NBS Special Publication 676-II

Second A Measurement Assurance Programs
Part II: Development and Implementation

Carroll Croarkin
Statistical Engineering Division
Center for Applied Mathematics
National Bureau of Standards
Gaithersburg, MD 20899

April 1985

DEPARTMENT OF COMMERCE ★ UNITED STATES OF AMERICA ★

Measurement Assurance Programs
Part II: Development and Implementation

Carroll Croarkin
Statistical Engineering Division
Center for Applied Mathematics
National Bureau of Standards
Gaithersburg, MD 20899

This document is a guide to the logical development of a measurement assurance program in which the tie between a measurement and its reference base is satisfied by measurements on a transfer standard. The uncertainty of values reported by the measurement process is defined; and the validation of this uncertainty for single measurements is developed. Measurement sequences for executing the transfer with NBS and procedures for maintaining statistical control are outlined for eight specific measurement situations with emphasis on characterizing parameters of the measurement process through use of a check standard.

Key Words: Calibration; check standard; measurement assurance; random error; statistical control; statistical methods; systematic error; uncertainty.

1. The Development of a Measurement Assurance Program

1.1 Historical Perspective

The development of measurement assurance at the National Bureau of Standards, over the more than eighty years that the nation's premier measurement laboratory has been in existence, has evolved hand in hand with the NBS central mission of providing quality measurement. We might date this evolution as starting with the early experiments on the velocity of light [1] [1]. Since then the principles of measurement assurance have reached realizations of all SI units and numerous derived units of measurement, and even now are influencing innovations in measurement science related to electronics and engineering.

As the reader familiarizes himself with the concepts of measurement assurance, he will come to realize that quality in calibration is dependent upon the inclusion of a check standard in the calibration scheme. The first application of this principle at NBS came in the area of mechanical measurements where a prescribed series of observations known as a weighing design, so called because of the obvious connection to mass weighings, defines the relationship among reference standards, test items and check standards. The first weighing designs published by Hayford in 1893 [2] and Benoit in 1907 [3] had no provision for a check standard, and the creation of suitable designs had to await general progress in the area of experimental design which characterized statistical activity at NBS in the nineteen fifties.

[1] The numbers in brackets refer to references cited at the end of this document.

As early as 1926 an NBS publication by Pienkowsky [4] referred to a standard one gram weight whose mass as "determined in the calibrations just as though it were an unknown weight" was used as a gross check on the calibrations of the other unknown weights. It remained until the nineteen sixties for the concept of measurement as a process to be described by repetitions on a check standard such as the one gram weight described by Pienkowsky. At that time calibrations of mass and length standards were formalized into measurement assurance programs with demonstrable uncertainty of reported values and statistical control of individual calibrations. A compendium of weighing designs for mechanical and electrical quantities with allowance for a check standard in each calibration sequence was published in 1979 (Cameron et al [5]).

Although many experimenters, past and present, have contributed to the quality of measurement science at NBS, the formulation of measurement assurance is the special province of the Statistical Engineering Division. Three members of this group, C. Eisenhart, W. J. Youden and J. M. Cameron, were largely responsible for fruition of the check standard concept, and the advent of electronic computers aided in the rapid application of this concept to NBS calibration programs. In 1962 a paper by Eisenhart [6] laid the groundwork for defining a repetition for a measurement process and assessing the uncertainties associated with such a process. This paper still serves as the primary treatise on the subject. Concurrently, Youden was implementing "ruggedness" testing in physical measurements [7], and at the same time he was introducing experimental design into interlaboratory testing [8].

In 1967 the first documentation of a measurement assurance approach appeared in print as an NBS monograph. The tutorial by Pontius and Cameron [9], treated the entire spectrum of mass measurement as a production process and began the dissemination of measurement assurance outside the NBS community. In the years since then, measurement assurance, both within and outside NBS, has been applied to basic SI units such as length as formulated in reference [10] and complex measurement areas such as dimensional measurements for the integrated circuit industry as formulated in reference [11]. Recently the measurement assurance approach has found its way into an ANSI standard for nuclear material control [12] with the use of "artifact reference mass standards as references for uranium hexafluoride" cylinders reported by Pontius and Doher [13].

1.2 Introduction

The development of a measurement assurance program evolves logically from the specific interpretation that we will give to the term "measurement assurance". The reader is asked to lay aside interpretations given to this term from previous experiences and to concern himself with what it means to have demonstrable scientific assurance about the quality of a measurement. For calibration activities, quality of a measurement is defined by its uncertainty, and the validity of an uncertainty statement for an individual measurement is guaranteed via the measurement assurance program as it is intended to

i) Tie a single measurement to a reference base; and

ii) Establish the uncertainty of the measured value relative to this reference base.

Firstly, in the case of basic SI units, a single measurement of a characteristic embodied in an object or artifact must be related to the defined unit for that quantity; for example, until recently the length of a gage block was defined relative to the wavelength of radiation of krypton 86 as realized through interferometry [14]. Because derived units of measurement can only be indirectly related to basic units, the measurement assurance concept is extended to such quantities by requiring that they be related to a reference base such as artifact standards or a measurement system maintained by the National Bureau of Standards. Secondly, a measurement assurance program must provide a means of maintaining statistical control over the measurement system thereby guaranteeing the validity of the uncertainty for a single measured value relative to its reference base (Cameron [15]).

The definition of measurement assurance is completed by an examination of the properties of measurement. A single measurement is properly related to national standards only if there is agreement between it and a value that would be achieved for the same quantity at NBS--meaning a value that would be arrived at from a sufficiently long history of like measurements at NBS. In actuality it is not possible to estimate the disagreement between a single measurement in a given laboratory and the long-term NBS value. However, if the measurement system of the laboratory is stable or as we say operating in a state of statistical control, the single measurement can be regarded as a random draw from another long history of measurements which also tend to a long-term value. The purpose of calibration is to eliminate or reduce the disagreement, referred to as offset, between a laboratory's long-term value for a measurement and the corresponding NBS long-term value by corrections to the measurement system and/or reference standards.

Where offset cannot be eliminated or reduced by calibration, it is a systematic error accruing to the laboratory's measurement system. Even where there is an accounting for such disagreement, the fact that NBS has imperfect knowledge about the long-term value from its own measurement system, based as it is on a finite though large number of measurements, means that the limits of this knowledge contribute another systematic error to the measurement system of the laboratory. In some special cases systematic and random errors that arise as NBS attempts to tie its measurement system to defined units of measurement may also become part of the systematic error for the laboratory.

The uncertainty that surrounds any single measurement describes the extent to which that single number could disagree with its reference base. The uncertainty includes all systematic errors affecting the measurement system; it also includes limits to random error that define the degree to which the individual laboratory, just as NBS, may be in error in estimating the long-term value for the measurement. Where the calculation of a long-term value for a measurement and limits to random error cannot be done directly, which is the usual case for calibration measurements, the long-term value is referenced to a long-term value of measurements made on an artifact(s) called a check standard.

Measurement assurance is attained when the determination of all sources of systematic error is coupled with statistical control of the measurement process as achieved by adapting quality control techniques to measurements on the check standard. Statistical control consists of comparing current check standard measurements with the value expected for such measurements and making decisions about the condition of the process based on the outcome of this test. The establishment of suitable check standards and implementation of statistical control procedures are discussed in the next two chapters with implementation for specific cases being outlined in chapter 4.

The determination of systematic error is made by intercomparing the laboratory's reference standard(s) or measurement system with national standards or a measurement system maintained by the National Bureau of Standards. This intercomparison can be interfaced with NBS in one of three ways. Firstly, the reference standards can be submitted to the usual calibration exercise wherein values and associated uncertainties are assigned to the reference standards by NBS. The only sources of systematic error that are identifiable in this mode are directly related to the reference standards themselves and to the NBS calibration process. The name "measurement assurance program" is not formally attached to such efforts because the NBS involvement is limited and measurement control is left entirely to the participant, but the goal of measurement assurance is certainly realizable by this route.

Secondly, systematic error can be identified by internal calibration of instrumentation or reference standards through use of a standard reference material distributed by NBS. Thirdly, systematic error can be determined by a formal program in which an NBS calibrated artifact, called a transfer standard, is treated as an unknown in the participant's measurement process. The difference between the participant's assignment for the transfer standard and the NBS assignment determines the offset of the participant's process or reference standards from NBS.

The National Bureau of Standards provides measurement assurance related services that utilize the latter two courses, especially the use of transfer standards, in selected measurement areas [16]. A standard reference material and a transfer standard are comparable in the measurement assurance context. The latter is referred to more frequently in this publication because transfer standards are more publicized in connection with measurement assurance.

The development of a program which satisfies the goals of measurement assurance begins with the measurement problem which must be related to physical reality by a statement called a model. Models covering three aspects of metrology are discussed in this chapter. The first of these, the physical model, relates the realization of the quantity of interest by a measurement process to the fundamental definition for that quantity. Physical models change with changes in fundamental definitions.

For example, until 1960, the standard of length was "the distance between two scratch marks on the platinum-iridium meter bar at the Bureau International des Poids et Mesures" [17]. Models for realizing length related the intercomparison between the international meter bar and the national meter bar and subsequent intercomparison between the national meter bar and gage block standards. In 1960 length was redefined in terms of the wavelength of radiation of krypton-86. The defining wavelength of 86Kr was related to the wavelength of a stabilized laser light[a], thus establishing the relationship of interference fringe patterns observed with the laser interferometer to the length of gage blocks standards. Length has recently been redefined in terms of the velocity of light. This latest change will necessitate another model relating standards to "the length of the path traveled by light in a vacuum during a (given) time interval" [17].

The calibration model describes the relationship among reference standards, items to which the quantity is to be transferred such as secondary or laboratory reference standards, and the instrumentation that is used in the calibration process. For example, calibration of gage blocks by electromechanical intercomparison with gage block standards that have been measured interferometrically includes a correction for the temperature coefficient of blocks longer than 0.35 inches [19]. The calibration model for these intercomparisons assumes a constant instrumental offset that is canceled by the calibration experiment as discussed in section 1.4.

Statistical models further refine the relationship among calibration measurements in terms of the error structure. Section 1.5 describes the type of error structure that is assumed for measurement assurance programs taking as an example the eletromechanical comparison of gage blocks according to the scheme outlined in section 4.

Modeling, usually the responsibility of the national laboratory, is emphasized in this chapter partly to lay the foundation for the remainder of the text and partly so that the reader can form some idea of the degree of success that can be expected from a measurement assurance program. It is an implicit assumption that the validity of any intercomparison, either between transfer standards and reference standards or between reference standards and the workload, depends upon all items responding to test conditions in fundamentally the same way as described by the models.

[a] "Direct calibration of the laser wavelength against 86Kr is possible, but is relatively tedius and expensive. The procedure used is a heterodyne comparison of the stabilized He-Ne laser with an iodione stabilized laser" (Pontius [18]).

This logic leads us the next major phase in development--the test of the measurement prescription as a device for transferring a quantity of measurement from the national laboratory to a laboratory participating in a measurement assurance program. The final phase--the application of quality control techniques to the measurement process ensures a continuing tie to the national system of measurement. Several activities can take place during each of these phases. These are listed either in section 1.6 under the role of NBS or in section 1.7 under the role of the participant although it is clear that in practice there is some overlapping of these responsibilities.

In summary, measurement assurance implies that the determination of systematic error and the assignment of values to standards has been done correctly at every step in the measurement chain and, moreover, that this is guaranteed by a statistical control program that is capable of identifying problem measurements at every transfer point in the chain. Accomodation to these principles may require modification of the laboratory's calibration procedures. Where an NBS transfer standard is used to directly calibrate reference standards, the same measurement process and control procedures that are used for this intercomparison should be used for the regular workload. Where the transfer standard is used to calibrate a laboratory's primary standards, a statistical control program should be implemented for this intercomparison, along with similar control programs for the intercomparison of the primary standard with the reference standards and for the intercomparison of the reference standards with the workload. Obviously, the effort required to maintain such a system is greater than is required to maintain a current calibration on the reference standards. Measurement assurance places a substantial portion of the burden of proof on the participant, where it should rightfully be, because it is the quality of his measurements that is of ultimate interest.

1.3 Models for a Measurement System

A measurement system that relies on an artifact demands that the artifact play "two essential roles in the system; it must embody the quantity of interest, and it must produce a signal, (such as the deflection of a pointer on a scale or an electrical impulse) which is unambiguously related to the magnitude or intensity of the specified quantity" (Simpson [20]). The first step that must be undertaken in constructing a measurement system is to reduce the artifact to an idealized model which represents those properties believed to be pertinent to the intended measurement.

This model of the measurement process, based on the laws of physics, in the broadest sense embodies our understanding of the physical universe. It is usually a software model or statement that relates the signal produced by the artifact and all procedures used to produce the desired measured value, called the measurement algorithm, to the realization of the physical quantity of interest taking into account any factors such as environmental conditions that affect this realization.

The integrated circuit industry is a case study of a measurement problem not properly defined in terms of an artifact model. Inability throughout the industry to measure optically the widths of chromium lines to the accuracies needed for producing photomasks for integrated circuits can be traced to misconceptions about the nature of linewidth measurements--misconceptions that led to reliance on a line-scale calibration for making such measurements, in the hope that a correct line-scale for the optical system would guarantee accurate linewidth measurements.

Before attempting to produce a linewidth standard, NBS explored the nature of the systematic errors that are inherent in line-scale and linewidth measurements (Nyyssonen [21]). Line-scale defines the internal ruler of an instrument i.e. it is basically a left-edge to left-edge or a right-edge to right-edge measurement for which any bias in detecting edge location is assumed to cancel out. Linewidth, a more difficult determination, measures the width of a physical object, in this case a chromium line. It is a left-edge to right-edge measurement in which any bias in detecting edge location is assumed to be additive (Jerke [22]).

This theoretical modeling was corroborated by an interlaboratory study which demonstrated that an optical imaging system, although properly calibrated for line-scale, would not necessarily produce linewidth measurements with negligible systematic errors. The study also demonstrated that the same system when properly calibrated using a linewidth artifact would produce linewidth measurements with negligible systematic errors (Jerke et al [23]).

A model is never complete or perfect, and the difference between the model and reality leads to "a particular type of systematic error which exists if the measurement algorithm is flawless. Failure to recognize this fact can lead to major wastes of resources since no improvement in the measurement algorithm can reduce this error" (Simpson [24]).

Thus even though NBS semiconductor research has greatly enhanced linewidth measurement capability, the accuracy of linewidth measurement is still constrained by the difference between the real edge profile of a chromium line and a theoretical profile (Nyyssonen [25]) upon which the model depends. The discrepancy between the edges of chromium lines on production photomasks and the theoretical model is a limiting factor in attaining measurement agreement among photomasks makers, and it will not be reduced by finer tuning of the optical imaging systems or more accurate standards. This points out a problem that exists in going from the calibration laboratory with carefully fabricated artifacts to the production line and prompts us to include a caveat for the claims of measurement assurance programs. This type of systematic error is kept at an acceptable level only if the measured items are close in character to the standards and theoretical model on which their assignments depend. The only strategy which can reduce model ambiguity identically to zero uses objects called "prototypes" and, in effect, takes a particular object and defines it to be its own model. As pointed out by Simpson [26],

> This amounts to saying that this object is the perfect and complete realization of the class of objects to which it belongs, and hence the model ambiguity is, by definition, identically zero. The only SI unit still using this strategy is mass where the Paris[b] Kilogram is the kilogram of mass, and the only objects where mass can be unequivocally defined are one kilogram weights made of platinum.

The comparison of a non-platinum kilogram with the Paris kilogram would produce a systematic error unless the comparison was done in vacuum. High accuracy mass calibrations in air are corrected for air buoyancy -- a correction that depends on the material properties of the weight, temperature on the weight at the time of weighing and the local pressure and humidity. Any ambiguity between the model that drives this correction and the Paris kilogram in vacuum contributes a systematic error to the calibration process although admittedly this error is negligible.

