source of variability, "corrective" action can be taken to reduce the magnitude of this particular systematic effect, with a resulting decrease in the "process" total s.d., as shown in figure 3. The results from the new algorithm are now free from identifiable sources of systematic variability.[1]

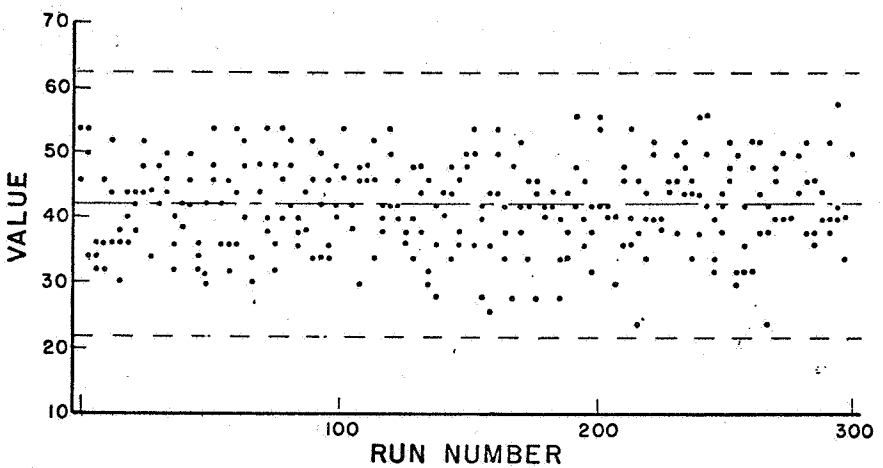


FIGURE 3

The random component of the uncertainty statement, $3\sigma_T$, relates to the limiting mean of the process. It is not only a quantitative statement with respect to a

---

[1] The corrective action was to set all coefficients of the cyclic terms to the same value. (This action is, in essence, changing the algorithm to correct for the identified source of variability.) Under this condition, the term by term correlation studies did not show strong evidence of correlations. The plots were similar to those from parameters B, C and D in figure 2.

"single" measurement and the process limiting means, but also with respect to the expected agreement between any two "single" measurements. If the process limiting mean is within $3\sigma_T$ of any "single" measurement, then for any two measurements, the $3\sigma_T$ limits, centered on the individual values, should overlap most of the time. If a quantitative uncertainty statement is intended, one must accept the process limiting mean as the "best estimate" of the process output.[1] It is seldom mean for each measurement which has to be made. One must look for other techniques if one intends to make an operationally verifiable uncertainty statement. Two such techniques involve redundancy and a "check standard" concept. When used in combination, these techniques will provide the desired data.

The use of redundancy in measurement is not new. Most important measurements are repeated several times, the announced value being the average of the results obtained. Some assessment of the process performance is made by computing the standard deviation of the collection of these repeated values. This standard deviation, however, reflects the process performance over a relatively short time, and as such is not necessarily a valid estimate of σ_T. By defining the announced value to be the average of n measurements, for a process in a reasonable state of control, the standard deviation of the n single measurements should also be well behaved. These standard deviations for each sequence of similar measurements can be combined to obtain a long term accepted within process standard deviation, σ_w. First estimates of σ_w may be based on only a few measurement sequences, however, in time the collection will provide a stable value for σ_w, which is a characteristic of the particular process, or algorithm. For each sequence of measurements, the computed within s.d. can be compared with the accepted σ_w to verify that the process is performing as expected.

The "check standard" concept provides a means to accumulate, usually at little cost in terms of measurement effort, a collection of data which will

---

[1] If the magnitude of the property being measured is temporally stable, the estimate of the limiting mean is the average of available collections of values. A somewhat different approach is used when the magnitude of the property is changing with time. This approach will be discussed later.

establish the accepted total standard deviation, σ_T. One form of a "check standard" is the inclusion of an extra object in each sequence of measurements used to establish the announced value. This object is chosen to be similar in all respects to other objects which must be measured. With a particular "check standard" in each similar measurement, the collection of values obtained will, in time, be similar to the collection shown in figures 1 or 3. For a long sequence of values, the standard deviation of the collection is a measure of σ_T. The total s.d., σ_T, determined in this manner is appropriate for use in expressing the random variability of the process. One could, for example, interchange the "check standard" with one of the "unknowns" and, after the same number of repeated measurements on the "unknown", the characteristics of the collection of values obtained would be the same as those which were from the original collection of values for the "check standard[1]."

In some cases, the collection of repeated measurements will indicate that the magnitude of the quantity of interest is not constant, but slowly increasing or decreasing. In this situation, the limiting mean or average value of the collection is not the best estimate of current, or future, values. One must predict an appropriate value for a particular time, together with the uncertainty of the predicted value. The prediction must be valid over some reasonable time increment in order to be useful. This must be done with care. A typical error in judgment is declaring the magnitude of the property to be changing without knowledge of either σ_w or σ_T. Usually significant rates of change are readily apparent in the collection

---

[1] It should be noted that the initial assignment to the "check standard" is of little importance. The value assignment which is made in accordance with the procedures of the algorithm, can be easily changed, one way or another, provided that the "check standard" is stable. The variability of the collection of values, however, is of major importance. The long term variability of the result is a characteristic of the particular process. Confidence in the total s.d., σ_T, increases as the collection of values for the "check standard" increases. The accepted value for σ_T can be used to assess the process performance for each sequence of measurements relative to the value obtained for the "check standard" in that measurement.

of historical data.[1] As the data base increases, confidence in the ability to predict both the value and the rate of change will increase.

To illustrate, in figure 4(a), a prediction line as a function of time has been fitted to three typical measured values. The predicted value is that expected one time increment beyond the existing data base. With a reasonable estimate of the process standard deviation, the uncertainty of the predicted value can be computed from the formula [23]:

$$3\sigma_\beta = 3\sigma\sqrt{\frac{1}{n} + \frac{(t-\bar{t})^2}{\Sigma(t_i-\bar{t})^2}}$$

where n = the number of points in the collection

t = time/date of the prediction

$\bar{t}$ = average time/date (location of the centroid of time span covered by the measurement, that is $\bar{t} = \Sigma t_i/n$

t_i = time/date associated with each of the n values

σ = process standard deviation (s.d. about the fitted line)

σ_β = s.d of predicted value at time t.

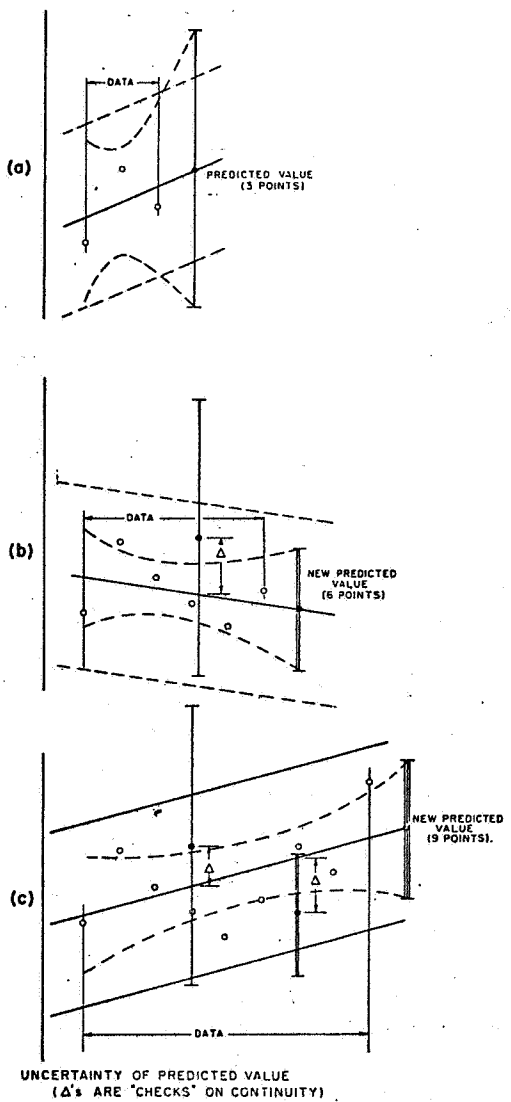
The uncertainty of the predicted value is large because the extrapolation interval is large relative to the time span of the data base.

With three additional data points, as shown in figure 4(b), a new prediction line is established. The uncertainty of the new predicted value is somewhat smaller since the extrapolation interval is a small fraction of the new data base. The uncertainty of the previous predicted value overlaps the prediction line as expected. Again, with three more data points, a new

---

[1] In addition to the obvious evidence which may be apparent in a control chart, there are several other ways to verify the existence of time dependent changes. If the s.d. about a "fitted" line is clearly smaller than the s.d. about the average value, there is a time dependent change occurring. Normally, in the process of fitting a set of data to a linear function of time, x, by the equation $y = ax + b$, the s.d. of the rate of change coefficient a, can be determined along with the value of a. The significance of the rate of change, a, can be determined relative to the s.d. of a.

prediction line is shown in figure 4(c). Again, for both of the previous predicted values, the uncertainty overlaps the new prediction line. As the historical data base increased, the uncertainty of the prediction over a relatively small time interval approaches the "uncertainty of the mean", $3\sigma_T(1/\sqrt{n})$. At this point, the rate of change should be well known.