1.4 Models for a Calibration Process

1.4.1 The Calibration Experiment

The exploration of the physical and mathematical models that relate a measurement to a quantity of interest leads to a measurement algorithm which defines a reference standard, instrumentation, environmental controls, measurement practices and procedures, and computational techniques for calibrating other artifacts or instruments with respect to the desired property.

> Calibration is a measurement process that assigns values to the response of an instrument or the property of an artifact relative to reference standards or measuring processes. This may involve determining the corrections to the scale (as with direct-reading instruments), determining the response curve of an instrument or artifact as a function of changes in a second variable (as with platinum resistance thermometers), or assigning values to reference objects (as with standards of mass, voltage, etc.) (Cameron [27]).

Calibration consists of comparing an "unknown" or test item which can be an artifact or instrument with reference standards according to the measurement algorithm. The calibration model, which addresses the relationship among measurements of test items and reference standards, must reflect the fact that the individual readings on the test items and reference standards are subject to systematic error that is a function of the measuring system and random error that may be a function of many uncontrollable factors.

[b] The international standard of mass resides at the Bureau International des Poids et Mesures in Sèvres, just outside Paris.

There are two common generic types of calibration models, additive models and multiplicative models. Reduction of systematic error by intercomparison with a reference standard involves estimating offset as either an additive factor Δ or a scale factor λ which in turn is used to assign a value to the test item relative to the known value of the reference standard. The choice of an additive or multiplicative model depends on the nature of the relationship among test items and reference standards and properties of the measuring system.

The calibration experiment is designed not only to assign values to test items that will account for systematic error between the requestor and the calibrator but also to estimate the magnitude of random errors in the calibration process. The nature of random error is discussed more fully in section 2.2, but suffice it to say for now that we are talking about small fluctuations that affect every measurement but are unmeasurable themselves for a given measurement. The statistical derivations in this manuscript assume that the random errors are independent and that they affect the measuring process symetrically i.e., that one is not predictable in size or direction from any other one and that the chances are equal of the resulting measurement being either too large or too small. It is also assumed that random errors for a given process conform to a law called a statistical distribution; quite commonly this is assumed to be the normal distribution, and the calibration experiment is designed to estimate a standard deviation which describes the exact shape of this distribution.

In the next three sections we list models that are in common usage in calibration work, and although the list is not exhaustive, it includes those models which form the basis for the calibration schemes in chapter 4. It is noted that the term "reading" or "measurement" in this context does not refer to a raw measurement, but rather to the raw measurement corrected for physical model specifications as discussed in the last section.

1.4.2 Models for Artifact Calibration[c]

In the simplest additive model for a calibration process, a test item x with a value X^*, as yet to be determined, and a reference standard R with a known or assigned value R^* are assumed to be related by:

$$X^* = \Delta + R^* \tag{1.4.1}$$

where Δ is small but not negligible. The method for estimating the offset Δ between the two artifacts depends upon the response of the calibrating instrument.

If the calibrating instrument is without systematic error, the instrument response x for any item X will attain the value X^* except for the effect of random error; i.e., the instrument responds according to the model

$$x = X^* + \varepsilon$$

[c] The models for artifact calibration are also appropriate for single-point instrument calibration.

where ε represents the random error term. In this case there is no need to compare the test item with a reference standard because the capacity for making the transfer resides in the calibrating instrument. Such is assumed to be the case for direct reading instruments. Normally the calibrating instrument is not invested with such properties, and one calibration approach is to select a reference standard that is almost identical to the test item and compare the two using a comparator type of instrument for which additive instrumental offset is cancelled out in the calibration procedure. Given that the comparator produces a measurement x on the test item and a measurement r on the reference standard, the response is assumed to be of the form:

$$x = \psi + X^* + \varepsilon_x$$

and

$$r = \psi + R^* + \varepsilon_r \qquad (1.4.2)$$

where ψ is instrumental offset and the ε_x and ε_r are independent random errors. An estimate[¶] of Δ is gotten by the difference

$$\hat{\Delta} = x - r , \qquad (1.4.3)$$

and the value of the test item is reported as

$$\hat{X}^* = \hat{\Delta} + R^* .$$

An inherent deficiency in relying on a single difference to estimate Δ is that it does not admit a way of assessing the size of the random errors. If the calibration procedure is repeated k times in such a way that the random errors from each repetition can be presumed to be independent, the model for k pairs of readings r_j, x_j $(j=1,\cdots,k)$ becomes

$$\begin{aligned} x_j &= \psi + X^* + \varepsilon_{x_j} \\ r_j &= \psi + R^* + \varepsilon_{r_j} \end{aligned} \qquad (1.4.4)$$

and the offset is estimated by

$$\hat{\Delta} = \frac{1}{k} \sum_{i=1}^{k} (x_i - r_i) . \qquad (1.4.5)$$

Given the further assumption that all the random errors come from the same distribution, the magnitudes of the random errors can be quantified by a standard deviation (see Ku [28] for a clear and concise discussion of standard deviations).

Another less frequently assumed response for a calibrating instrument allows not only for instrumental offset ψ but also for a non-constant error that depends on the item being measured. This type of response is sometimes referred to as non-linear behavior, and in this case two reference standards with known values R_1^* and R_2^* are required to estimate X^*. Given measurements r_1 on the first standard and r_2 on the second standard, the instrument response for the three artifacts is described by:

[¶] The caret (^) over a symbol such as Δ denotes an estimate of the parameter from the data. It is dropped in future chapters where the intent is obvious.

$$\begin{aligned} x &= \psi + \beta X^* + \varepsilon_x \\ r_1 &= \psi + \beta R_1{}^* + \varepsilon_{r_1} \\ \text{and} \quad r_2 &= \psi + \beta R_2{}^* + \varepsilon_{r_2} \end{aligned} \qquad (1.4.6)$$

where the parameter β is non-trivial and different from one, and ε_x, ε_{r_1} and ε_{r_2} are independent random errors.

Then the measured differences $x-r_1$ and r_2-r_1 are used to construct an estimate of Δ, namely,

$$\hat{\Delta} = (R_2{}^* - R_1{}^*)\cdot(x - r_1)/(r_2 - r_1). \qquad (1.4.7)$$

The calibrated value of the test item is reported as

$$\hat{X}^* = \hat{\Delta} + R_1{}^*. \qquad (1.4.8)$$

Equivalently, Δ can be estimated by

$$\hat{\Delta} = (R_1{}^* - R_2{}^*)\cdot(x - r_2)/(r_1 - r_2)$$

in which case

$$\hat{X}^* = \Delta + R_2{}^*.$$

In order to achieve symmetry in the use of the reference standards, before and after readings, x_1 and x_2, can be taken on the test items with the readings in in order x_1, r_1, r_2, and x_2. Then Δ is estimated by

$$\hat{\Delta} = \frac{1}{2}(R_2{}^* - R_1{}^*)\cdot(x_1 - r_1 - r_2 + x_2)/(r_2 - r_1), \qquad (1.4.8a)$$

and the value for the test item is given by

$$\hat{X}^* = \hat{\Delta} + \frac{1}{2}(R_1{}^* + R_2{}^*).$$

In comparing the models in (1.4.2) and (1.4.6) one sees that the former model amounts to the slope β of the response curve of the instrument being identically one. If this slope is in fact close to one, which is certainly a reasonable assumption for most instruments, any departure from this assumption will contribute only a small systematic error in the assignment to the test item because of the small interval over which the measurements are taken. For this reason (1.4.2) is the commonly accepted model for calibration processes that use a comparator system of measurement.

The model in (1.4.6) amounts to a two-point calibration of the response function of the instrument; it is not dependent on a small calibration interval; and it is commonly used for direct-reading instruments. Notice that for either model a valid calibration for the test item does not depend on the response parameters of the instrument as long as they remain stable.

A multiplicative model for calibration assumes that the test item X and the reference standard R are related by

$$X^* = \gamma R^* \tag{1.4.9}$$

and that the measuring instrument has a response function of the form

$$x = \beta X^* + \varepsilon_x \tag{1.4.10}$$

$$r = \beta R^* + \varepsilon_r$$

where β and ε_x and ε_r are defined as before. The model leads to an estimate of γ; namely,

$$\hat{\gamma} = x/r \ . \tag{1.4.11}$$

The calibrated value of the test item is reported as

$$X^* = \hat{\gamma}R^*. \tag{1.4.12}$$

1.4.3 Models for Instrument Calibration

Models for instrument calibration relate the response of the instrument to a known stimulus called the independent variable. Where non-constant response of the instrument over a range of stimuli can be either theoretically or empirically related to the stimulus, the relationship is called a calibration curve.

The model for a calibration curve assumes that a response X is offset from a known stimulus W by an amount $\Delta(W)$ that depends on W and that the relationship holds over the entire calibration interval within a random error ε. A relationship of the form

$$X = \alpha + \beta W + \varepsilon \tag{1.4.13}$$

where α and β may be unknown parameters is called a linear calibration curve.

Once the parameters of the calibration curve are known or have been estimated by an experiment, future responses can be related back to their corresponding stimuli. In the general case this inversion is not easy nor is the attendant error analysis very tractable because the calibration curve is used in the reverse of the way that the data are fitted by least-squares.

The only case where the solution is straightforward is the linear case where a series of readings $X_j (j=1,\cdots,n)$ at designated points $W_j{}^* \ (j=1,\cdots,n)$ are used to obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters. The best estimate of offset for the linear case is

$$\hat{\Delta}(W) = \hat{\alpha} + \hat{\beta}(W). \tag{1.4.14}$$

Methods for estimating the parameters and quantifying the random error are discussed by Mandel [29].

1.5 Models for Error Analysis

The models in sections 1.4.2 and 1.4.3 admit random errors that come from a single error distribution whose standard deviation is of interest in quantifying the variability in the calibration process. We now expand this concept to models that include two distinct types of random errors; a random error term for short-term repetitions that is usually attributed to instrument variability and a random error term that allows for changes that are dependent on the conditions of the calibration and as such are assumed to remain constant for a given calibration. These two types of errors give rise to two distinct error distributions with associated standard deviations which can be estimated from the calibration data. The former is usually referred to as a "within" standard deviation and is designated by s_w.

The latter referred to as a "between" standard deviation, meaning between calibrations and designated by s_b, is attributed to changes in the calibration process from day-to-day. These include environmental changes that are not accounted for by modeling, changes in artifact alignment relative to the standard, and other fluctuations that are not reflected in the within standard deviation. For example, the model in (1.4.4) can be rewritten in terms of measured differences d_j $(j=1,\cdots,k)$ as

$$d_j = x_j - r_j = X^* - R^* + \varepsilon_j \qquad (1.5.1)$$

where the subscript j denotes short-term repetition and the ε_j are independent random errors that come from a distribution with standard deviation s_w. When this model is expanded to allow for day-to-day changes, the model becomes

$$d_j = (X^* + \delta_X) - (R^* + \delta_R) + \varepsilon_j \qquad (1.5.2)$$

where δ_X and δ_R are assumed to be independent random errors that come from a distribution with standard deviation s_b.

The quantities s_w and s_b, while of interest in their own right, are components of a "total" standard deviation that includes both "within" and "between" type variations in the measurement process. It is this total standard deviation, whose structure is discussed at length in this and later chapters, that is of primary interest in measurement assurance. The reader can verify that the proposed approach to error modeling is compatible with a components of variance model [30] by considering model (1.5.2) which leads to the estimate of offset given in (1.4.5). In terms of the error structure this offset is

$$\hat{\Delta} = (X^* - R^*) + (\delta_X - \delta_R) + \frac{1}{k}\sum_{j=1}^{k} \varepsilon_j \quad .$$

It can be shown[‡] that a reported value based on a single (k=1) measured difference has standard deviation

$$s_r = (2s_b^2 + s_w^2)^{1/2} \quad .$$

[‡] The methodology for arriving at the standard deviation is not explained in this publication. See Ku [28], pages 312-314, for the computation of standard deviations when several independent errors are involved.

A reported value based on the average of k short-term differences has standard deviation

$$s_r = (2s_b^{\ 2} + s_w^{\ 2}/k)^{1/2} .$$

Notice that the contribution of the component s_b to the standard deviation s_r is not reduced by taking multiple measurements that are closely spaced in time. This is the reason for discouraging short-term repetitions in measurement assurance and insisting that the definition of the total standard deviation encompass a broad range of operating conditions in the laboratory--implications which will be addressed in some detail in later chapters.

In this manuscript the total standard deviation s_c is defined to be the standard deviation of a "check standard" value as estimated from repeated calibration of the check standard. Where the error structure for the check standard value is the same as the error structure for the reported value of the test item, the standard deviation of the reported value which we call s_r, is exactly s_c. Otherwise, s_r must be adjusted accordingly. For example, suppose that a test item X with unknown value X^* is compared with two reference standards R_1 and R_2 with known values R_1^* and R_2^* by consecutive readings x_1, r_1, r_2, x_2 as described in section 4.2.

The error model for the measured differences

$$d_1 = x_1 - r_1$$

and

$$d_2 = x_2 - r_2$$

can be written as

$$\begin{aligned} d_1 &= (X^* + \delta_1) - (R_1^* + \delta_2) + \varepsilon_1 \\ d_2 &= (X^* + \delta_3) - (R_2^* + \delta_4) + \varepsilon_2 \end{aligned} \qquad (1.5.3)$$

where it is assumed that δ_1, δ_2, δ_3 and δ_4 have standard deviation s_b and ε_1 and ε_2 have standard deviation s_w.

The offset is estimated by

$$\hat{\Delta} = \frac{1}{2}(d_1 + d_2) \qquad (1.5.4)$$

and in terms of the error model

$$\hat{\Delta} = X^* - \frac{1}{2}(R_1^* + R_2^*) + \frac{1}{2}(\delta_1 - \delta_2 + \delta_3 - \delta_4 + \varepsilon_1 + \varepsilon_2). \qquad (1.5.5)$$

A check standard defined as the difference between R_1 and R_2 is computed for each calibration by

$$c = (d_2 - d_1). \qquad (1.5.6)$$

In terms of the errors the check standard measurement can be written

$$c \;=\; (R_1{}^* - R_2{}^*) + (-\delta_1 + \delta_2 + \delta_3 - \delta_4 - \varepsilon_1 + \varepsilon_2) \qquad (1.5.7)$$

The error model (1.5.5) for the reported value

$$X^* = \hat{\Delta} + \frac{1}{2}\,(R_1{}^* + R_2{}^*), \qquad (1.5.8)$$

and the error model (1.5.7) for the check standard measurment c are comprised of the same error terms and differ structurally by a factor of two.

Explicitly, the standard deviation of the reported value X^* is

$$s_r = \frac{1}{2}\,(4s_b{}^2 + 2s_w{}^2)^{1/2} \qquad (1.5.9)$$

and the standard deviation of c is

$$s_c = (4s_b{}^2 + 2s_w{}^2)^{1/2} . \qquad (1.5.10)$$

Therefore,

$$s_r = \frac{s_c}{2} \qquad (1.5.11)$$

In practice s_c is estimated by check standard measurements from many calibrations (see chapter 4), and this estimate is used in (1.5.11) to compute s_r.

Where the check standard value is a least-squares estimate from a design or a function of measurements on more than one artifact, the computation of the standard deviation of a reported value is more complicated. In such a case, one must first estimate s_w from a single calibration and compute s_b from an equation for s_c such as (1.5.10). Then the standard deviation of the reported value can be computed from an equation such as (1.5.9).

1.6 NBS Role in the Development of a Measurement Assurance Program

1.6.1 Study of Operations at Participating Laboratories

Before undertaking the development of a measurement assurance program for disseminating a unit of measurement, NBS technical staff familiarize themselves with operations at potential user laboratories so that the program can be structured around the equipment, facilities and personnel available to the laboratories. Suggestions for equipment modifications and additions are made at this time. The range of operating conditions in the participating laboratories is checked for consistency with the model, and in order to determine whether or not the accuracy goals of the measurement assurance program are attainable, NBS is advised of the individual laboratory's measurement requirements and capabilities.