UNCERTAINTY OF PREDICTED VALUE
(Δ's ARE "CHECKS" ON CONTINUITY)

FIGURE 4

6.0 The Unit

With the process parameters, σ_W and σ_T, known, to complete the measurement, that is, to assign a number which relates a property of an "unknown" to some defined "unit", a representation of the "unit" must be included in the measurement process[1]. The process compares the "unknown" with this accessible "unit". Error bounds for agreement, or "closure", within a system of similar measurements must account for the "unit" error. Difference measurements are determined by the response of the process to the real magnitudes of the properties involved in the "unit" and the "unknown" while number assignment "assumes" the value assigned to the "unit" is correct.

Unit errors, unlike process performance parameters, cannot be operationally identified with a single measurement process, or with several measurement processes which use the same accessible "unit" and the same "equality rule", or measurement algorithm. Accessible units are a part of all practical measurement processes, few of which have access to the same realization of the unit. For a process operating in a state of control, the uncertainty is the sum of the limits for random error and the unit error, the unit error being a systematic error, or bias, beyond the control of the local process[2].

For each measurement process, the value assigned to the accessible unit is the product of one or more prior measurements which, in a geneological sense, extend back to some accepted definition of one or more units. While the accepted units are usually the units of the SI system or units which have exact relations to the SI

---

[1] A general interpretation of "unit" includes artifact replicas or other realizations of the units of the SI system, replicas of production objects, standard reference materials, and the like [9, 10].

[2] The condition for operating in a state of control were discussed earlier. A process is said to be operating in a state of control when, for each defined sequence of measurements, the computed s.d. is in agreement with the long term accepted σ_W, and the value of the "check standard" obtained is in agreement with the long term accepted value within the limits established by σ_T.

units, it must be emphasized that an SI unit definition is only the specification of the phenomenon, or artifact, which represents the unit [11]. There is no attempt to specify the manner in which the units are to be made accessible to the many measurement processes which must use them. The phenomena have been selected on the basis of temporal stability, and are subject to change as appropriate and agreed upon by the General Conference of Weights and Measures.

With the exception of mass, the only remaining artifact standard, the recommended realizations of the units are quite complex, and, once realized, are not in a form easily adaptable to practical measurement. For simplicity, mass will be used to illustrate unit error, but it must be remembered that with other units, several basically different processes, or algorithms, must be studied in the same manner. In the case of length, for example, one must go from the defining vacuum wavelength to the wavelength of the available light source in the environment in which the measurement is to be made, to the length of the objects such as gage blocks, to the length of scales such as meter bars, to dynamic "fringe counting" interferometers, and then to a wide variety of practical measuring equipment including both instruments and reference shapes.

All mass measurement algorithms, from those used with the defining unit to the most crude measurements, are based on the same principle, differing only in the degree of refinement. This is both convenient and practical; convenient because of the inherent simplicity of the process, and practical because of the availability of a wide variety of weighing equipment. The algorithm is based on relating the gravitational force acting on the "unknown" to the gravitational force acting on some representation of a suitable unit. Algorithms based on other principles are possible, however, suitable equipment has not been developed sufficiently to compete with the accepted algorithm[1]. Starting with a comparative measurement process, the realization of a particular algorithm-model concept, and the unit as embodied in the defining artifact, unit error can be illustrated by the manner in which values are assigned to replicas of the defining artifact.

---

[1] Since the result is defined by the particular algorithm, a considerable degree of confidence is gained when measurement of the same thing by two basically different algorithms are in agreement. Such a case in length measurements will be discussed later.

Assuming that the replicas are similar in all respects to the artifact, and that the mass differences between the replicas and the artifact are within the "on scale" capacity of the instrument, for a measurement process operating in a state of control, the results, y_i, of a large number, n, of difference measurements between the artifact and any one replica might be distributed as shown in figure 5(a).[1] Studies of the collection, y_i, such as correlation studies mentioned before, may result in algorithm improvements so that for the next series of measurements, the results, y_j, might be distributed as shown by (b). Now, being satisfied with the process performance, one can define the announced result to be the average of n "single" measurements. The distribution of the results, defined in this manner might be as shown by (c). Note, however, that the measurement effort to achieve this distribution requires m x n "single" measurements. In all cases, if m is sufficiently large, the limiting mean will not be changed[2].

The limiting mean, or "0" in figure 5, compares to some number relating to the difference in mass between the defining artifact and the replica. For a process operating in a state of control, the "best estimate" of the limiting mean is an average of the available measurement results:

$$\bar{y} = \frac{\sum_{1}^{n} y_i}{n}$$

The uncertainty of the limiting mean is:

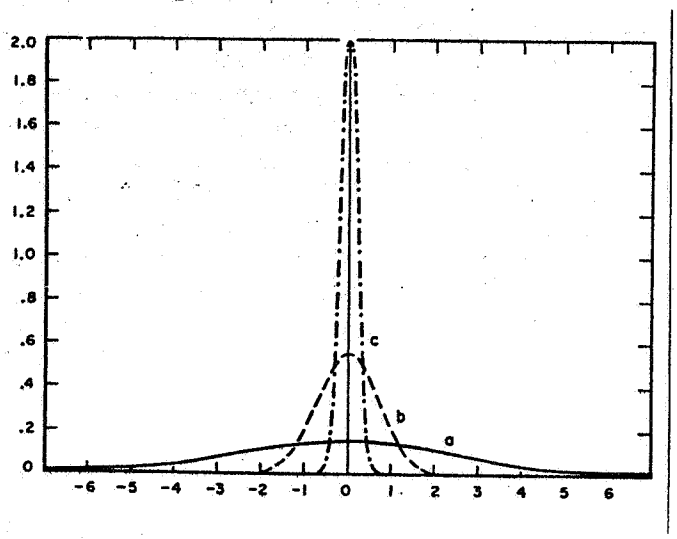
$$\mathrm{Unc}(\bar{y}) = 3\sigma_T(1/\sqrt{n})$$

---

[1] Normal distributions are shown here for convenience. The only requirement on the distribution is that it be reasonably symmetric.

[2] For a given set of conditions, the limiting mean reflects the "average" of all of the cyclic perturbation. Gross changes from "average" for perturbation of major influence on the process, if uncorrected in the algorithm, will obviously shift the limiting mean.

The expected shift in the computed average by virtue of adding one more measured value to the collection is on the order of $\pm 3\sigma_T/(n+1)$.



Distribution curves for three distributions, a, b and c, with the same mean and with distributions σ=2.5 for a, $\sigma=2.5/\sqrt{12}=.72$ for b, and $\sigma=.72/\sqrt{12}=.21$ for c.

FIGURE 5

The value reassigned to certain defining artifacts, such as the international prototype kilogram, is "exact" by definition. For replicas of such artifacts, the assigned value is an estimate of some process limiting mean and, as such, can never be "exact." The announced value for the "unknown" replica is the sum of two numbers, the number assigned to the defining artifact and the number assigned to the measured difference. One could use the result from a "single" measurement by the crude process, as shown by distribution (a), or from the improved process, as shown by (b), or an average of n "single" measurements, as shown by (c). Most of the time, each value, from whatever process, will not lie farther away from the limiting mean than $3\sigma_T$, or $3\sigma_T/\sqrt{n}$, as appropriate where σ_T is the total s.d. of a single measurement. Whether the value used is too big or too small cannot be determined, but, given sufficient time, it can be demonstrated that the limits are appropriate. A collection of replicas, "calibrated" in the above manner, together with the announced values and uncertainties serves to extend the unit to other facilities.

Figure 5 can also be used to illustrate the results of measurements of the difference between a particular

"calibrated" artifact and unknown by three different processes. Because the process responds to the actual magnitude of the properties embodied in the artifact and the unknown, the results from all three processes should tend to group about the same limiting mean. For all three processes, however, the number assignment for the unknown will be in error because the accessible unit, or the number assignment to the artifact, is not quite correct. The results from all these processes would be biased, or offset, in one direction or the other, by an amount which would rarely exceed the uncertainty of the value assigned to the artifact, the "unit error." The importance of the magnitude of the "unit error" depends upon one's ability to detect that it exists.

A measurement system consists of many different measurement processes, each with some suitable access to a unit. Some of these processes have characteristics like distribution (c), others have characteristics like distributions (a) and (b). The measurements from well characterized processes, regardless of the magnitude of the total s.d., would be considered in agreement if the results of measurements on a given object tend to support a single limiting mean. That is, the uncertainty limits associated with each reported value should tend to overlap a common limiting mean. If the local "unit errors" are large and not accounted for in the uncertainty of the stated result, for processes with the characteristic of distribution c, this condition would almost never occur. Each stated result would be biased by the magnitude of the local "unit error." The results from some facilities would always be high, and from others, always low with respect to the average of all of the results from similar facilities.

For realistic prediction limits for the agreement within the system, the local "unit error", that is the uncertainty of the value assigned to the "calibrated" replica, must be considered as a systematic error and added to the estimate of the local random error, $3\sigma_T$. For processes with characteristics similar to distribution a, the local "unit error" is still a bias which should be included in the uncertainty statement. In this case, however, if the magnitude of the local "unit error" is small relative to the random variability of the process, few could afford to make the number of measurements necessary to verify its existence.

The accessible unit can take many forms. It may be the sum of the values assigned to two objects, or the average of the values assigned to two objects, or the average value for a group of objects. Such an embodiment of the unit permits monitoring the relative stability of the items which in summation establish the unit. For most measurements, the accessible unit is a part of some instrument or measuring device, which in turn is some convenient realization of a unit-algorithm-model concept. Because of the simplicity of the manipulative procedures, and the ability to read, encode, or print out the desired quantity directly from the instrument, the importance of the unit-algorithm-model concept is frequently overlooked. It must be emphasized that the results from any process, be it a comparative process or a direct reading process, are a realization of particular unit-model-algorithm concepts. The usefulness of the result depends upon how these concepts relate to particular requirements.

To illustrate, consider the substitution balance and the large multiple lever scale, both of which can be used in a comparative mode, but are normally used as direct reading instruments. With the substitution balance, the direct reading measurement is a "comparison" between the object in question and the "calibrated" built-in weights of the instrument which are manipulated by the operator. For the multiple lever scale, the direct reading measurement is a "comparison" between the object in question and the position of a poise along a graduated beam, the mass of the poise and the lever ratio of the instrument being in essence a multiple of the mass unit. In the first case, the instrument assumes the role of the "unit" over the incremental difference between the built-in weights. In the second case, the instrument assumes the role of the "unit" over the capacity of the instrument. In both cases the task of establishing the "unit error" is a substantial one.

The instrument manufacturer normally assumes the responsibility for assuring that the initial instrument "unit error" is within some specified limit. Usually this can be accomplished by adjusting the instrument indications relative to the values of "known" weights made from materials which are reasonably similar to the materials of the built-in weights or the poise and beam. The interpretation of the instrument indication with respect to the object at hand, however, is the responsibility of the user. Failure to account for all of the factors in the algorithm may result in unexplainable discrepancies.

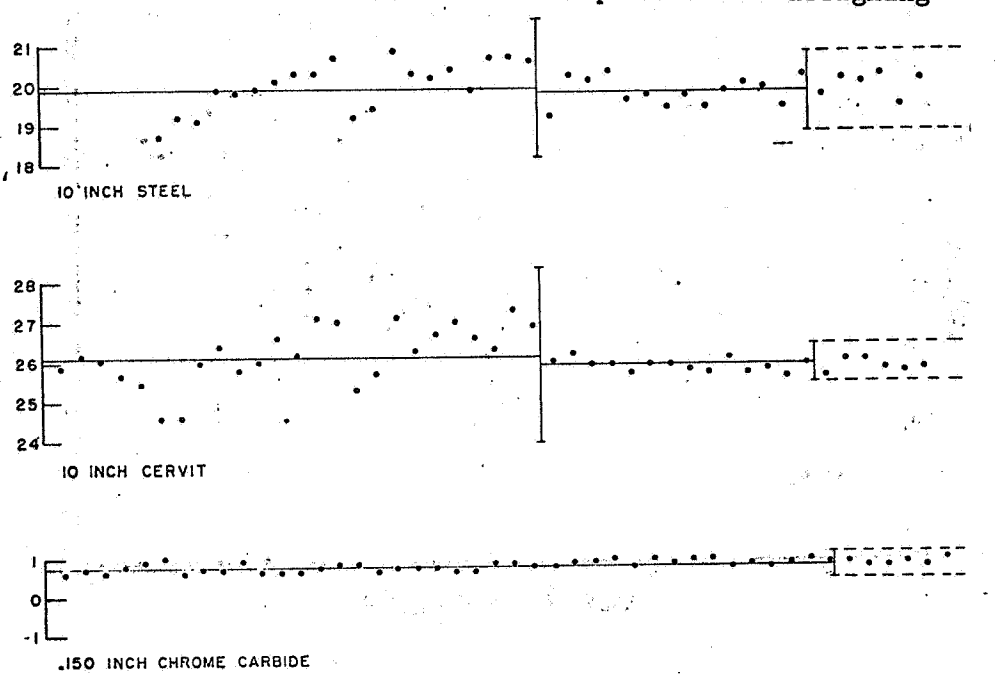
The unit error concept is applicable to all measurements which must be judged in accordance with established bounds for the limit of error. For direct reading instruments, it is prudent to have one or more "known" reference standards over the range of the instrument and in the neighborhood of the expected measurement result. In most cases, it is within the ability of the user to construct such standards, and in some cases, such action is necessary. For a wide variety of common measurements, suitable reference standards are available in the form of well characterized objects, instruments, and materials in which a considerable amount of effort has been spent to formulate the appropriate algorithms, and to reduce the magnitude of the unit error. The use of these items, in accordance with the appropriate algorithm and with measurement processes operating in a state of control, result in both an economic saving of measurement effort and an increase in the consistency of results. In many cases, the time and effort spent in trying to resolve inconsistent results is far greater than the time involved in the actual measurements.

7.0 The Practical Measurement Process

7.1 Verifying the Algorithm

The importance of the algorithm-model concept and the realization cannot be over emphasized. Model concepts and algorithms must be adjusted until the realization will provide consistent results which are adequate for the intended usage, for all materials which must be measured and for all environments in which the measurements must be made. This is not to say that all measurements should be made in accordance with the most complex algorithms, but rather that all of the factors in the algorithms have to be considered relative to each set of measurement requirements. When this is done, simplified instrumentation and procedures, if adequate, can be used with full confidence. When inappropriate algorithms are used, one is plagued with both real and imagined new sources of variability.

It is not enough to demonstrate consistency within one facility. The most severe test on the measurement system is to maintain consistency between facilities and environmental changes. For such tests on the system to have real meaning, each participating facility must first establish measurement processes which operate in a state of control. For example, figure 6 shows sequences of values for each of three "check standards" used in the NBS process for assigning



Early Interferometric Process Control Block Values
(Y, IN MICROINCHES PLOTTED IN SEQUENCE)

FIGURE 6

values to long gage blocks. These particular "check standards", or "control blocks", were chosen so that the difference between the .15 block and the 10 in blocks would emphasize length dependent variability, and the difference between the two 10 in blocks would emphasize temperature dependent variability. Clearly, in the beginning the variability of the results for the 10 in blocks was larger than the variability of the results for the .15 block. This suggested problems with the part of the algorithm which converts the vacuum wavelength to the effective wavelength.

A study of the values for the 10 in cervit block, using the techniques mentioned earlier, shows a correlation, figure 7, between the values and the vapor pressure in the environment at the time of the measurement. The corrective action, which resulted in a significant decrease in the variability of the values for both blocks, consisted of relocating the sensor to a position in closure proximity to the measuring position of the blocks. Further improvements have resulted in achieving approximately the same variability for each of the "control blocks."

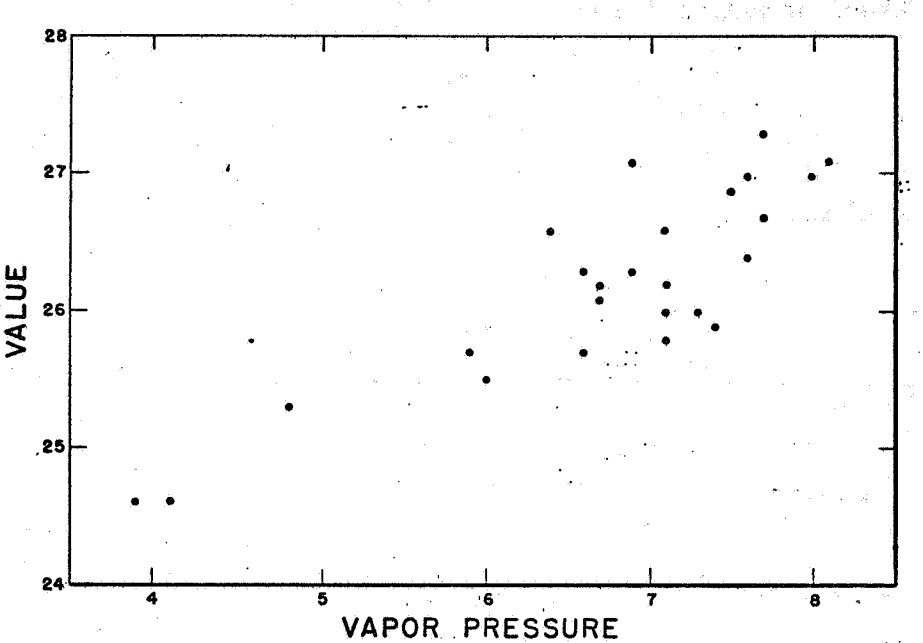
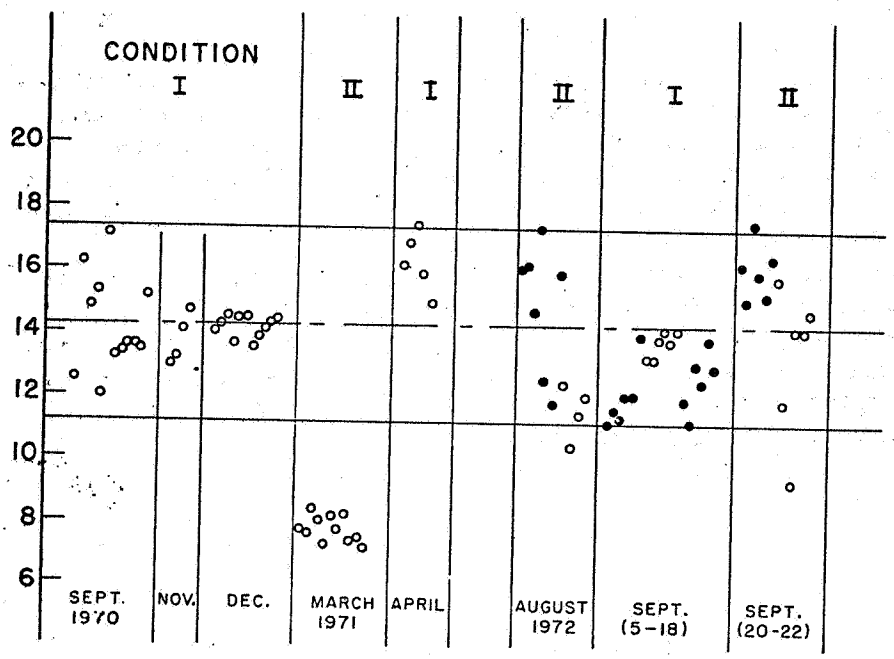