1.6.2 Identification of Factors Capable of Perturbing the System

It is the responsibility of NBS to identify and isolate those factors capable of seriously disrupting the measurement system so that equipment and procedures can be designed to offset the impact of such factors (Youden [31]). This is particularly important if the measurement assurance program is intended for an industrial setting rather than a controlled laboratory setting.

An example of this type of testing, called "ruggedness" testing is found in the NBS flowmeter program for liquids (Mattingly et al [32]). The effects of three types of perturbation on turbine meters were studied experimentally, and it was found that the profile of the flow entering the meter has a significant effect on meter performance. This research led to the development of a flow conditioner which can be inserted in an upstream section of pipe to regulate the profile of the flow entering the meter. Because flow profiles vary from laboratory to laboratory depending on the source of the flow, such a flow conditioner is appended to the turbine meters that are circulated in the industry as NBS transfer standards.

1.6.3 Design of Interlaboratory Exchanges

The purpose of the interlaboratory study or round-robin test that is usually sponsored by NBS at the inception of a measurement assurance program is to determine the extent and size of offsets from NBS that are typical in the target industry. Secondary goals are the evaluation of the adequacy of proposed procedures for resolving the measurement problem, critique of the format and content of directions from NBS, and study of the ease of implementation on the part of participants. Frequently a preliminary interlaboratory test designed to identify significant problem areas is followed by a more comprehensive study which incorporates modifications to artifacts and protocols based on experience gained in the preliminary test.

1.6.4 Development of a Stable Transfer Standard or Standard Reference Material

Either a standard reference material or a transfer standard is developed for each measurement assurance program that is sponsored by NBS. The standard reference material (SRM) is a stable artifact produced either commercially or in-house that is calibrated, certified and sold by NBS in fairly large numbers.[d] Standard reference materials are well known for chemical applications. Recently NBS has certified two separate dimensional artifact standards as SRMs, one a linewidth standard for the integrated circuit industry [NBS SRM-474] and the other a magnification standard for scanning electron microscopes [NBS SRM-484]. An SRM has the unique property that it can be used not only for determining offset from NBS but also as an in-house standard for controlling the measurement process.

[d] A listing of SRM's is contained in the catalog of NBS Standard Reference Materials, NBS Special Publication 260, 1979-80 Edition, available from the Office of Standard Reference Materials, NBS, Gaithersburg, MD.

The transfer standard is a calibrated artifact or instrument standard that is used for disseminating the unit of measurement. It is loaned to the participant to be intercompared with the participant's standards or instrumentation under normal operating conditions in order to determine offset from NBS.

Artifacts that are stable with relation to a physical quantity, such as the mass of an object, do not usually pose any special problems when they are used as transfer standards because they can be shipped from one place to another without a change in the quantity of interest. Transfer standards that are not easily transported are packaged in environmentally controlled containers, but additional redundancy in the form of multiple standards and observations is always included in the measurement assurance program whenever the stability of the transfer standard is in question.

1.6.5 Dissemination of Measurement Technology and Documentation

The participant in a measurement assurance program is entitled to draw upon the expertise and experience that resides in the sponsoring NBS technical group. Technical assistance is disseminated by way of NBS publications, ASTM, standards, ANSI standards, laboratory visits, telephone conversations and NBS sponsored seminars. In conjunction with the advent of a new program a series of seminars is usually offered to the public to explain the philosophy, theory, measurement technology and statistical analyses which form the basis for a measurement assurance program in that discipline.

Documentation for standard reference materials is available through NBS Special Publication Series 260. As part of a long range plan to upgrade its calibration services, the National Bureau of Standards has instituted documentation requirements for all calibration services. Documentation includes theory, laboratory setup and practice, measurement technique, maintenance of standards, specification of measurement sequence, protocol for measurement control and determination of final uncertainty. When these publications become available, they will provide the bulk of the documentation that is needed for implementing a measurement assurance program that is related to an NBS calibration service. Insofar as a measurement assurance program as implemented by the participant may differ from the NBS calibration program in regard to the number of standards, specification of measurement sequence, corrections for environmental conditions, estimation of process parameters, and methods for determining offset and uncertainty, additional user oriented documentation may be made available.

1.6.6 Establishment of Measurement Protocol for Intercomparisons with NBS

Measurement assurance programs currently in existence fall into two categories. The first category contains those services which are highly structured for the participant, with regard to the number of laboratory standards to be employed in the transfer with NBS, the number of repetitions to be made in the exchange, and the protocol to be used for establishing an in-house measurement control program. At this time only the Gage Block Measurement Assurance Program (Croarkin et al [33]) and the Mass Measurement Assurance Program fall into this category.

All other programs allow the participant considerable leeway in regard to the items mentioned above in order to make the service compatible with the unique situation in each laboratory. The advantage of operating within the constraints of equipment and staff resources that are already allocated to the laboratory's normal workload is obvious, especially where accuracy requirements are not difficult to meet. However, there are drawbacks. The data analysis must be tailored to each participant, imposing an additional burden on NBS staff, and responsibility for instituting a rigorous measurement control program is left entirely to the participant.

1.6.7 Data Analyses and Determination of Offset

The determination of offset and associated uncertainty as realized by intercomparison of laboratory reference standards with NBS transfer standards is accomplished in one of two ways:

i) The transfer standard(s) is sent to the participant as a blind sample, and the data from the intercomparison are transmitted to NBS. Based upon the value assigned to the transfer standard by NBS and associated uncertainty from the NBS process, new values with associated uncertainties are assigned to the laboratory standards along with the uncertainty that is appropriate for an item measured by the participant's process.

ii) the transfer standard along with the its assigned value and associated uncertainty are transmitted to the participant, and the analyses and determination of offset become the responsibility of the participant.

Data analyses relating to the regular workload and measurement control procedures in a laboratory are best left to the individual participant. These analyses provide important insights into the pecularities of a measurement process, and, consequently, these analysis are best done internally. Even where much or all of the data analysis is undertaken by NBS, participants are encouraged to develop facility in this area in order to make themselves independent from NBS in the future. Some participants in measurement assurance programs have automated the analysis of calibration data, decisions relating to process control, updating of data files and final determination of uncertainty on minicomputers in their laboratories.

1.7 Participant's Role in a Measurement Assurance Program

1.7.1 Staff Preparation

The success of a properly conceived measurement assurance program depends upon the enthusiasm and dedication of the personnel who are making the measurements and resolving problems that arise in day-to-day operations. The measurement assurance approach is a long-term commitment in terms of evolving a measurement control technique that continually checks on the state of control of the process. Before undertaking such a program, there should be reasonable assurance of continuity of personnel assigned to the project, and steps should be taken to guarantee that new personnel are sufficiently prepared for taking on the assignment before the departure of experienced personnel.

The success of such a program also depends on a certain depth of understanding on the part of the staff. Here we are talking not so much about the intricacies of a particular analysis, but about a basic understanding of scientific methodology, the philosophy of measurement assurance, and the relationship between the control techniques and the validity of the values reported by the measurement process and their associated uncertainties. To this end, NBS offers seminars in which the attendees are instructed in these principles, but some prior staff preparation may be necessary in order to benefit fully from these expositions. Courses at local community colleges are recommended for exploring scientific principles and gaining facility with fundamental mathematical and statistical manipulations.

1.7.2 Selection of a Check Standard

The selection of a check standard must be considered in the preliminary planning for measurement assurance program. In short, its purpose is to provide a continuing thread that characterizes the operation of the measurement process over changing laboratory conditions and over time with regard to both the variability of the process and the long-term average of the process. It is a basic tenet of measurement assurance that the response of the process to the check standard be sufficiently similar to the response of the process to the test items that the performance of the process at all times can be adequately monitored by monitoring the response of the process to the check standard. The value of the check standard at any given time is a decision-making tool, and unexpected behavior on its part is grounds for discontinuing the process until statistical control is resumed.

Careful consideration should be given to the type of artifact that would be suitable for this purpose. It should certainly be of the same character as the items that constitute the workload in the laboratory. For some processes, such as processes dealing with basic units of measurement, the selection is obvious; check standard artifacts are similar to reference standards in design and quality. In general, an artifact that is less stable than the reference standards will not be useful as a check standard if its instability is large enough to mask the properties of the measurement process.

The check standard should be thought of not so much as an artifact but as a data base because it is the measurements that are of interest and not the artifact per se. The check standard data base consists of measurements, properly corrected for environmental factors, or some function of those measurements that have been made on the artifact check standard or on the reference standards. For example, a test item that is compared to two reference standards has its assignment based on the average of the values assigned to the two reference standards. The check standard can be defined to be the difference between the measurements on the reference standards thus eliminating the need for an extraneous measurement or other artifact. Where a calibration involves only one reference standard, an artifact that is similar in response to the test items can be designated as the artifact check standard. This need not be a calibrated artifact, and the properties of the measurement process are ascribed to it as long as it is measured in the same time frame as the other items in the calibration process. Several check standards used separately or in combination may be employed when the stability of the reference standards, such as a bank of standard cells, is cause for concern.

Where reference standards exist at several levels, such as mass standards or length standards, check standards are maintained and monitored at each level. Where the quantity of interest is propagated over several levels from one standard such as a one ohm resistor, which is used to propagate resistances between one and ten ohms, the same check standard artifact may be employed at the different levels, but the data bases for the different levels are regarded as separate check standards.

An SRM makes an ideal check standard if it is not contaminated or otherwise degraded by heavy usage. In any case the artifact or artifacts on which the check standard base is built must be readily available to the measurement process over a long period of time.

The proliferation of check standards involves no small amount of work in maintaining the data base, and serious thought should be given to placement of check standards in the measurement echelon. For a new program, one should start with check standards at a few critical points and gradually increase these as experience is gained with the program.

1.7.3 Initial Experiments to Estimate Process Parameters

The establishment of an initial data base for the laboratory's check standards is the first order of business in a new measurement assurance program. Before one attempts to quantify offset, it must be demonstrated that a measurement process does in fact exist; i.e., that measurements from the process satisfy the requirements for statistical control. This presupposes that the process precision is well known and that this can be documented. If, in fact, the documentation of the process has been lax, or if a substantially new process has been instituted for the measurement assurance program, then measurements taken over as long a time period as practical should be made on the check standard(s) in order to estimate the long-term average of the process and the standard deviation. Procedures for obtaining these initial estimates are discussed in subsequent chapters.

A laboratory planning a transfer with NBS should undertake these experiments well in advance of the arrival of the NBS transfer standard so that any problems encountered in the measuring system can be rectified. This provides a shake-down period for procedures, equipment and software involved in the measurement assurance program. Once the transfer standards are intercompared with the laboratory's reference standards, the resulting measurements involving the check standard are compared with the initial data base to decide if the process is in control at that time, and the transfer between the laboratory process and the NBS process is accomplished only if the process is judged in control. Therefore, participants are urged to make the initial experiments as representative of laboratory conditions as possible and to request help from the sponsoring NBS group if measurement problems or procedural ambiguities exist so that delays with the transfer can be avoided.

1.7.4 Calibration Procedures

Accomodation to measurement assurance principles can mandate a change in calibration procedures within the laboratory. Most often such change will amount to additional redundancy in the design and/or change in the order of measurements. The laboratory should settle upon one calibration design for the transfer with NBS and the calibration workload. There is considerable advantage in doing this because the uncertainty determined from the transfer with NBS is only valid for that measurement process, and if the uncertainty is to have validity for the workload, the two measurement processes must be identical. There is a further advantage; the same statistical control program will suffice for both processes, and the check standard measurements from both sources can be combined into a single data base.

Another consideration is the manner in which systematic error is handled in the transfer experiment. Some measurement assurance programs are structured so that the determination of systematic error is made relative to the average of two or more reference standards as in section 4.2.4. For example, two reference gage blocks can be calibrated by intercomparison with two NBS transfer blocks by a design that assigns values relative to the average of the two reference blocks called the restraint. Systematic error is estimated as the difference between the restraint and the average computed for the two NBS blocks by the transfer experiment. The laboratory's restraint is then corrected for this offset. Meaningful values cannot be computed for the reference standards individually from the transfer experiment. Thus, the same design that is used for the transfer with NBS is employed in the calibration workload so that all assignments are made relative to the corrected restraint.

1.7.5 Process Control

The measurement assurance concept demands that a value be assigned to an artifact only when the measurement process is in control in order to guarantee the validity of the assignment and associated uncertainty statement. This means that statistical control is employed in the everyday workload of the laboratory as well as during the transfer with NBS. For highest accuracy work, comparable to calibrations at NBS, a check for control is made during every measuring sequence in which an artifact is calibrated by the system. Statistical control procedures based on check standard measurements along with the appropriate statistical tests are discussed in section 3.3.

The choice of a control procedure and its implementation are the responsibility of the participant. Those who are familiar with industrial quality control procedures and Shewhart type control charts should be able to adapt these methodologies to check standard measurements. A general discussion of control charts with examples is contained in chapter 5, and statistical control procedures for specific measurement situations are outlined in chapter 4.

1.7.6 Data Base Maintenance

A record of check standard measurements is kept separately from other laboratory records such as records of past calibrations. This permanent record should include all pertinent information relating to the measurement. For example, it normally includes an identification for the check standard, identification for the instrument, identification for the operator, day, month, year, identification for the type of statistical design used in the intercomparison, observed value of the check standard, environmental conditions that could affect the measurement such as temperature, pressure and relative humidity, standard deviation if applicable, and finally a flag denoting whether or not the check standard was in control on that occasion.

2. Characterization of Error

2.1 Introduction

It is the purpose of this chapter to introduce the reader to the concepts of random error, systematic error and uncertainty. It is the expressed purpose of measurement assurance to identify and quantify all sources of error in the measurement process, because in so doing, the worth of any value reported by the process can be stated in quantitative terms called an uncertainty. In a very real sense, a value assigned to an artifact only has meaning when there is an assessment of how well that number describes the property of interest (be it length, mass or whatever) in terms of its reference base. An uncertainty statement provides that assessment.

Error in measurement is categorized as either systematic, coming from a source that is constant and ever present in the measurement process, or random, coming from a source (or sources) that is continually fluctuating. Systematic error may be known or estimable for any given situation, but random error by its nature is never known for a given measurement situation. The point is that for a single measurement it may be possible to determine the size of the systematic error by intercomparison. On the other hand, the random error that is unique to a single measurement cannot be replicated because conditions of measurement cannot be repeated exactly. Therefore, it is common practice in metrology, as it is in process control [34], to quote limits to random error for all such experiments.

Classification of sources of error as either systematic or random is not always straightforward depending as it does on the way in which the potential source of error enters the measurement process, how it affects the output of that process, and the interpretation of the uncertainty. For example, the maximum observed difference between operators can define a systematic error for a system that is highly operator dependent and for which there are a restricted number of operators or, alternatively, a separate uncertainty statement can be issued for each operator's measurements. Measurement systems that routinely make use of many operators are better served by folding the effect of operator error into the total random error for that system.

At the National Bureau of Standards considerable attention is given to the classification of sources of error. For the participant in a measurement assurance program, systematic error is usually assumed to come from specific sources that are spelled out in this chapter, and remaining sources of error are assumed to be part of the random error of the participant's process and must be estimated as such.

2.2 Process Precision and Random Error

2.2.1 The Standard Deviation

A "measurement process" is said to exist for quantifying a physical attribute of an object, such as its length, only if the process is operating in a state-of-control (Eisenhart [35]). The fact is that, even for such a process, repeated measurements on the same object will not produce identical results. As long as the source of this disagreement is random in nature; i.e., its direction and magnitude not being predictable for any future measurement, the disagreement among measurements is referred to as the process imprecision. A measure of precision, such as the process standard deviation, quantifies this random error or scatter or, more aptly, describes the degree of agreement or closeness among successive measurements of the same object.