FIGURE 7

Figure 8 illustrates the results of measurements on the difference in length between two 16 in gage blocks. The initial value for the difference was established at NBS over the time interval Sept. 9 to Dec. 14, 1970. In a cooperating laboratory, the differences in March 1971 were considerably offset from the initial value. This offset was verified by repeating the measurements at the NBS in April. This discrepancy was not expected, and after considerable thought, it was attributed to the difference in the finish of the non-gaging surfaces of the block and the illumination level in the two facilities. The lab at NBS is almost dark, and the participating lab is a well lighted general purpose facility. The action which appeared to correct the problem consisted of wrapping the blocks in several layers of gold-coated mylar film, so as to achieve more nearly uniform thermal characteristics. While the measurements of 1972 did not necessarily confirm the action, at least the offset was no longer apparent.



16 in. BLOCKS, (2902-H155), PROCESS II
CONDITION I - LOW LIGHT LEVEL
CONDITION II - HIGH LIGHT LEVEL
○ UNLIKE FINISH
● "LIKE" FINISH (WRAPPED)

FIGURE 8

Youden suggests a technique for verifying the existence of systematic errors between facilities which are interested in measuring a particular quantity [12]. Two objects, or samples, similar in nature, are sent in turn to each of the participants. Independent measurements are made on each of the two samples, and the two results from each facility are plotted against each other. To quote, "Two median lines divide the graph paper into four quadrants. In the ideal situation where only random errors of precision operate, the points are expected to be equally numerous in all quadrants. This follows because plus and minus error should be equally likely. . . . . In any existing test procedure that has come to my attention the points tend to concentrate in the upper right and lower left quadrants. This means that laboratories tend to get high results on both materials, or low results on both materials. Here is evidence of individual laboratory biases."

Such a test, called TAPE 1, was made among the participants in the Mass Measurement Assurance Program [13].[1] Two pairs of stainless steel kilograms, designated X1, X2 and Y1, Y2, were circulated in such a way that a time interval from three to six months occurred between pairs for each facility. The results, including the results from NBS measurements with reference to the "defining artifact" kilograms, N1 and N2, are shown in figure 9. For each facility, the same measurement algorithm was used. The "unit error" of the local unit was on the order of 0.1 mg, and the process performance parameters of all facilities, σ_W and σ_T, were reasonably well known. The results from the first series of measurements were used, except repeats were required if an "out of control" situation existed. All of the results were in agreement on the basis of overlapping uncertainty limits.

---

[1] Time And Place Evaluation

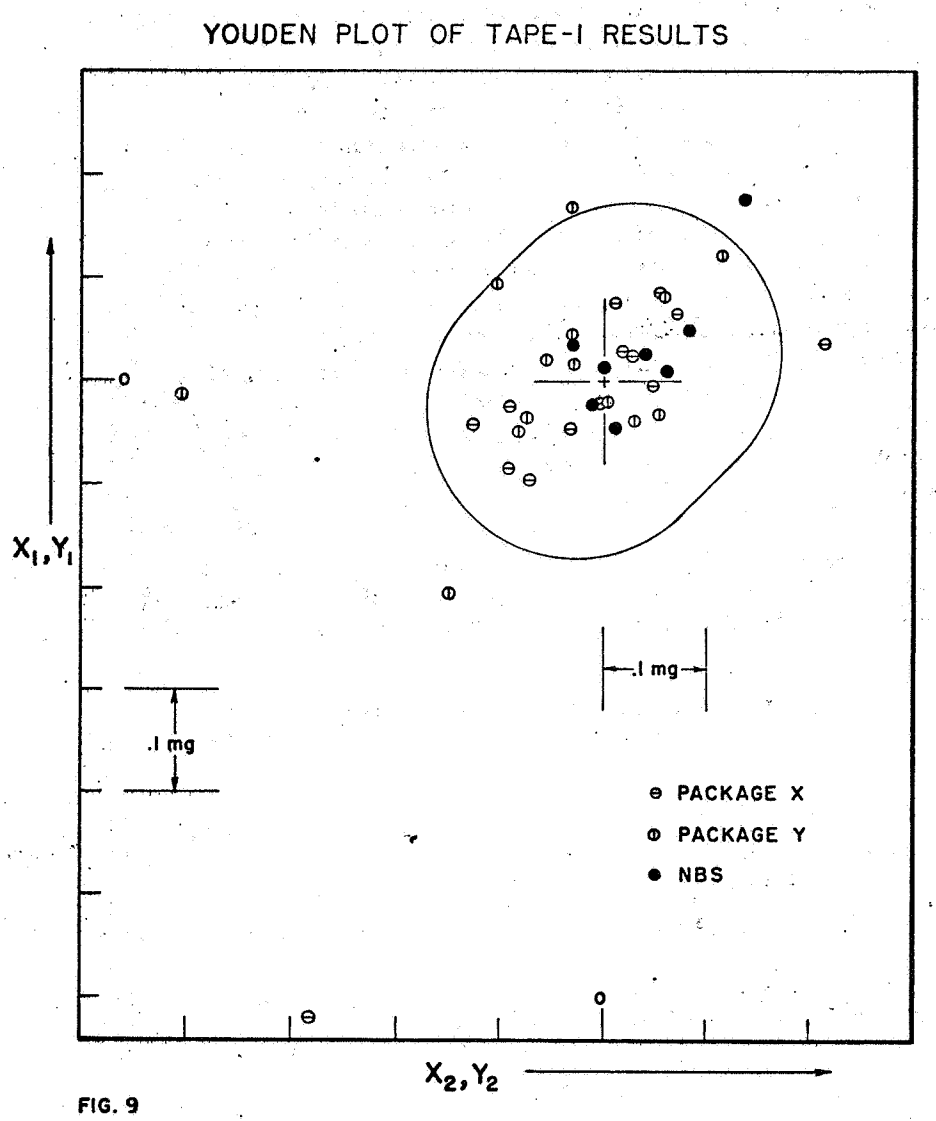


FIG. 9

For each measurement process, it would be expected that a sequence of such paired values would fall within a circle, the diameter of which would be a function of the process total standard deviation. For the NBS values, with the exception of one outlier, this appears to be the case. For other processes of comparable precision, one would expect the values to group about a line of 45° slope in a manner commensurate with the "unit error" of the local unit. Again, this is the case. The conclusion is that, in addition to an overall system agreement better than a part in 10^6, the measurement algorithm and its realization is reasonably correct for the items which were measured.