The term process precision as used in this publication is not limited to the characterization of the behavior of the particular measuring device per se, but it is intended to describe the total configuration of operator, environmental conditions, instrumentation and whatever other variables go into making any given measurement. As it is rarely possible to measure an item submitted for calibration over a representative set of environmental and working conditions in the laboratory, redundancy is obtained from measurements made on a check standard that is introduced into the measurement sequence on a routine basis. It is assumed that the check standard is similar in response to the test item and that the process precision can be estimated from the measurements made on the check standard.

The simplest measure of process precision is the range--the difference between the largest and smallest measurements in the group. The range is a satisfactory measure of precision when the number of measurements is small, say less than ten. It "does not enjoy the desirable property" (Ku [36]) of tending toward a limiting value as more measurements are taken; it can only increase and not decrease. Therefore, it is desirable to find a measure of precision which takes into account the information in all the measurements and which tends to a limiting value as the sample size increases if we are to use this measure to describe the process behavior as a stable phenomenon.

The standard deviation is such a measure. Small values for the standard deviation are indicative of good agreement and large values are indicative of poor agreement. Because it is necessary to distinguish different kinds of variability that contribute to the total process variability, it is likewise necessary to define different kinds of standard deviations. We routinely identify two levels of standard deviations in calibration work.

These two levels are described briefly in the first chapter where we are dealing with the models covering the error structure among measurements. Reiterating, the first type of standard deviation is a measure of the variability of the measurement process over a short period of time, usually the time necessary to complete one calibration using a particular sequence of measurements called a statistical design. This measure is called the "within standard deviation." Its usage as a check on the internal consistency of an individual calibration experiment is explained in chapter 3 and chapter 4 along with formulas and examples.

The second type of standard deviation that we are dealing with in measurement assurance, and by far the more important of the two, is the total standard deviation s_c. This latter measure includes both the "within" component of variability s_w and a "between" component of variability s_b, which the reader will recall explains incremental changes that can take place from calibration to calibration. The relationship among these quantites is assumed to be of the form

$$s_c = (s_w{}^2 + s_b{}^2)^{1/2} \; .$$

Therefore, the total standard deviation, including as it does both "within" and "between" components of variability, should accurately reflect both the short-term and long-term random errors that are affecting the measurement process.

The limits to random error quoted in the uncertainty statement are computed from the total standard deviation thus assuring that the conditions of a single calibration do not invalidate this measure of the quality of the reported value. As has been noted previously, the total standard deviation, not generally being available from the calibration data, is based on repeated check standard measurements that are structured to include all possible sources of random error. This is accomplished by monitoring the check standard over a long period of time and over the full range of environmental factors for which the uncertainty statement is assumed to be valid.

The total standard deviation depends on the physical model. The most familiar form

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - \overline{c})^2 \right)^{1/2} \qquad (2.2.1)$$

where the arithmetic mean is

$$\overline{c} = \frac{1}{n} \sum_{i=1}^{n} c_i \qquad (2.2.2)$$

assumes that check standard measurements $c_1,\cdots,c_n$ are independent of time and that the effect of other variables is negligible.

The term (n-1), called the degrees of freedom associated with s, is an indication of the amount of information in the standard deviation and is always reported along with the standard deviation.

2.2.2 Pooled Standard Deviation

If several standard deviations with small numbers of degrees of freedom are computed from the same process, they will vary considerably among themselves. It goes without saying that the standard deviation that is quoted in the uncertainty statement must have a sufficient amount of information to guarantee that it is a valid measure of process precision. The question is, "How much redundancy is sufficient?" As a general rule, fifteen degrees of freedom is a minimum for the initial computation of the standard deviation. As the measurement assurance program progresses, the standard deviation is recomputed to take advantage of the increased data base, and assuming that the process is stable, this will assure a more reliable value of the standard deviation. A standard deviation based on as few as two data points can be combined with other similar estimates that have been obtained on separate occasions for the same process to obtain what is called a "pooled" standard deviation. If the individual standard deviations are $s_1, \cdots, s_k$ with degrees of freedom $\nu_1, \cdots, \nu_k$, respectively, the pooled standard deviation is

$$s_p = \left(\frac{\nu_1 s_1{}^2 + \cdots + \nu_k s_k{}^2}{\nu_1 + \cdots + \nu_k} \right)^{1/2} . \qquad (2.2.3)$$

The degrees of freedom associated with s_p is $\nu = \nu_1 + \cdots + \nu_k$.

2.2.3 Limits to Random Error

Limits to random error can be computed with a given probability if the distribution of random errors is known. Limits, so stated, depend upon assumptions concerning the average value and spread of the underlying distribution. For a calibration process it is assumed that random errors of measurement have an equal chance of being negative or positive such that their average value is zero. It is also assumed that the spread of the distribution is adequately estimated by the total process standard deviation.

Limits to random error for a single value from the measurement process are constructed so that the probability is $(1-\alpha)$, for α chosen suitably small, that if the measurement algorithm were to be repeated many times, the average outcome of these experiments would fall within $\pm\ s_c \cdot t_{\alpha/2}(\nu)$ of the reported value, where s_c is the total process standard deviation, ν is the number of degrees of freedom in s_c, and $t_{\alpha/2}(\nu)$ is the $\alpha/2$ percent point of Student's t distribution. (See Ku [37] for a further discussion of Student's t distribution.) Critical values for Student's t are given in Table I for $\alpha = 0.05$ and $\alpha = 0.01$ and degrees of freedom $\nu = 2(2)120$.

Frequently a precise probability interpretation for the limits to error is not needed and, in fact, will not be possible if it cannot be demonstrated that the underlying probability distribution for the data is exactly a normal distribution. In metrology the limits to random error are often taken to be three times the standard deviation. Other technical areas may use two standard deviations. The bounds, plus and minus three standard deviations, are statistically robust (with respect to the coverage of the distribution) in that if the experiment were to be repeated, the chance of reporting a value outside of these bounds would be extremely small. This, of course, assumes that the random errors affecting the experiment come from a distribution that is close in character to the normal distribution and that enough data have been collected to provide a reliable estimate of the standard deviation. The examples given in this chapter use three standard deviation limits.

2.3 Systematic Error

2.3.1 Conventional Calibration

Systematic error takes into account those sources of error, peculiar to the measurement system, that remain constant during the calibration process and explain a difference in results, say, between two different measuring systems trying to realize the same quantity through a large number of measurements. Some obvious examples are: uncertainties in values assumed for reference standards, uncertainties related to the geometry or alignment of instrumentation, differences between operators, differences between comparable systems, etc. The size of the possible discrepancy is estimated, either empirically or theoretically, but its direction is not always known.

In order to define systematic error for a calibration process, it is necessary to define the steps in a calibration echelon that relate the measured value of the quantity of interest back to its basic SI unit or to a national standard. NBS, except in the case of international comparisons, occupies the premier position in the U.S. calibration echelon. Thus the first transfer point in this calibration echelon involves the intercomparison of a laboratory reference standard with the national standard maintained by NBS which may be an artifact standard or an instrument. The second transfer point involves the intercomparison of the laboratory reference standard with an unknown which in turn can be a working standard from the same laboratory or an artifact standard from a lower level calibration laboratory or a finished product. The calibration chain is extended in this way until the final product has been calibrated by an intercomparison involving it and a standard which can be traced back to the National Bureau of Standards.

Systematic error is assessed at every transfer point and passed along to the next lower level in the calibration chain. Thus, the total systematic error for the measurement process that delivers the final product is an aggregate of systematic errors from all transfer points. Systematic error must be defined very specifically for each transfer point in terms of the long-term values for measurements from two systems, and it must also include an estimate of the amount by which the higher level system, such as NBS, may be in error in estimating its long-term value.

The purpose of each transfer point is to reduce or eliminate systematic errors at that level. If we look at an exchange between a laboratory and NBS, a potentially large source of systematic error comes from the values assigned to the laboratory's reference standards. Calibration of the reference standards at NBS can eliminate offset from this source, but the calibration itself is still a source of systematic error whose magnitude depends on how well NBS was able to conduct the calibration as measured by the uncertainty associated with the calibrated values.

The rationalization for assessing a systematic error from this source is that the values for the reference standards remain constant as they are used as a reference for assigning values to other artifacts or instruments. At least they remain constant until they are recalibrated at NBS, and the assignments resulting from their use are all affected in the same way, being either too low or too high, even though the direction and exact magnitude of this error are not known. Thus, uncertainties for values of reference standards are regarded as a systematic error in the laboratory's process (Youden [40]).

Systematic error associated with the uncertainty of a reference standard is assessed proportional to the nominal value of the test item and the nominal value of the reference standard. For example, if a one kilogram standard is used in a weighing design to calibrate a 500g weight, the systematic error from this source is one-half of the uncertainty associated with the assignment for the kilogram standard.

If the value for a test item is reported relative to the average of two reference standards R_1 and R_2, all artifacts being of the same nominal size, and if the assignments for R_1 and R_2 are independent, the systematic error from this source is assessed as

$$U = \frac{1}{2}\left(U_{R1}{}^2 + U_{R2}{}^2\right)^{1/2}$$

where U_{R1} and U_{R2} are the uncertainties for R_1 and R_2 respectively. Where the assignments to R_1 and R_2 are not done independently

$$U = (U_{R1} + U_{R2})/2.$$

2.3.2 Measurement Assurance Approach

A laboratory participating in a measurement assurance program measures a transfer standard(s) from NBS as if it were an unknown item using the reference standards and instrumentation that constitute the measurement system in that laboratory. The resulting value for the transfer standard, be it based on one measurement or on several repetitions in the laboratory, is compared with the value assigned the transfer standard by NBS. The relationship between the laboratory's assignment and the NBS assignment for the transfer standard defines an offset which is used to correct the values for the laboratory's reference standards.

This approach has an advantage over the usual calibration route as far as identifying systematic error in the laboratory. Either method suffices for identifying errors related to the values of the reference standards, but given that the reference standards are properly calibrated, the particular conditions of their usage in the laboratory may invite systematic errors that are unsuspected and unidentifiable. The dependence of optical systems on operator was mentioned in an earlier chapter, and systematic error caused by operator effect may be significant for other types of systems as well. Also, instrumentation can differ enough that the reference standards alone are not sufficient for eliminating systematic error. Of course, both of these sources of systematic error might be identifiable by proper experimentation, but it would be difficult to assess the magnitude of such errors without the measurement assurance program. Other factors that are probably not identifiable within the laboratory itself are systematic errors related to lack of proper environmental control or incorrect measurement of temperature and humidity.

Two sources of systematic error are always present in a measurement assurance program. The uncertainty associated with the value of a transfer standard is one. Because another transfer point has been effectively added to the calibration chain, the limits to random error associated with the transfer measurements in the participating laboratory define another systematic error for the laboratory.

2.3.3 Calibration Curve

A more complex situation arises when the purpose of the program is to calibrate an instrument over a range for all continuous values. In this case transfer artifacts are provided at selected points covering the range of interest, and the intercomparisons are used to establish a functional relationship between the instrument and the NBS system. The assignment of values is based on this functional relationship. For example, systematic errors in linewidth measurements produced by an optical imaging system can be reduced relative to the NBS prototype optical system [38] from measurements made on an NBS dimensional artifact. (This artifact is a glass substrate with a series of chromium lines at spacings spanning the range of interest.)

Measurements made on individual lines on the artifact define a functional relationship between the two systems, and a least-squares technique is used to derive a best fitting curve to the measured values as a function of the NBS values. The empirical fit is called the calibration curve.

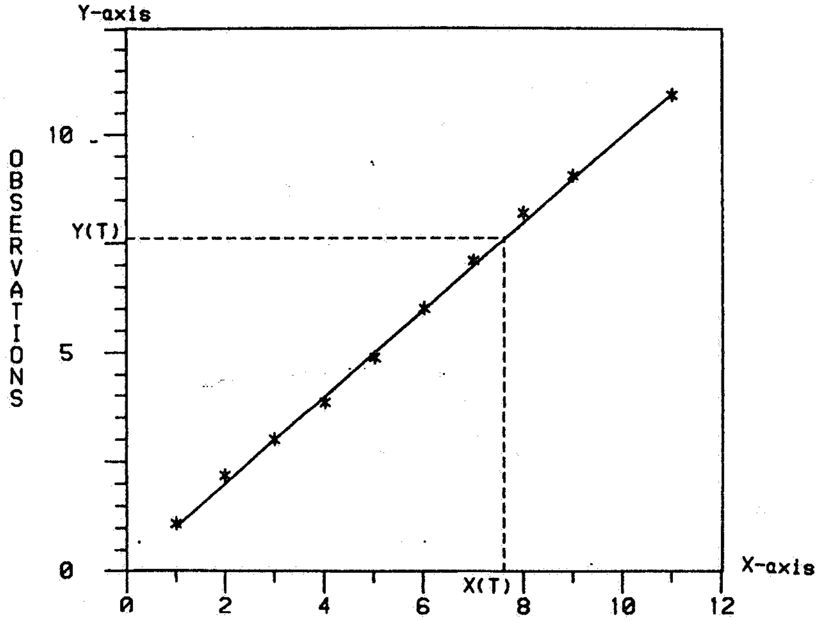


Figure 1
Schematic diagram of a linear calibration curve showing the relationship between an observed value Y(T) and its calibrated value X(T)

In figure 1, each optical measurement is plotted against the corresponding NBS value, and the calibration curve fitted to all the measurements is shown by the solid line. The offset between the user's system and the NBS system is reduced by relating any future measurement back to the NBS value. Schematically, for a future value Y(T) as shown on the Y-axis, a dotted line is drawn through Y(T) parallel to the X-axis. At the point where it intersects the calibration curve another dotted line is drawn parallel to the Y-axis, and its point of intersection on the X-axis, X(T), is the corresponding calibrated value relative to NBS.

Because the functional relationship is not known exactly but is estimated by a series of measurements, the calibration curve can be in error. A discussion of the effect of this error on the final uncertainty of a calibrated value is beyond the scope of this treatise. The reader is referred to Hockersmith and Ku [39] for a discussion relating to quadratic calibration curves and to Croarkin and Varner [40] for a discussion relating to linear calibration curves.

2.4 Uncertainty

2.4.1 Definition

The uncertainty statement assigns credible limits to the accuracy of the reported value stating to what extent that value may differ from its reference base. In practice it quantifies the magnitude of any possible discrepancy between the value actually obtained in the laboratory and the value which would be obtained at NBS for the same property of an object. An uncertainty provides both a measure of the worth of the values reported by the measurement laboratory and an estimate of the systematic error accruing to any organization that makes use of these values.

The uncertainty statement is composed of i) all sources of sytematic error that contribute to the offset from the reference base and ii) a limit to random error that quantifies the variability that is inherent in the measurement process as it transfers from a "known" or calibrated artifact or measurement system to an "unknown".

2.4.2 Combination of Random and Systematic Error

Once the systematic errors and the limits to random error have been estimated, they are combined into a single number which is called the uncertainty. Much controvery arises over the proper way to combine systematic and random errors in an uncertainty statement. Basic premises concerning measurement and its uncertainty as espoused by Youden [41], Eisenhart et al. [42] and others have long been adopted by NBS calibration services and are recommended for measurement assurance programs. A different philosophy that has recently been advanced by the Bureau International des Poids et Mesures is discussed in reference [43]. Basically the question revolves around whether systematic errors should be added linearly or combined in quadrature and around whether the systematic error and the limit to random error should be added linearly or combined in quadrature. For example, if there are several sources of systematic error $S_1, \cdots, S_k$, adding the systematic errors linearly assumes the worst possible combination of errors and gives a total systematic error of total systematic error S where

$$S = S_1 + S_2 + \cdots + S_k \,. \qquad (2.4.1)$$

Combining the systematic errors in quadrature produces a total systematic error for those sources of

$$S = (S_1{}^2 + S_2{}^2 + \cdots + S_k{}^2)^{1/2} \,. \qquad (2.4.2)$$

Recommended practice for measurement assurance programs is to combine in quadrature systematic errors that are known to be independent as in (2.4.2), to add linearly systematic errors that may not be independent as in (2.4.1), and to combine systematic and random errors linearly.