The more precise facilities participated in an extension of this, called TAPE 2, which involved tantalum, aluminum and stainless steel kilograms [14]. The results were as shown in figure 10. The ellipse shown is the "expected agreement" based on the previous test. There were difficulties from the beginning in achieving "in control" process performance. While the results for the stainless steel kilogram were as expected, there is clearly a correlation between the results for the tantalum and the aluminum kilograms in which the first order dissimilarity is in displacement volume, and also, perhaps, surface characteristics. This suggests both local process problems in measuring the parameters necessary to compute the air density and algorithm problems associated with the actual computation of the air density, and the accounting for surface effects, if need be. These studies are continuing, primarily, in order to construct an appropriate algorithm for such a situation. Until this is done, there is no real assurance in the ability to assign consistent mass values to objects made from materials other than that normally used to make

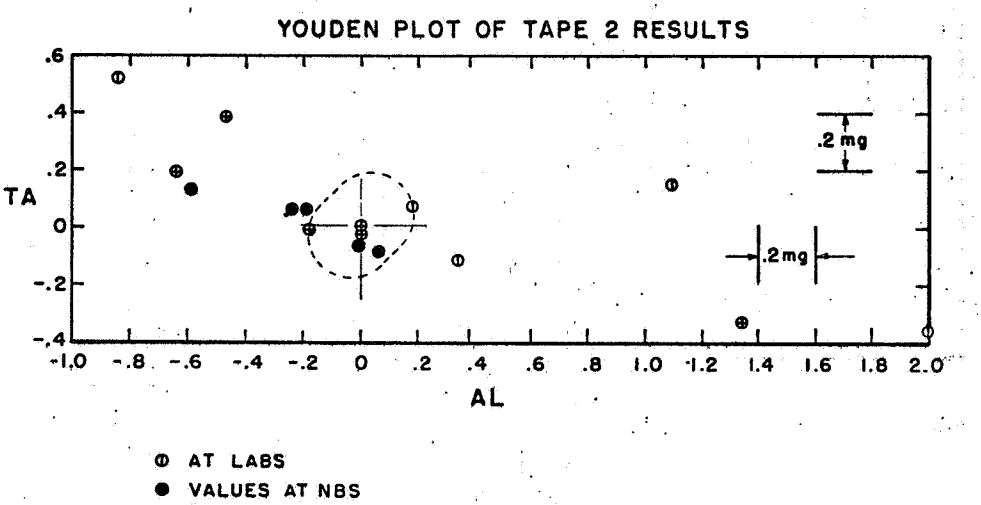


FIGURE 10

One form of an accessible length unit is the gage block and its assigned length. Two different processes are involved in determining the announced value. One process, the interferometric process, determines the lengths of selected blocks, the algorithm being based

on a model which "declares" the length of the block to be the separation between a defined point on the top gaging surface of the block and the surface of a platen "wrung" to the bottom surface of the block, the platen being made from the same material as the block. The procedures are tedious and time consuming. The total s.d. of the process is relatively large. Few facilities have the capabilities for such measurements, since many measurements are required to reduce the "unit error" imposed on the process in which the "calibrated" blocks will be used.

The second process, the mechanical comparison process, can only determine the difference between the reference blocks and the "unknown" blocks. The algorithm is based on essentially the same model, except the block is not "wrung" to the reference plane. The procedures for this process are simple and comparisons can be done rapidly. The total s.d. is somewhat less than that of the interferometric process. There are many such processes in every day use. The announced value can be computed from the value assigned to the reference block and the "unknown." The uncertainty of the announced value, at NBS, is the sum of the "unit error", or the uncertainty from the interferometric process, and the $3\sigma_T$ limits for the transfer process. This announced uncertainty is the "unit error" for the process in which the calibrated blocks will be used.

Confidence in the two-process system is considerably enhanced by demonstrating that both processes are in agreement, within their respective uncertainties. For example, the difference in length as computed from the values assigned to two blocks by the interferometric process can be measured directly in the comparison process. For similar block materials, the agreement is as shown in table 1. When different materials are involved, both process algorithm-model concepts must be reviewed carefully.

Table 1[1]

Comparison ((.)-(..)), Process I and Process II, January 1974

Nominal Size	Process I (.)-(..)	UNC	Process II (For 1-1-74) (.)-(..)	Total S.D.	D.F.	3S.D. Mean	(Delta)
5	27.95	.95	28.460	.489	90	.153	-1.29
6	32.70	1.14	33.476	.525	97	.513	- .78
7	23.53	1.37	23.562	.543	106	.546	- .03
8	43.94	1.54	44.212	.753	309	.129	- .27
10	61.95	1.97	61.373	.729	88	.231	.58
12	56.65	2.36	56.908	1.104	89	.348	- .26
16	55.64	3.26	57.282	1.072	71	.378	-1.24
20	-15.33	4.20	-16.484	1.054	94	.324	1.15

[1] In the above table, the designation (.) and (..) is used to differentiate between two specific blocks of the same nominal length. The Process I ((.)-(..)) has been computed from interferometric measurements. The Process II ((.)-(..)) has been measured directly by a mechanical comparison process. UNC is the uncertainty of the interferometric difference. The degrees of freedom, D.F., is the number of independent measurements in the collection used to determine the total S.D, (Delta) illustrates the closure between the results from the two different processes, {Process I ((.)-(..)) - Process II ((.)-(..))}.

In the interferometric process, the "differential" phase shift at the reflecting surfaces of the block and the platen have been "defined" out by requiring both the block and the platen to be of the same material. In like manner, in the transfer process, "differential" penetration of the probes into the surfaces is "defined" out for like materials. Where different materials are involved, such as quartz, cervit, and the various carbides, the comparison process algorithm must consider "differential" penetration. Normally one would compute the "differential" penetration correction, to be included in the algorithm, using a form of the Hertz equations and the necessary parameters [15]. In this case, the result can be checked with measurements using the interferometric process. The difference between the values for cervit blocks as determined interferometrically on a quartz platen, and as determined relative to alloy steel reference blocks was approximately 2 microinches, an amount in excess of

the normal process variability[1]. Until this discrepancy is resolved, the uncertainty of the values assigned by NBS to blocks from materials other than alloy steel includes an allowance to account for the doubt associated with differential penetration.

Fortunately, a complete algorithm for one particular process is usually appropriate for most similar processes, provided that the algorithm accounts for the range of condition, and materials, which must be made. The same is true for processes operating in a state of control. If it can be demonstrated that, for a particular algorithm, a particular process will operate in a state of control over the expected environmental changes, and with the various objects which have to be measured, one is reasonably sure that all similar processes which use the same algorithm can also be made to operate in a state of control. For each process, however, it is necessary to determine the appropriate process performance parameters.

---

[1] This is to say only that under conditions of gage block comparison, it may not be possible to determine the appropriate parameters for the Hertz equations. In many cases there is no competing process which can be used to verify the results of the algorithm independently. In the case of the measurement of "thread wires", the Hertz equations are used to determine an unstressed diameter from measurement data taken from a stressed condition. In use, the same relations are used to determine a stressed diameter from an unstressed condition. If the stressed condition in measurement and in use are identical, or nearly so, minor differences between the Hertz corrections and the actual differences may be of no importance. The same is true for other corrections which may be included in the algorithm, such as thermal coefficients of expansion. However, if the objects are to be used under conditions which differ significantly from the conditions under which the values have been assigned, it may be necessary to verify the appropriateness of the algorithms used.

7.2 Performance Parameters

In the beginning, the value for σ_w and/or σ_T is unknown, therefore the question of how much effort is required to establish a reliable value for the process to establish a reliable value for the process standard deviation must be considered. One cannot normally afford to make a sequence of repeated independent measurements sufficiently large to determine the long term process standard deviation. As an alternate procedure, one relies on estimates of the standard deviation, s, computed from collections of n repeated measurements. When intercomparison designs are used, estimates of σ_w are obtained from each sequence of measurements. With single measurements, as is the normal situation for production measurements, initially, and at reasonable intervals, sequences of n measurements must be repeated to establish s, and subsequently the long term process standard deviation, σ_T. The variation in independent values s_1, s_2... of the estimated standard deviation is surprisingly large if the s_i are based on small numbers of degrees of freedom. (The degrees of freedom are (n-1) when only an average is involved).

Although the distribution of s is not symmetrical about the standard deviation, sigma, for small degrees of freedom (e.g. less than 10), the standard deviation of the distribution of s, expressed as a percentage of sigma, is a useful guide in determining the number of observations required to determine a reasonable value for s. Table 2 gives the percentage standard deviation associated with s based on different degrees of freedom [16].

Table 2

Degrees of Freedom	Standard Deviation of s as a Percentage of σ
1	60.3
2	46.3
3	38.9
4	34.1
5	30.8
6	28.2
7	26.2
8	24.5
9	23.9
10	22.1
12	20.2
15	18.1
20	15.7
25	14.0
50	10.0

It is suggested that a sequence of 8 to 10 repeated independent measurements should be sufficient to provide an initial estimate of s for the production process. For a process operating in a state of control, using either comparison designs, or repeated sequences of single measurements, the estimates s_i can be "pooled", or combined, to determine the total process s.d. The total process s.d. should be a reasonably stable value for cumulative degrees of freedom in excess of 50.

The word "precision" is usually associated with the magnitude of the total process variability. For example, distribution a in figure 5, would be called a "more precise" process than either distribution b or d. However, the total process s.d. for a single measurement is a process characteristic, whereas the variability exhibited in a collection of data and the associated standard deviation are functions of the process definition. For example, a collection of data involving the same object and the same instrumentation

treated in several different ways is shown in figure 11 [17]. Figure 11(a) illustrates the results from a group of measurements in which a "single" measurement" is defined as the difference between the standard, S, and the unknown, X. In figure 11(b) the same data has been used but the "single measurement" has been defined as the average of two measurements of the difference between S and X. In figure 11(c) the result is the value obtained from an intercomparison design.