2.4.3 Final Statement

Because there is no universal agreement on setting limits to random error, such as two or three standard deviation limits, and also because there is no universal agreement either at NBS or internationally as to how the systematic and random components should be combined, it is recommended that for maximum clarity the composition of the uncertainty statement be fully explained. The explanation should include a statement of the limits to random error, a list of sources of systematic error, and a description of the way in which they have been combined. An example of an uncertainty statement from an NBS calibration process is:

> The apparent mass correction for the nominal 10 gram weight is +0.583mg with an overall uncertainty of ±0.042mg, using three times the standard deviation of the reported value as a limit to the effect of random errors of measurement, the magnitude of systematic errors from all known sources being negligible.

The chain of uncertainty as propagated through a calibration echelon starts with the uncertainty assessed at NBS which consists of all sources of error, both systematic and random, associated with that process including the uncertainty of its reference standards relative to basic units of measurements. If the calibration echelon involves one or more standards laboratories, the total uncertainty as assessed at each echelon becomes a systematic error for the next lower echelon laboratory, and the uncertainties at each level are propagated in like manner. In the next section the propagation of uncertainties for a laboratory that uses an NBS calibrated artifact as a reference standard is compared with the propagation of uncertainties for a laboratory that calibrates its own measuring system through the use of an NBS transfer standard.

2.5 Uncertainty of Reported Values

2.5.1 Uncertainty via Conventional Calibration

The uncertainty associated with a value reported for a test item by a measurement process that is operating in a state of statistical control using a reference standard calibrated by NBS is

$$U = 3s_r + U_{STD} \cdot \qquad (2.5.1)$$

This assumes that the standard is not changed during transport and that environmental and procedural factors are not different from the conditions of calibration. The standard deviation of the reported value s_r depends on the total standard deviation s_c, the error structure for the reported value as discussed in section 1.5, and the number of measurements made on the test item. The quantity U_{STD} is the uncertainty associated with the reference standard as stated in the NBS calibration report.

Note that where the reported value is an average of p measurements, the usual standard deviation of an average, $s_r/\sqrt{p}$, sometimes called the standard error, will apply to the reported value only if the p repetitions were made over the same set of environmental conditions that were sampled in the calculation of the total standard deviation. In a calibration setting where repetitions are done within a day or two, the standard deviation of a reported value depends upon a between component of variability s_b and a within component s_w as explained in section 1.5.

2.5.2 Uncertainty via a Transfer Standard

Where a laboratory has calibrated its own reference standard using an NBS transfer standard, rather than using a reference standard calibrated at NBS, another echelon has effectively been added to the calibration chain. The uncertainty of that transfer must be assessed, and it contributes another systematic error to the process of subsequently assigning values to test items.

The uncertainty of a transfer involving a single transfer standard compared with a single laboratory standard is

$$U_{tr} = 3s_t + U_T \qquad (2.5.2)$$

and the uncertainty associated with a value reported for a test item is

$$U = 3s_r + 3s_t + U_T = 3s_r + U_{tr} \qquad (2.5.3)$$

where s_r is the standard deviation associated with the reported value of the test item as discussed in the last section; s_t is the standard deviation associated with the value assigned to the laboratory's reference standard via measurements made on the transfer standard; and U_T is the uncertainty assigned to the transfer standard by NBS.

Admittedly there can be some concern about qualifying a laboratory's systematic error by means of an NBS transfer standard because of the additional systematic error that this imposes on the uncertainty statement. This fact is inescapable, but the resulting uncertainty statement is, in fact, a realistic expression of the errors affecting the process whereas the usual calibration route does not provide a way of assessing systematic errors that may be affecting measurements, other than those directly involving the artifact standard.

The uncertainty, U_T, associated with a transfer standard will usually be smaller than U_{STD}, the uncertainty associated with a calibrated artifact. The calibration workload at NBS is at least one step removed from the NBS primary standard, and the size of U_T relative to U_{STD} can be reduced by eliminating this step in assignments to transfer standards. For example, transfer standards for voltage measurements are compared directly to an NBS primary reference bank that is in turn compared on a monthly basis to the Josephson effect, which provides a realization of the volt. The regular calibration workload is compared with a secondary bank of cells that is compared to the primary bank on a daily basis.

Transfer standards that are assigned values at NBS based on secondary standards are calibrated several times over a long time period in order to reduce the contribution of random error to the uncertainty of the assignment. For example, values for gage blocks that comprise the transfer set from NBS are averages of approximately nine electro-mechanical calibrations completed over a two year period. Furthermore, because s_t can be made small by sufficient repetition and careful excution of the transfer, the total uncertainty in (2.5.3) can be kept close to the uncertainty in (2.5.1) or at least small enough to meet the goals of the measurement assurance program. See figure 2 for a graphic comparison of uncertainties via measurement assurance and conventional calibration routes.

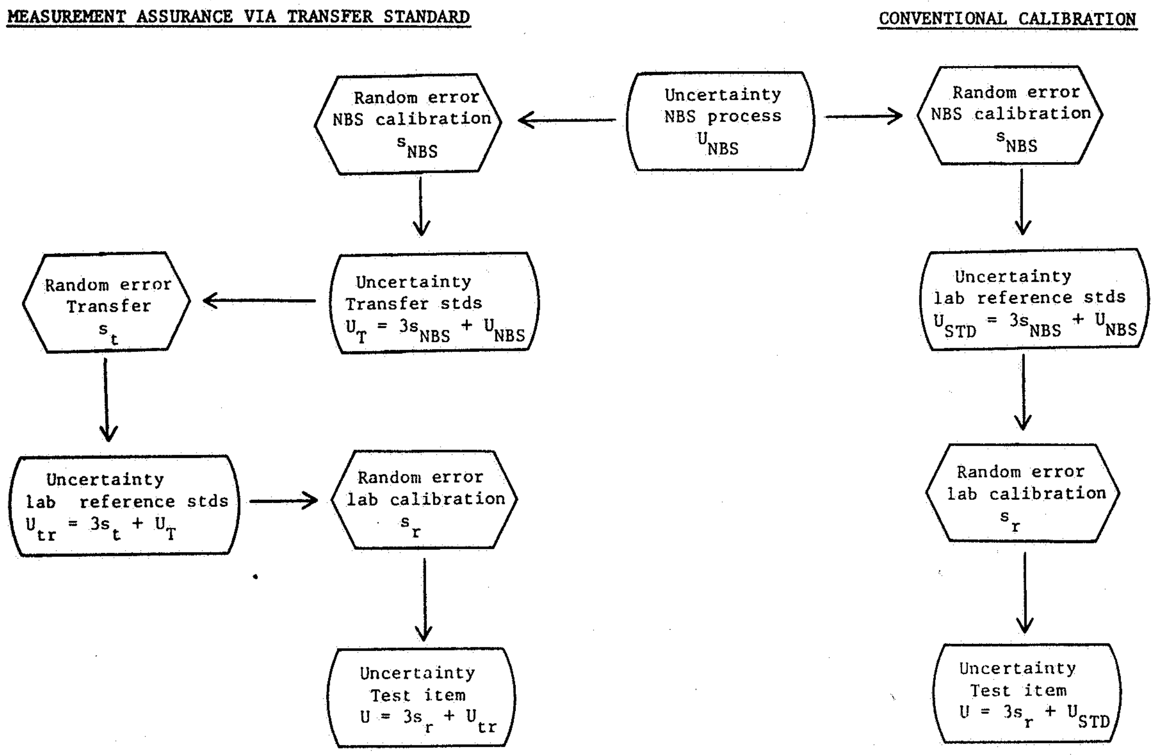

MEASUREMENT ASSURANCE VIA TRANSFER STANDARD
CONVENTIONAL CALIBRATION
Random error NBS calibration s_{NBS}
Uncertainty NBS process U_{NBS}
Random error NBS calibration s_{NBS}
Random error Transfer s_t
Uncertainty Transfer stds $U_T = 3s_{NBS} + U_{NBS}$
Uncertainty lab reference stds $U_{STD} = 3s_{NBS} + U_{NBS}$
Uncertainty lab reference stds $U_{tr} = 3s_t + U_T$
Random error lab calibration s_r
Random error lab calibration s_r
Uncertainty Test item $U = 3s_r + U_{tr}$
Uncertainty Test item $U = 3s_r + U_{STD}$

Figure 2
Diagram showing propagation of uncertainties from NBS process to final uncertainty for test item via measurement assurance route compared to the conventional calibration route

2.5.3 Example of an Uncertainty Statement

The principles of this chapter are illustrated by a preliminary experiment at NBS that eventually led to the development of a linewidth standard. Three sources of systematic error were identified in the NBS photometric process that related linewidth measurement to the fundamental definition of length through line-scale interferometry.

The uncertainty from the interferometric process, resulting from random errors associated with making the interferometric determinations and negligible systematic error, translated into a systematic error in the photometric process of 0.01µm. The maximum differences that were observed between the two operators and two instruments that were employed in the NBS system translated into systematic errors of 0.005µm and 0.020µm respectively.

Values assigned to linewidth artifacts were averaged from four photometric readings, and the standard deviation of each assignment was reported as s_r. The limits to random error were taken to be three times the standard deviation of the assignment. An error budget showing the various components contributing to the total uncertainty is shown below.

Components of Uncertainty

Limit to Random Error = $3s_r$			± 0.040µm
Systematic errors:			
	a. Operator differences	± 0.005µm	
	b. Instrument differences	± 0.020µm	
	c. Uncertainty from interferometry	± 0.010µm	
	Total systematic errors		± 0.035µm
Total Uncertainty[§]			± 0.075µm

Based on this analysis NBS assigned a total uncertainty of ± 0.075µm to artifacts that were calibrated by this system. If such an artifact were to be used by a laboratory for calibrating its optical imaging system, this uncertainty would become a systematic error for that process.

[§] It is suggested that uncertainties be stated to no more than two significant figures and that the last decimal place in the reported value of the measured item correspond in place value to the last decimal place in the uncertainty statement.

3. The Check Standard in a Measurement Assurance Program

3.1 Introduction

A check standard provides a means of characterizing the behavior of a measurement process by way of repeated measurements on the same artifact, combination of artifacts, or instrument over a substantial period of time and over fluctuating environmental conditions. It should be thought of as a data base of such measurements rather than as an artifact per se because it is the measurements, or some function of those measurements, corrected according to the model specifications, that actually describe process performance.

The structure of the check standard measurement depends on whether the calibration procedure is based on a single measurement or a calibration design. In some cases the check standard may be a function of readings on two reference standards, thus eliminating the need for an additional artifact. Check standard measurements of the following types form the basis for the measurement assurance programs in the next chapter.

1) Measurements made on a single artifact as close in time as possible to the measurements on the reference standard and the test item.

2) Differences between the observed values of two reference standards whose assigned values are the basis for assigning a value to a test item.

3) Computed value for single artifact from a statistical design involving k intercomparisons of reference standards, test items and artifact check standard.

4) Computed value of difference between two reference standards from a statistical design involving k intercomparisons of reference standards and test items.

5) Measurements made on a calibrated artifact by a direct reading instrument.

6) Calibrated value of a single artifact from a calibration process that uses a ratio technique.

3.2 Process Parameters Defined by the Check Standard

Measurement processes have two properties that are critical to a measurement assurance program. Measurements of a stable quantity tend to a long-term average which may not be the same average that would be achieved if a different laboratory produced the measurements. As discussed in detail in the last chapter, these measurements while tending to an average, will not be identical because of inability to reproduce conditions of measurement exactly, and this latter property is referred to as process variability or imprecision. Process parameters are quantities that describe the long-term value and the process precision from redundant measurements on a check standard.

The statistic for characterizing the long-term value is simply the arithmetic average of the check standard measurements and is referred to as the "accepted value of the check standard." The check standard measurements supplant the ideal set of measurements that could be made on a test item if it were in the laboratory for a sufficiently long period of time. The average of those hypothetical measurements is, of course, the quantity that is of primary interest, but because such is not at our disposal, we define the process in terms of the accepted value of the check standard. This statistic defines a local base for the measurement process which is intimately related to any discrepancy between the reference base and the average of the measurements that could be made on a test item, and any change in the local base is reason to suspect that this systmatic error has changed.

The statistics for characterizing the process precision are: i) a total standard deviation computed from the same check standard measurements and ii) a within standard deviation computed from each calibration design or group of repetitions for cases where the calibration experiment reports a value based on more than a single measurement on a test item. Within standard deviations are pooled according to (2.2.3) into a single value called the "accepted within standard deviation" which reflects variations that typically take place in the measurement process during the course of a calibration.

If the check standard measurements are properly structured, the accepted total standard deviation reflects the totality of variability in the measurement process. The scatter of check standard measurements will be characteristic of measurements of a test item observed over a period of time in the calibration setting only if both types of measurements are affected by the same sources of error. Then the accepted total standard deviation computed from the check standard measurements can be used to compute the standard deviation for a value reported by the calibration process. Evidently, this computation depends on the type of measurements that are designated as check standard measurements and on the model for the calibration process. Specific examples are discussed in chapter 4.

Before embarking on a full-scale measurement assurance program, the participant conducts a series of experiments to establish a data base of check standard measurements. Accepted values for the process parameters are computed from this data base, and it is emphasized that these experiments should cover several weeks' time and should number at least fifteen to obtain reasonable estimates. The calibration schemes or designs for producing the check standard data must be identical to the procedures for calibrating test items in the workload and measuring transfer standards from NBS.

The importance of the initial check standard measurements dictates that they describe the system in its normal operating mode. Care should be exercised to guarantee that this is indeed the case, so that the standard deviation will be appropriate for an uncertainty statement constructed at any time in the future. This is done by varying the conditions of measurement to cover a representative range of laboratory conditions including operator and environmental variations. These measurements should be scrutinized for outliers because even one significant outlier in a small data set can seriously bias the estimates of the process parameters--perhaps causing an out-of-control condition when the transfer standard is being characterized in the laboratory and invalidating the transfer.

Methods for identifying outliers are highly dependent on underlying distributional assumptions. Several methods for detecting outliers are discussed in ASTM Standard E178[f], but for the foregoing reason, they may not be effective given a limited number of check standard measurements. A plot of the data points is usually satisfactory for detecting outliers. Each check standard measurement should be plotted against a time axis, thus creating a preliminary control chart, and measurements which are obviously aberrant should be deleted from the data set. On the other hand, the data should not be edited in order to achieve seemingly better precision because this will cause failures in the control mechanism at a later time. If a large number of points are suspected as outliers, say more than five percent, the check standard measurements do not constitute a strong data base, and the cause of large variations should be investigated and rectified before proceeding with the measurement assurance program.

3.3 The Check Standard in Process Control

Each check standard measurement is subjected to a statistical test for control, and the outcome of that test is used as a mechanism for accepting or rejecting the results of the measurement process. This presupposes that there is, in fact, a process that is in control, that sufficient data from the process exists to quantify this control, and that the behavior of future measurements is predictable from past behavior of the process. This test is exactly analogous to control chart methodology wherein values that fall inside control limits based on historical data are said to be in control, and values that fall outside the control limits are judged out-of-control.

The technique that is used for control is called a t-test wherein a test statistic is computed from the current check standard measurement, the accepted value of the check standard, and the total standard deviation. This test statistic, when large in absolute value compared to a critical value of Student's t distribution, is indicative of lack of control.

The critical value $t_{\alpha/2}(\nu)$ depends on ν, the number of degrees of freedom in the accepted total standard deviation, and on α, the significance level. The significance level α, the probability of mistakenly flagging a check standard measurement as out-of-control, should be chosen by the participant to be suitably small, say between 0.10 and 0.01, so that the number of remeasurements that must be made because of a chance failure is kept at an acceptable level.

Once the control procedure is installed in the laboratory, the assignments generated by the calibration process are accepted as valid within the stated uncertainty as long as the check standard measurements remain in control. Action is required whenever a check standard measurement is out-of-control. The immediate action is to discard the results of the calibration. Of course, at this point one is faced with a dilemna about what future actions should be taken in regard to the calibration process. Because of the probability of chance failure, exactly α, it is reasonable, while discarding the results of the calibration, to repeat the calibration sequence, hoping that check standard measurements will be in control.