It is immediately apparent that the variability and thus the standard deviation associated with each treatment is markedly different even though the total measurement effort is the same. The uncertainty of the value established for X relative to S relates the correctness of the long term average of the respective total collection of data while the precision of each treatment refers just to the variability of the collection of data about their central values. The uncertainty of the average in each case is the same.

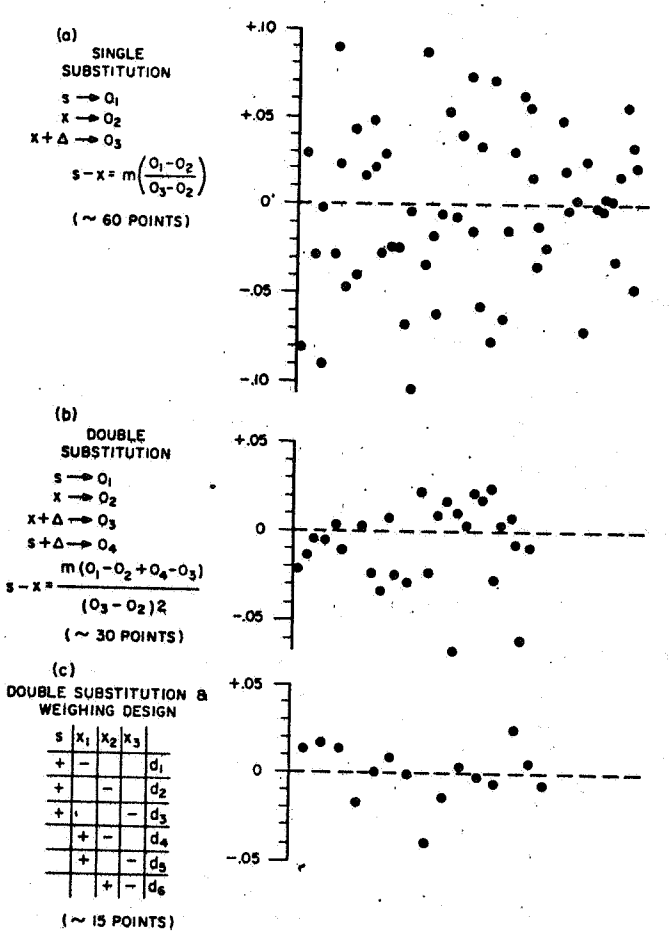


FIGURE 11 Process precision as a function of defined measurement methods.

The above is not intended to imply that the precision of a process is not important but rather that the precision must be considered carefully in terms of the process requirements. For a given algorithm, the more precise the process, the less the measurement required to produce satisfactory results. On the other hand, as the process precision is increased, the variability of the data reflects more and more disturbances from various sources. Sometimes such variability can be handled by "rounding" but for complete explanation, algorithm modification is necessary. For the research process, algorithm modification is a necessity. For the production process, accounting for variability which is not significant with respect to the requirements is a wasted effort.

Aside from initial estimates, the long term total process standard deviation is the variability associated with repeated measurements on a "check standard." The "check standard" can be introduced into a measurement process in a variety of ways. For a comparative measurement, the "check standard" may be the difference between two objects used to introduce the unit, such as a pair of mass standards or a pair of gage blocks. It may be a selected object, similar to other "unknowns" which is measured at frequent intervals until such time that a reasonably stable estimate of the total s.d. is obtained, and then, perhaps at less frequent intervals to verify the validity of the accepted total s.d. For a production process, it may be a dedicated artifact, similar in nature to the product, which is measured on a regular routine schedule. Properly chosen, the variability of the collection of values is a real measure of the total process variability. The collection of values becomes supporting evidence for the total s.d., and, if necessary, a measurement of the "check standard" at any time should demonstrate that the prediction limits are valid.

7.3 Unit Error

The magnitude of the "unit error" relative to the total s.d. is important to the proper interpretation of the measurement result relative to the requirement. For comparative measurements, where the unit is introduced into the process by means of "calibrated" reference artifacts, it may be possible to achieve a situation in

which the "unit error" is some small fraction of the total s.d. In this case, the random variability between the results from several measurement processes can sometimes be reduced by having all processes introduce the unit by means of the same "calibrated" reference artifact, or perhaps a suitable well characterized reference material.

In both of these cases, it is necessary that all of the processes use essentially the same algorithm as was used to establish the value assigned to the "calibrated" artifact. Failure to do so may lead to illusory results.

In the case of direct reading equipment, the unit is introduced by the instrument, and as a consequence, the determination of the magnitude of the unit error is a sizeable task. One class of instrument, such as a substitution balance, relies on a linear relationship to subdivide the interval between discrete settings. Other instruments, such as the gage block comparator, used a linear relationship to subdivide a fixed interval which, in turn, can only be used to determine differences. Some elastic devices rely on a relationship which may or may not be linear over the whole range of the instrument [18]. In these cases a linear indicating scale is often imposed on a nonlinear response in such a way as to distribute the nonlinearity over the range of the instrument. Instrument "calibration" requires careful attention to the nature of the instrument, the manner in which it is used, and how the results are to be interpreted.

Assigning a value to a discrete instrument setting with reference to a known "standard" is exactly similar to the task of assigning a value to any other unknown with reference to that "standard." The uncertainty of the value assigned to the setting is the sum of the "unit error" of the reference standard and of the instrument itself. This summation becomes the "unit error" of the instrument. If the instrument is sufficiently stable, the magnitude of the "unit error" embodied in the instrument can be reduced by the use of reference standards in which the "unit error" or uncertainty is small with respect to the instrument s.d., and by using the average of a number of independent "calibration" runs.

The proportional part of the instrument indication can be evaluated with two known "objects" which differ in magnitude by an amount somewhat less than the smallest incremental instrument setting. For example, in mass a small known "sensitivity" weight, Δ, is chosen in such a manner that $(S+\Delta)>X>S$. From the three observations:

$$(S+\Delta) \rightarrow O_1$$
$$X \rightarrow O_2$$
$$S \rightarrow O_3$$

one can form the relation:

$$X = S + K(O_2 - O_3) = S + \left(\frac{\Delta}{O_1 - O_3}\right)\left(O_2 - O_3\right)$$

where $K = \left(\frac{\Delta}{O_1 - O_3}\right)$, the mass value per division of instrument reading scale.

Traditionally, for independence, the value of K as determined in a particular comparison has been used in the computations. If, however, K is a stable property of a particular instrument, the use of an average K will result in a smaller s.d.[1]

With certain instruments, operator or servo, adjustment of the instrument configuration to obtain a defined "null" position is necessary before the indication can be recorded. For the operator, this can be a tedious and time consuming operation, particularly if the time constant of the instrument is long. Either, or combinations of, reducing the instrument sensitivity, and "rounding" the indication are used to alleviate this condition in many practical measurement processes. "Rounding" occurs when the operator is instructed to "read to the closest marked interval", or to adjust to within some defined limits around the desired "null" positions, or by purposely dropping digits in the instrument indicating system. The sensitivity may be reduced to the point that all variability is masked. For both of these cases, the "unit error" of the instrument must be considered carefully.

If, with the instrument operating properly and with no intentional "rounding", a sequence of measurements on a "known", similar in all respects to other "unknowns" which are to be measured, produces a sequence of identical results, the instrument is not sufficiently sensitive to detect normal process variability. In this case, the random component of the uncertainty is

---

[1] The same is true for the practice of resetting the zero of an instrument. If the instrument "drifts" because of some environmental problem, the effects of drift can be reduced somewhat by resetting zero immediately before each measurement. On the other hand, if the change in zero is random in nature, the practice of continually readjusting zero will result in a larger s.d.

zero. That is, the instrument unit error is the uncertainty of the measurement. The instrument unit can be established by adjusting the reading scale so that the indication agrees "exactly" with the value assigned to the "known." By adding, or removing small increments from the "known" one can establish the magnitude of change necessary to cause a change in the instrument indication. This change, together with the uncertainty of the value assigned to the "known", is the instrument "unit error." The second element of the uncertainty of the result relates to the manner in which the instrument is used. If the operator "rounds" to the closest graduation, the value of the graduation interval must also be included in the uncertainty statement.

It is not always possible to adjust the instrument reading scale as suggested above. In this case, the residual difference between the instrument reading and the value assigned to the "known" is a systematic error, or bias. Such a bias may exist at each incremental setting of the instrument. The bias, or reading scale offset, for a particular setting applies to all measurement data over the on-scale range of the instrument for that particular setting. With suitable tests, both the magnitude and the direction of the offset can be determined and incorporated in the measurement algorithm.

"Rounding" by dropping digits also introduces a bias, or offset. In this case, for all X within the interval $I_1 \leq X < I_2$, the recorded value for X is I_1, thus the recorded value for X may be low by as much as the value of the increment $|I_1 - I_2|$. The consequences of this type of rounding will be discussed later. In some cases, it may be possible to establish an "unrounded" value for X. If the instrument indication is I_1 for X, one can add a small summation, Δ_1, to obtain an indication I_2. Continuing, add a second summation, Δ_2, to obtain an indication I_3. Assuming the instrument response to be linear over the range I_1 to I_3, the "unrounded" estimate of X is then:

$$X = I_1 + \frac{\Delta_1 - \Delta_2}{\Delta_2}$$

In this case Δ_2 is the amount necessary to change the indication by one unit, I_1 to I_2. For indication I_1, with $I_1 < X < I_2$, Δ_1 is the increment which must be added to obtain I_2. There are numerous variations of this procedure.

As it becomes necessary to announce smaller uncertainties for the measurement results, more sensitive, or precise measurement, processes must be used in order to detect, with some assurance, small changes in the property of interest. In these cases the increment of the normally used rounding practices may be smaller than the observable process variability. Evidence that such is the case can again be obtained from sequences of repeated measurements, as for example, a sequence of 10 numbers, 8 of which are identical, and with one each plus one "increment" and minus one "increment." A proper assessment of the "unit error" and the process performance characteristics cannot be derived from such a distribution. This is not to say that such procedures will not produce adequate results, but rather that uncertainty associated with the result must be based on a more comprehensive study of the process.

One technique which is applicable to a variety of measurement processes is that of changing the mode of operation for the purpose of determining the total process s.d. By changing from a direct reading to a comparative mode of operation, the rounding effect of the instrument indicating system can effectively be bypassed. With ingenuity, in addition to the total process standard deviation, one can also determine the instrument "unit error" for the incremental settings as well as the proportional subdivision of the increment between settings.

For example, consider the case of the large multiple lever weighing scale. With the instrument settings in a fixed position, a "known" weight can be placed on the platform and, by adding summations of small "known" weights, the position of the weigh beam can be brought to some arbitrarily defined "null" position[1]. The large weight can be removed and replaced on the platform, again adjusting the summation to obtain a "null" position. Reordering the changes in the summation of small weights for a sequence of such measurements provides a set of data which can be used to determine an estimate of the total process standard deviation. The method is directly extensible to the comparison of two weights, and the "calibration" of one weight with respect to the other. Proper selection of weights to be added or removed provides a means to evaluate both the incremental instrument settings as well as the proportional subdivision. While this method of testing is not identical to the manner in which the instrument is normally used, the process characteristics determined are appropriate for determining the uncertainty for all modes of operation.

---

[1] This may require the addition of a suitable pointer and a small linear scale at the tip end of the weigh beam as shown in reference [19].

7.4 Measurement Requirements

Production measurement requirements may be stated in several ways, such as: (1) determine the magnitude of the property within some specified uncertainty limits; (2) determine that the magnitude of the property does not deviate from some desired magnitude in excess of some specified limits; or (3) determine the magnitude of the property with an uncertainty appropriate to the requirement that the announced result will not disagree with the result from another facility in excess of some specified limits. The goal of measurement assurance efforts is to provide evidence that the performance of a given process is adequate with respect to any one, or all, of these requirements. Fundamental to this assurance is a realistic estimate of the "unit" error, and demonstratable evidence to support the total process s.d. It should also be evident that it is of primary importance to verify that the required limits are valid with respect to the manner in which the measurements are to be used.

A generalized uncertainty statement might be as follows:

$$\text{Uncertainty} = \begin{bmatrix}\text{Instrument}\\ \text{"unit error"}\end{bmatrix} + \begin{bmatrix}\text{Unaccounted}\\ \text{for S.E.}\end{bmatrix} + \begin{bmatrix}\text{Random}\\ \text{variability}\end{bmatrix}$$

For any given situation, any one of the three elements may be predominant. The instrument unit error is fixed by the mode of operation. While the magnitude can be established, the direction cannot. Generally, this term can be reduced by changing to a comparative mode of operation in which the "unit error" is the uncertainty of the reference standard used. The "unaccounted for S.E." includes "rounding", uncorrected bias, and using inappropriate algorithms. Some of these may be known both in magnitude and direction. It is generally possible to reduce the term to the point that such effects are no longer identifiable in the end results. The "random variability" is a function of the total process s.d., σ_T[1]. The magnitude depends upon the amount of effort

---

[1] Operator skills are included in all nonautomated measurement processes. The least important are those associated with "setting zero", estimating tenths of divisions, and the like. It is reasonably easy to teach operators to make unbiased estimates of pointer positions between two numbered intervals. In many cases devices such as digital "readouts" and card printers replace this function. It is most important for the operator to clearly understand the principles of the instrument as they relate to the measurement task.

expended in a given measurement. The direction cannot be determined. Considering the economy of measurement effort, for conditions in which relatively large uncertainties are acceptable, the magnitude of the second term in the above relationship should predominate. Practices which are adequate for the requirements may mask completely the last term. In the limit, with sufficiently well known standards available, the last term is the predominate one. In this case, the residual systematic effects after correction for known sources of variability and the effects from unknown sources are the components of the random variability.

The problems associated with testing for compliance with specified tolerances are somewhat more complex. Tolerance may be associated with process control, as for example, the exceeding of specified tolerance limits may be a signal for change of process parameters, e.g. change in part size as related to cutting tool wear. Tolerances may also specify a band within which one assigns a common number to indicate the magnitude of a particular property. In the latter instance, if the variability between items from the production process is small relative to the tolerance band, the process could be adjusted so that the average value is near the tolerance limit. This introduces an unexpected bias. One normally expects the average product value to be in reasonable agreement with some specified nominal value. Such a condition could go undetected if the procedure used for testing for compliance is not sufficiently precise, and the tests occur at random over long time intervals. On the other hand, if the production process is not capable of meeting the tolerance specifications, there may be endless haggling over the question of compliance or noncompliance irrespective of the suitability of the product in its intended usage.

If the acceptance of a tolerance structure is appropriate to a certain requirement, the same philosophy must be extended to the problem of determining compliance with the structure for the objects with values at or near the tolerance limits. As an example, for some classes of weights, two tolerance limits are specified, one for adjustment and one for maintenance. While it is normally assumed that the maintenance tolerance is associated with a certain allowance for wear, it also tends to prevent rejection for noncompliance, or re-adjustment based on the assumption of change, when tests are made by various processes with unknown performance characteristics.

For tolerance tests with processes in which the predominant error is systematic, that is, the sum of the "unit error" and "unaccounted S.E.'s", the tolerance limits should be narrowed by the magnitude of the systematic error of the process. Any value which lies within the narrowed limits could safely be considered to be indicative of an in-tolerance situation. If the random variability is the major component of the process variability two actions are required. The tolerance limits must be narrowed by the "unit error" of the instrument, or the local accessible unit, and a "rule" must be established to define compliance relative to the narrowed limits. A simple rule for judging compliance is the same as above, any value which is within the narrowed limits could be considered as indicative of a within tolerance situation. If the risk associated with an out-of-tolerance condition is large, the total s.d. of the process should be some fraction of the narrowed tolerance limits. In most cases, a requirement that the s.d. of the process should be on the order of one-fifth of the narrowed limits should be satisfactory. In any event, rejection on the basis of an out-of-tolerance condition on the order of a fifth of the tolerance band negates the whole philosophy of the convenience of a tolerance structure.

Any meaningful comparison of measurement results from different processes must be predicated on the fact that both measurement processes are well characterized. That is, significant systematic effects must be accounted for and the uncertainty associated with each process must be realistic. For two such processes with very nearly the same total s.d., the expected limit for disagreement between results from a single measurement would be $(3\sqrt{2})\sigma_T$. Thus for some specified limit on the agreement, L:

$$(3\sqrt{2})\sigma_T \leq L \text{ or } \sigma_T \leq \frac{L}{3\sqrt{2}}$$

from which it follows that the individual process s.d. should be on the order of 1/5 of the specified limit. If more than two processes are involved in measurements of the same item, and the requirement is that only one time in a hundred shall any two results disagree in excess of L, the individual process s.d.'s should not exceed those shown in table 3 [21].

Table 3

No. of Processes	Desired Process s.d.*
5	(L/4.6)
10	(L/5.16)
20	(L/5.65)
30	(L/5.91)

* The denominator is the value for the upper one percent point in the distribution of the range.

To illustrate the manner in which the random and systematic components of uncertainty are combined, consider the task of assigning a value to a 10 000 lb artifact, and the problem of assigning an uncertainty for the announced value. Such an artifact might be used as the "accessible unit" in a particular measurement process. For this example, all of the measurement processes are well characterized, and the announced value for any measurement is defined as the average of n "single measurements" with total s.d. σ. The accessible unit for the operation is a 50 lb reference artifact with known value and uncertainty, U_s. The measurement operations are, in sequence:

(1) Each of ten 50 lb artifacts with reference to the "standard"

(2) Each of five 500 lb artifacts with reference to the summation 500 lb established in (1)

(3) Each of four 2 500 lb artifacts with reference to the summation 2 500 established in (2)

(4) The 10 000 lb artifact with reference to the summation 10 000 lb established in (3).