[f] ASTM Standard E178 is available from the American Society for Testing Materials, 1916 Race Street, Philadelphia, Pennsylvania 19103.

In this happy event, one assumes that either something was amiss in the initial calibration, such as insufficient warm-up time for the instrument, or that one was the victim of chance failure. In either case it is permissible to accept the more recent result and proceed as usual. In the event of repeated successive failures or numerous failures over time, one must conclude that a major disruption in the calibration process is affecting the process offset, such as a change in a laboratory reference standard, and the calibration process should be shut down until the problem can be rectified and control reestablished. Each calibrration experiment is intended to reveal the offset of a test item or the client's process relative to NBS, and this offset will not be correctly estimated by the calibrating laboratory if the long-term average for its measurements is not constant relative to the reference base. Therefore, a failure of the check standard test implies that offset has not been eliminated or accounted for by the calibration experiment.

A consideration in choosing α is that the significance level for process control should be the same as the significance level for determining the limits of error in section 2.3. Smaller values of α, the probability of having to remeasure unnecessaarily, that is because of chance failure, correspond to larger associated limits of error. Thus the cost of remeasurement must be weighed against the impact of a larger uncertainty. Values of $\alpha = 0.05$ or $\alpha = 0.01$ are recommended.

An alternative to a critical value based on the t-distribution, as explained in section 2.3, is a factor such as three or two which can be used for computing limits to random error and testing for control. The factor three corresponds approximately to $\alpha = 0.003$ for the normal distribution and is well established in quality control applications. There are no hard and fast rules for picking either a significance level α or a factor such as three for the control procedure, but once it is chosen, it plays a large part in determining the frequency of remeasurement and the magnitude of the uncertainty for the process.

The measurement assurance procedures that are outlined in the next chapter are based upon a critical value of three in almost all cases. Those wishing a more stringent control procedure can substitute the appropriate value of $t_{\alpha/2}$ in the appropriate equations. In calibration work, the purpose of the control procedure is to flag those measurements which are clearly out-of-control, and a critical value of three is suitable for many situations. This approach is the current practice of many calibration services at NBS. Moreover, limits based on the factor three work well, covering a large percentage of the distribution of possible values of the test statistic, even where the test statistic is not strictly distributed as Student's t which is the case for some of the more complicated constructions in the next chapter.

If the measurement sequence allows for a within standard deviation, the ratio of this within standard deviation to the accepted within standard deviation is compared to a critical value based on Snedecor's F distribution (see Ku [44] for a discussion of the F test). A ratio that is large compared to the critical value is indicative of lack of control during the course of the measurement sequence.

The critical value $F_{\alpha}(\nu_1,\nu_2)$ depends on; ν_1, the number of degrees of freedom in the current within standard deviation; ν_2 the number of degrees of freedom in the accepted within standard deviation; and α, the significance level discussed in preceding paragraphs. Critical values of $F_{\alpha}(\nu_1,\nu_2)$ are tabulated in Table II for α=0.01, ν_1=1(1)10(2)30(10)120 and ν_2=10(1)20(2)30(5)120.†

The t-test and F test are invoked simultaneously--the failure of either test constituting grounds for discarding the measurement on the test item or transfer standard. The combination of these two tests is a powerful means of detecting shifts in the long-term average of the process as it defines systematic error.

The efficacy of the check standard as a device for guaranteeing that the process is functioning properly and that, therefore, the test items are assigned values with negligible offset relative to NBS, depends on the relationship among the measurements made on the test items, the measurements made on the reference standards and the measurements made on the check standards. The strongest case for measurement assurance exists when all assignments are statistically interrelated as in a statistical design. When the assignments are by nature statistically independent, it is essential that the measurements be temporally related by completing the measurement sequence in as short a time as possible.

There is really no guarantee that a predictable response on the part of the check standard assures a good measurement on the test item if it is possible for the process to change appreciably during the intervening time between the check standard measurement and the other measurements. However, a strong case for confidence in a measurement process exists for a process that is continuously in control. Furthermore, out-of-control findings for the check standard are almost unfailingly indicators of measurement problems because the control limits are specified so that the probability of a single value being out-of-control is extremely small.

The question of how often the process should be checked for control can only be answered in terms of the goals of the measurement program. A criterion based on economic considerations must balance the tradeoff between the cost of making additional measurements to ensure accuracy and the costs incurred when inaccurate measurements are allowed to occur. In order to achieve the highest level of measurement assurance, check standard measurements should be incorporated in every calibration sequence. When this is not possible or not necessary, a check for control should be incorporated in start-up procedures and repeated at intervals thereafter that depend on the level of system control that is desired and on past experiences with the control procedure.

A system that is always in control when checked can be presumed to remain in control between checks, and the time between check standard measurements can be lengthened. Conversely, the same presumption cannot be made for a system that is occasionally out-of-control, and the time between check standard measurements should be shortened if one is to determine how long the system can operate in-control.

†The notation 10(2)30, for example, indicates that the values go in steps of two from ten to thirty.

3.4 The Transfer with NBS

During the transfer between the participating laboratory and NBS, current check standard measurements that result from the transfer experiments are compared with the accepted value of the check standard by a t-test in order to ascertain whether or not there has been a significant change in the long-term average of the process. If the check standard measurements are continually out-of-control while the transfer standard is in the laboratory, the transfer measurements are invalid, and the transfer experiment should be discontinued until the initial check standard measurements are repeated and new accepted values are established. Isolated failures can be treated as they are treated in the calibration workload, and offending measurements that cannot be repeated are deleted from the transfer data.

Similarly, the within standard deviation computed from the transfer measurements is compared with the accepted within standard deviation by an F-test. If possible, sufficient repetitions spaced over a period of time are also included in the procedures for measuring the transfer standards so that the standard deviation for the transfer can be compared to the accepted total standard deviation.

After the completion of the transfer with NBS, the tests for control are continued for the calibration process. When an out-of-control condition is encountered in this mode, the measurement process is discontinued until control is restored which may amount to simply repeating the measurement sequence on the test item and check standard. When it is obvious that the process mean has shifted because of repeated out-of-control findings for the check standard, signifying that the offset from NBS has changed, it is time for another intercomparison with NBS. Theoretically one may be able to analyze the amount of change in the offset, but it seems judicious at this point to reestablish the values of the laboratory's reference standards.

3.5 Updating Process Parameters

After the control procedure has been in place for a year or more, sufficient data should be available so that the process parameters can be updated. The mechanics for doing this depend on the degree of automation that exists in the laboratory and on the computing capability at its disposal. In a sophisticated program one compares the accepted value for the check standard and the accepted total standard deviation with values computed from the check standard data that has been accumulated since the last update. If the two sets of data are essentially in agreement, updated process parameters are computed based on all check standard measurements. In cases where the process has changed significantly in regard to these parameters, the past historical data are discarded, and new process parameters are computed from the most recent data. For computer systems such as micro-computers with limited storage capacity, it may be feasible to retain only a fixed number of check standard measurements. Obviously the number should be sufficient for obtaining reliable estimates. The data file is continually updated by deleting the oldest measurement and adding the newest-thereby always keeping a fixed number of check standard measurements in the data file with which to compute the process parameters.

4. Implementation of Measurement Assurance for Specific Cases

This chapter contains the basic outlines for implementing measurement assurance programs for eight specific measurement situations where the sequence of measurements that constitute an intercomparison depends upon the number of reference standards, the number of test items and the number of redundant measurements to be employed in each intercomparison.

The essential elements that specify the measurement situation for each plan are as follows:

4.1 A comparator process in which one reference standard is compared to a test item and a check standard.

4.2 A comparator process in which a test item is compared to each of two reference standards, and control is maintained on the difference between readings on the two reference standards.

4.3 A comparator process in which three test items are compared to two reference standards in a statistical design, and control is maintained on the difference between the two standards.

4.4 A comparator process for mass calibrations illustrating the use of a 1, 1, 1 design and a 5, 3, 2, 1, 1, 1 design with provision for a check standard for each series.

4.5 A comparator process in which four test items are compared to four reference standards, without direct intercomparison between the test items or reference standards. Control is maintained on the difference between two reference standards.

4.6 Direct reading of the test item with the instrument as the standard. Control is maintained by repetitions on a calibrated artifact.

4.7 Simultaneous measurement of a group of test items relative to a bank of reference standards where a check standard is always included among the test items.

4.8 A ratio technique for one or more test items and one or two reference standards. Control is maintained on calibrated values of an artifact check standard.

Calibration as a process of intercomparing a test item with a reference standard and assigning a value to the test item based on the accepted value of the standard is frequently carried out by a comparator process. For high precision work, the comparator process makes use of an instrument or device which is capable of handling only very small differences between properties of similar objects such as a mechanical comparator for comparing gage blocks of the same nominal length or an electrical bridge for detecting very small differences between resistances. Where individual readings, in scale units, are taken on the unknown and the reference standards and converted to the appropriate units, a value can be assigned to the test item only through the difference between the reading on the test item and the reading on the reference standard (See section 1.4.2). The calculated difference between the two readings is the "measurement of interest" and the number of such differences determines the redundancy in a measurement scheme.

Where the calibration experiment produces only a difference measurement, such as the difference in emf between two saturated cells as measured by a potentiometer, the term "reading on an unknown" or "reading on a standard" does not have a literal interpretation but refers to the logical intercomparison of the items. In either case, a value is assigned to an unknown relative to the known value of one or more reference standards. This known value is referred to as the restraint.

Where there are a small number of unknowns and reference standards, the calibration experiment may consist of all possible intercomparisons that can be made on the collection of items; this would amount to k(k-1)/2 difference measurements for k items being intercompared two at a time. A calibration design consists of a subset of all possible intercomparisons such that, given a restraint or assigned value for the reference standards, the series of intercomparisons can be solved for the unknowns. The method for finding a solution is least-squares, and the resulting values for the unknown items are least-square estimates.

Several factors dictate the choice of intercomparisons that constitute the design. Obviously, it is desirable to keep the number of intercomparisons small. Designs are usually structured so that precision in the assignments to the test items is the same for all items of the same nominal size and so that precision in this sense is optimized for a given number of intercomparisons. Other optimality criteria that are discussed in the statistical literature in references [45] and [46] may be of interest to the reader.

Calibration can also be carried out using a direct reading device or instrument in which case the device is regarded as the standard, and values, already in the appropriate units, are assigned directly to the test items. Such a device, for example an interferometer, can also be used in a comparator mode in which case the difference between a reading on the test item and a reading on the standard is regarded as the measurement of interest.

The eight measurement plans that are discussed in this section have been adapted to both mechanical and electrical measurements. Plan 4.1 is the simplest scheme for a comparator process and may be appropriate when accuracy requirements are moderate. It does not afford a high degree of protection because the linkage between the measurement on the test item and the measurement on the check standard is not as strong as it is for the other comparator schemes. Plan 4.2 affords a higher degree of protection against incorrect measurements by requiring redundant measurements on each test item. This plan is well suited to mechanical measurements and is currently utilized in the Gage Block Measurement Assurance Program. The program is illustrated with data from one participant in section 4.2.7.

Plans 4.3 and 4.5 involve calibration designs that are particularly appropriate for voltage and resistance measurements. The designs have a provision for estimating a so-called left-right effect which is an important circuit parameter for voltage measurements. The discussion of plan 4.5, which is illustrated with data from the NBS Volt Transfer Program, explains the steps to be followed in process control using a check standard that is either stable or is drifting linearly with time.

Plan 4.4 describes a measurement assurance program for guaranteeing the accuracy of very precise weighings by means of two designs which are routinely used in the NBS mass calibration program. Weighing designs for different combinations of weights along with designs for mechanical and electrical measurements involving more standards and test items are described by Cameron et al [47]. Designs for eliminating temporal effects are described by Cameron and Hailes [48].

Surveillance testing as a means of ensuring the self-consistency of a weight set is described in detail in a recent publication by Jaeger and Davis [49]. The basic idea is to compare a given weight against a collection of other weights in the set whose nominal sum equals the first weight. The authors develop measurement assurance methods for monitoring the difference calculated from the comparison and resolving it with values assigned to the individual weights.

Plan 4.6 is probably the simplest and involves only direct readings on the test items. It is appropriate for large volume workloads that utilize an instrument standard such as interferometer, digital voltmeter, or electronic balance where there is a need to monitor or guarantee the accuracy of the instrument as a matter of course.

Plan 4.7 is appropriate for assigning values to test items or instruments relative to a bank of standards where the calibration consists of subjecting all items including the reference standards to the same stimuli, usually simultaneously. Control is maintained by a check standard which is included as a test item in each measurement sequence. Applications include watthour meter calibration where test meters and reference meters are connected to the same power source and very low pressure calibration where several pressure gages are confined in a vacuum chamber with a reference pressure gage.

By necessity, the analyses are outlined in a straightforward manner, and problems involving drifting reference standards or check standards must be considered separately. It is obviously impossible to anticipate the spectrum of complications that may arise in a given measurement area, and these analyses, offered as a simplistic approach to sometimes difficult problems, are intended to provide a starting point for measurement assurance.

Each measurement assurance program that is presented in this chapter relies upon a check standard concept as discussed at length in the last chapter, and the check standard measurements are crucial to the steps that constitute such a program; namely i) establishment of process parameters; ii) routine process control; iii) evaluation of systematic error by transfer with NBS; iv) determination of uncertainty for test items; v) update of process parameters.

The first four steps are outlined in detail for each program, and the fifth step relating to updating and maintaining the data base was discussed in generality in section 3.5.

4.1 Comparator Process for One Test Item, One Reference Standard, and One Check Standard

4.1.1 Measurement Sequence

This scheme is appropriate for a comparator process where the intercomparison of the test item X with the reference standard R is immediately followed by the intercomparison of an artifact check standard Y with the reference standard R in the sequence X, R, Y, R. The readings are denoted by x, r_1, y, r_2 respectively. This measurement sequence should be followed for all calibrations for which statistical control is to be achieved. The value of the check standard for one such sequence is defined from the reading on the artifact check standard and the duplicate readings on the reference standard as

$$c = y - \frac{1}{2}(r_1 + r_2) \ . \qquad (4.1.1)$$

All aspects of a measurement assurance program involving this design are explained and illustrated for gage blocks in reference [50].

4.1.2 Process Parameters

Initial values of the process parameters are obtained from n such measurement sequences, where $c_1, \cdots, c_n$ are the observed values of the check standard. The accepted value of the check standard is the mean of the check standard measurements; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i \ . \qquad (4.1.2)$$

The accepted total standard deviation for the check standard is

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n}(c_i - A_c)^2\right)^{1/2} \qquad (4.1.3)$$

with ν = n-1 degrees of freedom.

The model assumed for the calibration process is the additive model (1.4.2). Under this model the error structure for the value of the test item and the error structure for the check standard measurement are identical. Thus s_c also estimates the standard deviation of the reported value of the test item which is shown in (4.1.6).

The control limits[h] that are appropriate for future check standard observations are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \; .$$

4.1.3 Control Procedure

The control procedure applied to each calibration depends on a test statistic t_c that is computed from the observed value of the check standard c for that measurement sequence by

$$t_c = \frac{|c - A_c|}{s_c} \; . \qquad (4.1.4)$$

If

$$t_c < 3 \qquad (4.1.5)$$

the process is in control, and the value of the test item is reported as

$$X^* = x - \frac{1}{2}(r_1 + r_2) + R^* \qquad (4.1.6)$$

where R^* is the value assigned to the reference standard.

If

$$t_c > 3,$$

the calibration of the test item is invalid and must be repeated.

[h] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; $t_{\alpha/2}(\nu)$.

4.1.4 Transfer with NBS

The transfer with NBS is accomplished by p repetitions of the measurement sequence in which a transfer standard takes the place of the test item in each repetition. Process control as defined by (4.1.5) should be confirmed for each repetition. Any sequence that is out-of-control should be repeated until control is reestablished or else that repetition is deleted from the transfer. If the value assigned to the transfer standard by NBS is T^* with uncertainity U_T, the uncertainty of the transfer is

$$U_{tr} = \frac{3s_c}{\sqrt{p}} + U_T \ . \qquad (4.1.7)$$

The offset Δ of the laboratory process from NBS is

$$\Delta = \frac{1}{p} \sum_{j=1}^{p} X_j{}^* - T^* \qquad (4.1.8)$$

where $X_1{}^*,\cdots,X_p{}^*$ are values calculated according to (4.1.6) for the transfer standard for each of the p repetitions.

This offset is judged significant if

$$\frac{\sqrt{p}\,|\Delta|}{s_c} > 3 \ , \qquad (4.1.9)$$

and in such case the assigned value of the reference standard becomes $R^* - \Delta$.

The assigned value of the reference standard is unchanged if

$$\frac{\sqrt{p}\,|\Delta|}{s_c} < 3 \ .$$

4.1.5 Total Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by <u>one</u> calibration sequence is

$$U = U_{tr} + 3s_c. \qquad (4.1.10)$$

4.2 Comparator Process for One Test Item and Two Reference Standards

4.2.1 Measurement Sequence

This scheme involving duplicate measurements on the test item is appropriate for a comparator process where the assignment for the test item is made relative to the average of the values assigned to the two reference standards, called the restraint R^*. The intercomparison of the test item X with each of two reference standards, R_1 and R_2, in a trend eliminating design (Croarkin et al [51]) is accomplished by the sequence X, R_1, R_2, X, and the readings are denoted by x_1, r_1 r_2, x_2 respectively. The difference measurements are:

$$d_1 = x_1 - r_1$$

$$d_2 = x_2 - r_2$$

There is no artifact check standard for this design, and a check standard value is defined for each sequence as the calculated difference between the readings on the two reference standards as

$$c = d_2 - d_1 \qquad (4.2.1)$$

The value c is structured so as to reflect the maximum variation that occurs in the measurement sequence between the first and the last readings on the test item and not just the variation that occurs between the readings on the two reference standards.

4.2.2 Process Parameters

Initial values of the process parameters are obtained from n such measurement sequences yielding check standard values $c_1, \cdots, c_n$. The accepted value of the check standard is given by the mean of the check standard values; namely,

$$A_c = \frac{1}{n} \sum_{i=1}^{n} c_i \; . \qquad (4.2.2)$$

The total standard deviation of the check standard is defined by

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \qquad (4.2.3)$$

with $\nu = n-1$ degrees of freedom.

The control limits[i] that are appropriate for future observations on the check standard are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \; .$$

[i] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

The model assumed for the process is the additive model (1.4.2). The error structures for the check standard measurement and the reported value of the test item are worked out in detail in section 1.5 where it is shown that the standard deviation for the reported value of the test item is $s_c/2$.

4.2.3 Control Procedure

The control procedure applied to each calibration depends on a statistic t_c that is computed from the observed value of the check standard c for that measurement sequence where

$$t_c = \frac{|c - A_c|}{s_c} . \qquad (4.2.4)$$

If

$$t_c < 3 \qquad (4.2.5)$$

the process is in control, and the value of the test item is reported as

$$X^* = \frac{1}{2}(d_1 + d_2) + R^* \qquad (4.2.6)$$

where the restraint $R^* = \frac{1}{2}(R_1^* + R_2^*)$, and R_1^* and R_2^* are the assigned values of the reference standards.

If

$$t_c > 3,$$

the calibration of the test item is invalid and must be repeated.

4.2.4 Transfer with NBS

The transfer with NBS can be accomplished with two tranfer standards T_1 and T_2. In this mode p_1 repetitions of the measurement sequence are made with T_1 taking the place of the test item and p_2 repetitions of the measurement sequence are made with T_2 taking the place of the test item. This produces a total of $p_1 + p_2$ repetitions for the transfer. Process control as defined by (4.2.5) should be confirmed for each repetition. Any sequence that is out-of-control should be repeated until control is reestablished or else that repetition is deleted from the transfer. If the values assigned to the transfer standards by NBS are T_1^* and T_2^* with uncertainties U_{T1} and U_{T2} respectively, the uncertainty of the transfer is

$$U_{tr} = \frac{3}{4}\left(\frac{p_1+p_2}{p_1 \cdot p_2}\right)^{1/2} s_c + U_T \qquad (4.2.7)$$

where

$$U_T = \frac{1}{2}\left(U_{T1}^{\ 2} + U_{T2}^{\ 2}\right)^{1/2} .$$

The offset Δ of the laboratory process from NBS is defined only in terms of the restraint; i.e., the average of the two reference standards. It is computed from the p_1 values assigned to the first transfer standard according to (4.2.6); namely, $X_1^*,\cdots,X_{p_1}^*$ and the p_2 values assigned to the second transfer standard according to (4.2.6); namely, $X_1^{**},\cdots,X_{p_2}^{**}$.

$$\Delta = \frac{1}{2p_1}\sum_{i=1}^{p_1} X_i^* + \frac{1}{2p_2}\sum_{i=1}^{p_2} X_i^{**} - \frac{1}{2}(T_1^* + T_2^*) \qquad (4.2.8)$$

The offset is judged significant if

$$\tilde{t} > 3, \qquad (4.2.9)$$

where

$$\tilde{t} = \frac{4\sqrt{p_1 \cdot p_2}\ |\Delta|}{\sqrt{p_1+p_2}\ s_c} \qquad (4.2.10)$$

and in such case the assigned value of the restraint is changed to $R^* - \Delta$. The restraint is unchanged if $\tilde{t} < 3$.

4.2.5 Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by (4.2.6) from one calibration sequence is

$$U = U_{tr} + \frac{3s_c}{2} . \qquad (4.2.11)$$

4.2.6 Example from the Gage Block Measurement Assurance Program

Two sets of eighty-one gage blocks from NBS were sent to industrial participants for the purpose of assigning values to their laboratory reference standards. Before the transfer blocks left NBS, each participant conducted a minimum of six experiments in which his two sets of reference standards were compared to a set of test blocks according to the measurement scheme in section 4.2.1. Because six measurements are not sufficient for estimating a standard deviation, the data were analyzed by groups, with about twenty blocks constituting a group.

In order to check a large data set for outliers, such as the data accumulated on the gage block check standards, it is sometimes possible to make use of the information in the individual standard deviations. Because the measurements are assumed to all come from the same process, a standard deviation that is large compared to the other standard deviations in the group suggests an outlier in the check standard measurements for that nominal size.

If there are k block sizes in a group, the test statistic is the ratio of a single standard deviation s_i to a quantity that has been pooled from the remaining standard deviations in that group; namely, s_j $(j=1,\cdots,k;\ j\neq i)$. The test statistic is

$$F = \left(s_i/s_{p_i}\right)^2$$

where

$$s_{p_i} = \left(\frac{1}{k-1}\sum_{j\neq i} s_j^{\,2}\right)^{1/2}$$

and s_i has ν_1 degrees of freedom and each pooled standard deviation has ν_2 degrees of freedom. If all s_i have the same number of degrees of freedom ν, then $\nu_1 = \nu$ and $\nu_2 = (k-1)\cdot\nu$. If for α chosen suitable small,

$$F \;>\; F_\alpha(\nu_1,\nu_2)$$

where $F_\alpha(\nu_1,\nu_2)$ is the upper α percent point of the F distribution with ν_1 and ν_2 degrees of freedom, the standard deviation in question is considered significant, and the individual measurements for that check standard are inspected for an outlier--the outlier being either the largest or the smallest measurement.

Consider the standard deviations in exhibit 4.2.1 which were computed from check standard measurements for nine nominal sizes. The individual measurements are plotted in figure 3 as deviations from the mean for each nominal size as a function of nominal size. Test statistics computed for each nominal size show that the standard deviation for the 0.122000 inch check standard is significantly larger than the others, and figure 3 verifies that the smallest observation is not consistent with the other data for that size and is thus labeled an "outlier."

Exhibit 4.2.1 - Standard deviations from check standard measurements
Values in microinches

Nominal Length (Inches)	Std Devs	Degrees of Freedom	Pooled Std Devs	Degrees of Freedom	Test Statistic
	s_i	ν_1	s_{p_i}	ν_2	F
0.117000	0.445	5	0.723	40	0.38
0.118000	0.288	5	0.733	40	0.15
0.119000	0.952	5	0.659	40	2.09
0.120000	0.382	5	0.727	40	0.28
0.121000	0.616	5	0.707	40	0.76
0.122000	1.303	5	0.579	40	5.06†
0.123000	0.539	5	0.715	40	0.57
0.124000	0.674	5	0.700	40	0.93
0.125000	0.472	5	0.721	40	0.43

† $\left(s_i/s_{p_i}\right)^2 > F_{.01}(5,40)$ where $F_{.01}(5,40) = 3.51$ from Table II.

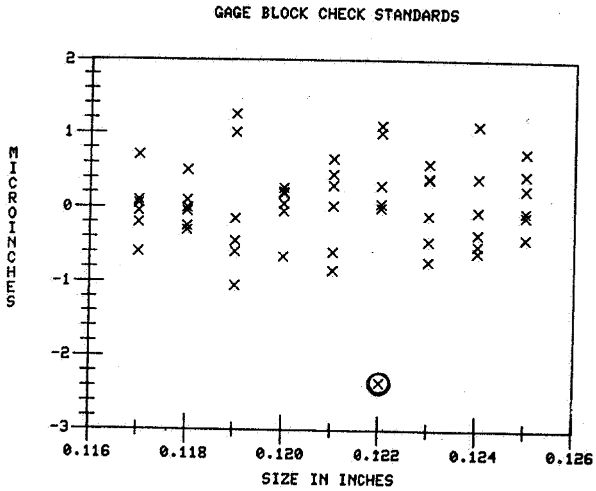


Figure 3
Deviations (microinches) from the mean versus nominal length (inches) for groups of six check standard measurements showing a single outlier

The initial data taken by the participants in the measurement assurance program were inspected for outliers by this method. All outliers were deleted from the data base before calculating the accepted values and standard deviations of the check standard measurements. A subset of the data for one participant is featured in exhibit 4.2.3 with the number of blocks being restricted to five for the purpose of the illustration. The exhibit shows the data from the initial experiments, with a check standard for each repetition computed according to (4.2.1) and initial values for the process parameters A_c and s_c computed using (4.2.2) and (4.2.3) respectively. After the initial data set was edited for outliers, the transfer blocks were sent to the participant. The values assigned to the transfer standards by NBS and the value for the participant's restraint are listed in exhibit 4.2.2.

Exhibit 4.2.2 - Participant's restraint and NBS values for transfer standards
Values in microinches

Nominal	Restraint	Transfer Stds		Uncertainties		Total[§]
	R^*	T_1^*	T_2^*	U_{T1}	U_{T2}	U_T
0.1006	1.30	-0.63	-0.56	2.17	2.06	2.12
0.1008	0.80	3.21	3.14	2.17	2.06	2.12
0.1010	2.65	2.33	2.52	2.17	2.06	2.12
0.1020	0.45	0.35	0.19	2.17	2.06	2.12
0.1030	-0.05	-2.09	-2.32	2.17	2.06	2.12

[§] The systematic errors associated with the transfer standards are added linearly instead of in quadrature because the assignments T_1^* and T_2^* are not independent. Thus $U_T = (U_{T1} + U_{T2})/2$.

Exhibit 4.2.3 - Readings on unknown X and reference standards R_1 and R_2
Corrections to nominal size in microinches

Nominal (inches)	Reps	Readings				Check Standard	Mean	Total S.D.	D.F.
		x_1	r_1	r_2	x_2	c	A_c	s_c	ν
0.1006	1	53.9	52.7	46.8	53.9	5.9	5.80	0.616	5
	2	54.8	49.9	45.0	54.5	4.6			
	3	56.0	50.0	44.1	56.1	6.0			
	4	56.3	50.0	44.1	56.3	5.9			
	5	55.1	49.7	43.7	55.1	6.0			
	6	55.0	50.0	43.9	55.3	6.4			
0.1008	1	51.1	51.1	49.2	50.5	1.3	2.33	0.554	5
	2	53.0	49.9	47.9	53.1	2.1			
	3	54.2	50.1	47.5	54.3	2.7			
	4	54.4	50.2	47.5	54.3	2.6			
	5	53.2	49.9	47.2	53.2	2.7			
	6	53.3	50.0	47.3	53.2	2.6			
0.1010	1	52.0	50.1	49.1	52.5	1.5	1.70	0.593	5
	2	54.8	48.8	47.7	54.7	1.0			
	3	55.5	50.0	48.4	55.5	1.6			
	4	55.4	50.0	48.3	55.4	1.7			
	5	55.5	51.0	48.2	55.5	2.8			
	6	55.8	50.0	48.2	55.6	1.6			
0.1020	1	52.1	50.1	48.1	52.2	2.1	2.07	0.339	5
	2	57.3	51.1	49.0	57.2	2.3			
	3	57.0	50.0	48.3	57.0	1.7			
	4	57.2	50.1	48.4	57.1	1.6			
	5	55.3	50.0	47.6	55.3	2.4			
	6	55.1	49.9	47.6	55.1	2.3			
0.1030	1	53.9	49.2	48.5	54.1	0.9	0.73	0.361	5
	2	58.8	50.0	49.0	58.8	1.0			
	3	59.4	50.0	49.1	59.5	1.0			
	4	59.4	50.0	49.1	59.4	0.9			
	5	59.3	50.0	49.5	58.9	0.1			
	6	59.7	50.2	49.6	59.6	0.5			
							Pooled	0.507	25

Each transfer block was intercompared twice with the participants reference standards by the same scheme used to obtain the initial data, resulting in a total of $p_1 + p_2 = 4$ repetitions. The data for each repetition are shown in exhibit 4.2.4. The readings on the reference standards are designated by r_1 and r_2, and the duplicate readings on a transfer standard are designated by x_1 and x_2. The exhibit also lists the check standard that was computed for each repetition, the test statistic t_c, and the value reported for the NBS transfer standard according to (4.2.6).

Notice that on three occasions the check standard measurement failed the test for control defined by (4.2.5). Because the data were analyzed at NBS after the transfer standards left the participant's laboratory, it was not possible to repeat those sequences, and they were deleted from the transfer data thereby reducing the number of valid repetitions for those block sizes.

Exhibit 4.2.4 - Readings on transfer standards T_1 and T_2
Corrections to nominal size in microinches

Nominal (inches)	Stds	Reps	Readings				Check Std	Test Statistic	Transfer Std
			x_1	r_1	r_2	x_2	c	t_c	X*
0.1006	T_1	1	51.2	55.2	48.0	50.8	6.8	2.0	0.70
	T_1	2	51.3	55.2	48.9	51.2	6.2	0.8	0.50
	T_2	1	50.8	54.9	48.1	51.3	7.3	3.0§	----
	T_2	2	51.2	55.2	48.9	51.3	6.4	1.2	0.50
0.1008	T_1	1	56.5	55.3	52.4	56.3	2.7	0.7	3.35
	T_1	2	56.2	55.1	52.5	56.2	2.6	0.5	3.20
	T_2	1	56.1	55.1	52.3	56.4	3.1	1.5	3.35
	T_2	2	55.7	55.0	52.5	55.8	2.6	0.5	2.80
0.1010	T_1	1	54.0	54.9	53.2	54.0	1.7	0	2.60
	T_1	2	53.5	55.0	52.8	53.5	2.2	1.0	2.25
	T_2	1	53.9	54.9	53.3	53.9	1.6	0.2	2.45
	T_2	2	53.8	55.0	52.8	53.9	2.3	1.2	2.60
0,.1020	T_1	1	54.9	54.3	52.1	54.7	2.0	0.1	2.05
	T_1	2	55.0	55.1	52.5	55.0	2.6	1.0	1.65
	T_2	1	54.6	54.3	52.1	54.6	2.2	0.3	1.85
	T_2	2	55.2	55.1	52.5	55.2	2.6	1.0	1.85
0.1030	T_1	1	52.9	53.9	52.9	52.8	0.9	0.3	-0.60
	T_1	2	53.9	54.9	52.4	53.9	2.5	3.5§	----
	T_2	1	52.4	53.9	52.9	52.5	1.1	0.7	-1.00
	T_2	2	53.4	54.9	52.4	53.4	2.5	3.5§	----

§Failed the test for control. The Gage Block Measurement Assuance Program uses a critical value of 3.