The uncertainties for each sequence are as follows:

(1) For a 50 lb artifact:

$$U_{50} = U_s + 3(\sigma_{50}/\sqrt{n})$$

For the summation 500 lb:

$$U_{\Sigma 500} = 10U_s + (10)(3)(\sigma_{50}/\sqrt{n})$$

(2) For the 500 lb artifact:

$$U_{500} = U_{\Sigma 500} + 3(\sigma_{500}/\sqrt{n})$$

For the 2500 lb summation:

$$U_{\Sigma 2500} = 5U_{500} + (5)(3)(\sigma_{500}/\sqrt{n})$$

(3) For a 2 500 lb artifact:

$$U_{2500} = U_{\Sigma 500} + 3(\sigma_{2500}/\sqrt{n})$$

For the summation 10 000 lb

$$U_{\Sigma 10000} = 4U_{2500} + (4)(3)(\sigma_{2500}/\sqrt{n})$$

(4) For the 10 000 lb artifact

$$U_{10000} = 4_{\Sigma 10000} + 3(\sigma_{10000}/\sqrt{n})$$

Since the value for the 10 000 lb artifact will be used as a constant in the next process, the uncertainty of (4) above is the "unit error" of that process. It is a ± systematic error which must be added to the results of comparisons with the "calibrated" 10 000 lb artifact.

The above procedures can be extended establishing the uncertainty of an inventory, provided that the measurements are made with a well characterized process, and that the instrument "unit error" is very nearly the same over the range of objects which must be weighed. For an inventory of m objects, the uncertainty of Σm is:

$$U_{\Sigma m} = m(\text{instrument unit error}) + 3\sqrt{m}\ (\sigma_T)$$

In the above relation, the value for each m is a single "direct reading" with standard deviation σ_T.

In cases where the "direct reading" mode of operation is used, and the summation (Instrument "unit error" + Uncorrected S.E.'s) completely mask the process variability, the "inventory" uncertainty is:

$$U_{\Sigma m} = m\left(\begin{matrix}\text{instrument} \\ \text{unit error}\end{matrix} + \begin{matrix}\text{uncounted} \\ \text{S.E.'s}\end{matrix}\right)$$

Both terms in this relation must be considered carefully. In some cases, it may be possible to adjust the inventory value for bias associated with the instrument "unit error." If digital rounding of the type mentioned earlier occurs, with a large m and with the objects distributed over several rounding increments, the indicated inventory will be less than the actual inventory by the amount m(rounding increment/2). On the other hand, if the inventory consists of "net" values, computed from "gross" and "tare" values which have been obtained using the same digital rounding increment, the computed inventory summation is not biased. In some cases, it may be possible to reduce bias from digital rounding by relocating the rounding increment relative to the indicating scale. If the increment can be set so that, for some objects, the indicated value is biased in one direction, and for others, the bias is reversed, the effect on the total inventory will be reduced.

For those who interpret measurement results, it must be emphasized that the most one can expect of a sequence of repeated measurements from a typical measurement process, operating in a reasonable state of control, is a reasonably symmetric distribution about some limiting mean. While the results of repeated measurements from a few select processes which are very well characterized, which are generally located in a controlled environment, and which have been in operation for long periods of time may be nearly normally distributed about the mean, such a distribution is not necessarily a characteristic of all measurement processes. It is suspected that the basis for the assumption of normal distribution extends far back in history, where the variability associated with the operator and his ability to estimate the proper instrument reading was the largest source of variability in the processes under study. With appropriate training variability associated with operator "reading" error is generally of no consequence. In a well characterized mass measurement in which comparison designs are used, it can be demonstrated that attempts by the operator to "override" the normal process variability in order to obtain a smaller within-group standard deviation are evident in a loss of control on the value obtained for the "check standard."

8.0 Summary

The concept of a measurement process as a production process is relatively new, having evolved in the last ten years. There have been significant contributions from many sources which have served to refine the initial ideas. While some of the techniques may not be appropriate for certain highly specialized measurement processes, it is felt that the concepts are applicable to practically all measurement processes. For certain types of general measurement processes, which must operate in a variety of environments, and which must accommodate a variety of materials and properties, the techniques have been invaluable in understanding the manner in which measurement processes operate in a "real" world.

The Measurement Assurance Programs, associated with this concept, are only new words and minor refinements of something that has been going on for a long time -- doing whatever is necessary to provide assurance that the measurements are adequate for the intended usage. While these programs emphasize the importance of verifiable evidence, such evidence is not always expensive to obtain. Suitable "check standards" and some form of redundancy can be incorporated in almost all measurement processes. Simple analytical techniques will either verify that the process is performing as expected, or that it is not. Fundamental to the whole approach is the need to understand in the beginning just what the measurement operation is supposed to accomplish.

Any one measurement process is a particular realization of an accepted unit-algorithm-model concept. Meaningful quantitative evaluation of the product, that is, the measurement results, must be based on consistency, both with itself and with other measurement processes. The important questions are: "If I did it over again, right away, next week or next year, would the new result agree with my current result?" and "If someone else repeated the measurement with his process, would his result agree with mine?" The answers to both questions are predictable and verifiable for most measurement processes.

References

[1] Carnap, R., Philosophical Foundations of Physics, Chapters 6 and 7, (Basic Books, Inc., New York, 1966).

[2] Campbell, Norman R., An Account of the Principles of Measurement and Calculation (Longmans, Green and Co., Ltd., New York, 1928).

[3] Eisenhart, Churchill, Realistic Evaluation of Precision and Accuracy of Instrument Calibration Systems, J. Res. Nat. Bur. Stand. (U.S.), 67C (Eng. and Instr.), No. 2, 161-187 (Apr.-June 1962).

[4] Pontius, P.E. and Cameron, J.M., Realistic Uncertainties and the Mass Measurement Process, Nat. Bur. Stand. (U.S.), Monogr. 103, 17 pages (1967).

[5] Pontius, P.E., The Measurement Assurance Program -- A Case Study: Length Measurements, Part I, Nat. Bur. Stand., (U.S.), Monogr. to be published 1974.

[6] Cameron, J.M. and Hailes, G.E., Designs for the Calibration of Small Groups of Standards in the Presence of Drift, Nat. Bur. Stand. (U.S.), Tech. Note 844, 32 pages (1974).

[7] Almer, H.E., National Bureau of Standards One Kilogram Balance NBS No. 2, J. Res. Nat. Bur. Stand. (U.S.), 76C (Eng. and Instr.), Nos. 1 and 2, 1-10 (Jan.-June 1972).

[8] Schweitzer, W.G.Jr., Kessler, E.G.Jr., Deslattes, R.D., Layer, H.P., and Whetstone, J.R., Description, Performance, and Wavelengths of Iodine Stabilized Lasers, Applied Optics, Vol. 12, p 2927, December 1974.

[9] Calibration and Test Services of the National Bureau of Standards, Nat. Bu. Stand. (U.S.), Spec. Publ. 250 (Dec. 1970).

[10] Catalog of Standard Reference Materials, Nat. Bu. Stand. (U.S.), Spec. Publ. 260 (Apr. 1973).

[11] Page, C.H. and Vigoureux, P., The International System of Units (SI), Nat. Bu. Stand. (U.S.), Spec. Publ. 330 (Apr. 1972).

[12] Youden, W.J., Graphical Diagnosis of Interlaboratory Results, Industrial Quality Control, Vol. XV, No. 11 (May 1959).

[13] Progress Report - TAPE I, "Overlap" No. 2 (July 1969)*.

[14] Progress Report - TAPE II, "Overlap" No. 6 (March 1974)*.

[15] Puttock, M.J. and Thwaite, E.G., Elastic Compression of Spheres and Cylinders at Point and Line Contact, National Standards Laboratory Technical Paper No. 25, Commonwealth Scientific and Industrial Research Organization, Australia (1969).

[16] Natrella, M.G., Experimental Statistics, Nat. Bur. Stand. (U.S.), Handbook 91, pp. 2-12, 2-13, U.S. Government Printing Office (1963).

[17] Pontius, P.E., Measurement Philosophy of the Pilot Program for Mass Calibration, Nat. Bur. Stand. (U.S.), Tech. Note 288, 39 pages (1968).

[18] Anderson, G.B. and Raybold, R.C., Studies of Calibration Procedures for Load Cells and Proving Rings as Weighing Devices, Nat. Bur. Stand. (U.S.), Tech. Note 436, 22 pages (1969).

[19] Briggs, C.A. and Gordon, E.D., Weighing by Substitution, Nat. Bur. Stand. (U.S.), Technologic Paper No. 208 (Feb. 1922).

[20] Volodarskii, V.Ya, Rozenberg, V.Ya and Rubichev, N.A., Effect on the Precision of Measurements of the Discrepancy Between the Investigated Object and its Model, (translated from Izmeritel'naya Tekhnika, No. 7, pp. 18-20, July 1969) Measurement Techniques, No. 7 page 907 (1969).

[21] Pearson, E.S., Percentage Limits for the Distribution of Range in Samples from a Normal Distribution, Biometrika, Vol. 24, pp. 404-417 (1932).

[22] Blair, B.E., Editor, Time and Frequency: Theory and Fundamentals, Nat. Bur. Stand. (U.S.), Monogr. 140, Ch. 8, pp. 153-204 (1974).

[23] See reference [16], Ch. 5.

---

* Copies of Overlap may be obtained from P. E. Pontius, Mass and Volume Section, Room A123 MET, National Bureau of Standards, Washington, D. C. 20234.