Offsets from NBS were computed for each block size by (4.2.8) and were tested for significance by (4.2.9). The participant was advised to change the value of the restraint for those block sizes which showed a significant offset from NBS. The uncertainty of the current transfer with NBS was computed. Results are reported in exhibit 4.2.5. The participant was further advised that the uncertainty appropriate for his process was $U = 4.73$ microinches as calculated by (4.2.11).

This uncertainty is valid for calibrations conducted according to the measurement scheme in Section 4.1.1 with the value of the restraint as stipulated as long as the process remains in control. Another transfer with NBS will be scheduled in two years to check on the state of the measurement assurance program, and it is anticipated that thereafter transfers with NBS will become increasingly rare. Specific blocks that shows signs of change can be recalibrated or replaced in the interim.

Exhibit 4.2.5 - Offsets from NBS and corrected restraints
Values in microinches

Nominal (inches)	Number of Repetitions	Offset	Test Statistic	Corrected Restraint	Uncertainty of Transfer
	$p_1 + p_2$	Δ	$\tilde{t}$	$R^* - \Delta$	U_{tr}
0.1006	3	1.14	7.3†	0.16	2.59
0.1008	4	0.00	0.0	0.80§	2.50
0.1010	4	0.05	0.4	2.65§	2.50
0.1020	4	1.58	12.5†	-1.13	2.50
0.1030	2	1.40	7.8†	-1.45	2.66

†The test statistic $\tilde{t} > 3$ indicating that the offset from NBS is significant and that the laboratory restraint should be decreased by the amount Δ.

§The restraint is unchanged because the offset is not significant.

4.3 Comparator Process for Three Test Items and Two Reference Standards

4.3.1 Measurement Sequence

In this scheme, which is particularly suitable for electrical measurements, the small difference between two items, such as the difference between the electromotive forces for two saturated cells, constitutes a measurement. The assignments of values to test items are done relative to two reference standards. The statistical design leads not only to equal precision in the assigned value for each test item, but it is also structured so that any position effect in the electrical connection, called left-right effect, is cancelled (Cameron & Eicke [52]). The theory of least-squares estimation which governs the solution of this type of design is explained by Cameron in reference [53].

The design is composed of a subset of all possible difference measurements that could be made on the two standards and three test items. The total number of measurements that could be made in order to achieve left-right balance on such a complement of standards and test items is twenty, and the design is parsimonious in that it requires a subset of ten of the possible measurements while still achieving equal precision for each assignment.

The reference standards are designated by R_1 and R_2, the test items by X, Y, and Z and the corresponding intercomparisons on each by r_1, r_2, x, y, z respectively. The order of measurements is given below:

$$\begin{aligned} d_1 &= r_1 - r_2 \\ d_2 &= r_2 - x \\ d_3 &= x - y \\ d_4 &= y - z \\ d_5 &= z - r_1 \\ d_6 &= y - r_1 \\ d_7 &= r_2 - y \\ d_8 &= z - r_2 \\ d_9 &= x - z \\ d_{10} &= r_1 - x \end{aligned} \tag{4.3.1}$$

The left-right effect is estimated by

$$\hat{\zeta} = \frac{1}{10} \sum_{i=1}^{10} d_i \ . \tag{4.3.2}$$

The differences of the reference standards from their average as estimated by least-squares are:

$$\hat{r}_1 = \frac{1}{10}(2d_1 - d_2 - d_5 - d_6 - d_7 + d_8 + d_{10})$$

$$\hat{r}_2 = \frac{1}{10}(-2d_1 + d_2 + d_5 + d_6 + d_7 - d_8 - d_{10})$$

and the corresponding differences for the test items are:

$$\begin{aligned} \hat{x} &= \frac{1}{10}(-3d_2 + 2d_3 + d_5 + d_6 - d_7 + d_8 + 2d_9 - 3d_{10}) \\ \hat{y} &= \frac{1}{10}(-d_2 - 2d_3 + 2d_4 + d_5 + 3d_6 - 3d_7 + d_8 - d_{10}) \\ \hat{z} &= \frac{1}{10}(-d_2 - 2d_4 + 3d_5 + d_6 - d_7 + 3d_8 - 2d_9 - d_{10}). \end{aligned} \tag{4.3.3}$$

The within standard deviation for each design is

$$s_w = \left(\frac{1}{\nu}\sum_{i=1}^{10} \xi_i^{\,2}\right)^{1/2} \tag{4.3.4}$$

with degrees of freedom $\nu = 5$.

The individual deviations ξ_i from the least-squares fit are defined by:

$$\begin{aligned} \xi_1 &= d_1 - \hat{r}_1 + \hat{r}_2 - \hat{\zeta} \\ \xi_2 &= d_2 - \hat{r}_2 + \hat{x} - \hat{\zeta} \\ \xi_3 &= d_3 - \hat{x} + \hat{y} - \hat{\zeta} \\ \xi_4 &= d_4 - \hat{y} + \hat{z} - \hat{\zeta} \\ \xi_5 &= d_5 - \hat{z} + \hat{r}_1 - \hat{\zeta} \\ \xi_6 &= d_6 - \hat{y} + \hat{r}_1 - \hat{\zeta} \\ \xi_7 &= d_7 - \hat{r}_2 + \hat{y} - \hat{\zeta} \\ \xi_8 &= d_8 - \hat{z} + \hat{r}_2 - \hat{\zeta} \\ \xi_9 &= d_9 - \hat{x} + \hat{z} - \hat{\zeta} \\ \xi_{10} &= d_{10} - \hat{r}_1 + \hat{x} - \hat{\zeta}. \end{aligned} \tag{4.3.5}$$

This design can be used for measurement situations where there is no left-right effect to be estimated. In this case, the equations for the deviations ξ_i do not have the term ζ, and the degrees of freedom associated with s_w is $\nu = 6$. All other computations remain the same.

The value of the check standard for one such sequence is defined as the difference between the estimated values of the two reference standards for the sequence as

$$c = \frac{1}{5} (2d_1 - d_2 - d_5 - d_6 - d_7 + d_8 + d_{10}) \qquad (4.3.6)$$

4.3.2 Process Parameters

Initial values of the process parameters are obtained from n such designs, yielding check standard values $c_1,\cdots,c_n$ and within standard deviations $s_{w_1},\cdots,s_{w_n}$. The accepted value of the check standard is defined as the mean of the check standard values; namely,

$$A_c = \frac{1}{n} \sum_{i=1}^{n} c_i \quad . \qquad (4.3.7)$$

The accepted value of the within standard deviation, describing short-term phenomena that affect the measurements within the design, is the pooled value

$$s_p = \left(\frac{1}{n} \sum_{i=1}^{n} s_{w_i}^2 \right)^{1/2} \qquad (4.3.8)$$

with degrees of freedom $\nu_1 = \nu \cdot n$.

The total standard deviation of the check standard is defined as

$$s_c = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_i - A_c)^2 \right)^{1/2} \qquad (4.3.9)$$

with $\nu_2 = n-1$ degrees of freedom.

The model assumed for the process is the additive model (1.4.2). Under this model the error structure for the check standard measurement and the error structure for the reported value of an individual test item are such that the appropriate standard deviation for a value reported for a test item is

$$s_r = \frac{\sqrt{3}}{2} s_c \ .$$

The control limits[k] that are appropriate for future check standard values are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \text{ .}$$

4.3.3 Control Procedure

A test statistic t_c that depends on the observed value of the check standard c is computed for each design by

$$t_c = \frac{|c - A_c|}{s_c} \text{ .} \qquad (4.3.10)$$

The control procedure depends upon this test statistic and the within standard deviation s_w for that design. A dual control procedure is applied as follows:

If $$t_c < 3 \qquad (4.3.11)$$

and if $$s_w < s_p \sqrt{F_\alpha(\nu, \nu_1)} \qquad (4.3.12)$$

for α chosen suitably small, the process is in control and values of the test items are reported as

$$\begin{aligned} X^* &= \hat{x} + R^* \\ Y^* &= \hat{y} + R^* \qquad (4.3.13) \\ Z^* &= \hat{z} + R^* \text{ .} \end{aligned}$$

The restraint is defined as $R^* = \frac{1}{2}(R_1{}^* + R_2{}^*)$ where $R_1{}^*$ and $R_2{}^*$ are the assigned values of the reference standards.

If $$t_c > 3,$$

the calibration of the test items is invalid and must be repeated.

4.3.4 Transfer with NBS

Given three transfer standards T_1, T_2, and T_3, the transfer with NBS could be accomplished in one of several ways such as including only one transfer standard in each design. The most straightforward way is to let the transfer standards take the place ot the test items X, Y, and Z in the design. The calibration design is repeated p times, and process control should be confirmed for each repetition as defined by (4.3.11) and (4.3.12).

[k] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

Any design that is out-of-control should be repeated until control is reestablished or else that design is deleted from the transfer. If the values assigned to the transfer standards by NBS are T_1^*, T_2^*, and T_3^* with uncertainties U_{T1}, U_{T2}, and U_{T3} respectively, the uncertainty of the transfer is

$$U_{tr} = \frac{3}{2\sqrt{15p}} s_c + \frac{1}{3}\left(U_{T1}^2 + U_{T2}^2 + U_{T3}^2\right)^{1/2}. \qquad (4.3.14)$$

A characteristic of the design that is not always recognized is that the offset Δ of the laboratory process from NBS is defined only in terms of the restraint and not in terms of individual reference standards. The reference standards should not be used separately and, if one standard is replaced, the value of the remaining standard and the replacement standard must be reestablished in relationship to NBS.

Given the p values assigned to each transfer standard by (4.3.13); namely,

$$X_1^*,\cdots,X_p^*$$
$$Y_1^*,\cdots,Y_p^*$$
$$Z_1^*,\cdots,Z_p^*,$$

the offset is computed as

$$\Delta = \frac{1}{3p}\sum_{i=1}^{p}(X_i^* + Y_i^* + Z_i^*) - \frac{1}{3}(T_1^* + T_2^* + T_3^*). \qquad (4.3.15)$$

The offset is judged significant if

$$\tilde{t} > 3 , \qquad (4.3.16)$$

where

$$\tilde{t} = \frac{2\sqrt{15p}\ |\Delta|}{s_c}. \qquad (4.3.17)$$

and in such case the assigned value of the restraint R^* is changed to $R^* - \Delta$.

The restraint is unchanged if $\tilde{t} < 3$.

4.3.5 Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by (4.3.13) from one design is

$$U = \frac{3\sqrt{3}}{2} s_c + U_{tr}. \qquad (4.3.18)$$

4.4 Comparator Process for Mass Calibrations with One Check Standard for Each Series

4.4.1 Measurement Sequence

High precision mass determination is done by a sequence of intercomparisons that relate the mass of an object to the laboratory's kilogram reference standards which in turn are related to the Paris kilogram. An entire weight set may require several series of intercomparisons in order to assign values to all weights. The weights in each series are intercompared by a statistical design that prescribes the weighings. Each weighing involves a mass difference between two nominally equal weights or groups of weights. Values assigned thereby are least-squares estimates from the design. Provision for a check standard is included with the weights for each series. The reader is referred to Cameron et al. [5] for the statistical theory governing weighing designs; to Jaeger and Davis [54] for the physical theory; to Varner [55] for a description of the NBS software for mass determination; and to Appendix A in this publication for a description of the matrix manipulations needed for a solution to general weighing designs and the propagation of standard deviations and uncertainties through several series.

Normally the first series involves two kilogram reference standards, R_1 and R_2, a test kilogram X_{10}, and a summation Σ_1 of other weights totaling one kilogram nominally. The restraint is the average of the values assigned to R_1 and R_2, and the check standard is defined as the difference between R_1 and R_2 as estimated from the design.

The value assigned to the summation Σ_1 by the first series constitutes the restraint for the second series with the individual weights in the summation being calibrated separately in the second series. For example, if a 500 gram, a 300 gram, and a 200 gram weight make up the summation totaling one kilogram, those weights are assigned values in the second series of intercomparisons. Two series are needed to calibrate a weight set consisting of 1kg, 500g, 300g, 200g, and 100g weights, say. A summation of weights Σ_2 which becomes the restraint for third series is included in the second series if the weight set is to be extended to 50g, 30g, 20g, and 10g weights, and the calibration is extended to lesser weights in like manner.

The weighing designs for two such series are described generically as a 1, 1, 1, 1 design and a 5, 3, 2, 1, 1, 1 design representing the ratios of the weights in the series to each other. A design consists of a subset of all possible intercomparisons that can be made on the group of weights with several factors dictating this choice. A design is always constructed so that the standard deviation of reported values for weights of the same nominal size are equal. The number of intercomparisons is kept small, less than twenty, so that the weighings can be completed with thermal effects being minimized. Furthermore, the number of weights that one is willing to have on the pan at one time and the maximum load of the balance have some bearing on the choice of observations.

Two designs satisfying these criteria are shown below for calibrating the aforementioned weight set. These designs are used routinely in the NBS calibration program. Six observations designated by $d_1, \cdots, d_6$ suffice for the first series. A check standard for the first series is constructed by differencing the values of R_1 and R_2 that were estimated from the design. The second series has eleven observations designated by $d_1, \cdots, d_{11}$. Notice that a 100g weight designated as C is included in this design as a check standard. An observation for a single pan balance is defined as the mass difference between the weights marked by (+) and the weights marked by a (-) as defined by Jaeger and Davis [54].

Design for 1st Series

Obs	1kg	1kg	1kg	1kg
	R_1	R_2	X_{10}	Σ_1
d_1	+	-		
d_2	+		-	
d_3	+			-
d_4		+	-	
d_5		+		-
d_6			+	-

Design for 2nd Series

Obs	500g	300g	200g	100g	100g	100g
	X_5	X_3	X_2	X_1	Σ_2	C
d_1	+	-	-	+	-	
d_2	+	-	-		+	-
d_3	+	-	-	-		+
d_4	+	-	-			
d_5	+		-	-	-	-
d_6		+	-	+	-	-
d_7		+	-	-	+	-
d_8		+	-	-	-	+
d_9			+	-	-	
d_{10}			+	-		-
d_{11}			+		-	-

4.4.2 Process Parameters

The check standard for the first series is defined as

$$c_1 = (1/4)\{2d_1 + d_2 + d_3 - d_4 - d_5\} . \qquad (4.4.1)$$

The check standard for the second series is defined as

$$c_2 = (1/920)\{4d_1 - 111d_2 + 119d_3 + 4d_4 - 108d_5 - 102d_6 - 102d_7 + 128d_8 - 10d_9 - 125d_{10} - 125d_{11}\}. \qquad (4.4.2)$$

The within standard deviation for the first series is

$$s_{w_1} = \left(\frac{1}{4} \sum_{i=1}^{6} \xi_i{}^2 \right)^{1/2} \tag{4.4.3}$$

with $\nu_1 = 4$ degrees of freedom.

The deviations ξ_i that are needed to compute s_{w_1} are defined by:

$$\begin{aligned}
\xi_1 &= d_1 - (1/4)\ [2d_1 - d_2 - d_3 + d_4 + d_5] \\
\xi_2 &= d_2 - (1/4)\ [-d_1 + 2d_2 - d_3 - d_4 + d_6] \\
\xi_3 &= d_3 - (1/4)\ [-d_1 - d_2 + 2d_3 - d_5 - d_6] \\
\xi_4 &= d_4 - (1/4)\ [d_1 - d_2 + 2d_4 - d_5 + d_6] \\
\xi_5 &= d_5 - (1/4)\ [d_1 - d_3 - d_4 + 2d_5 - d_6] \\
\xi_6 &= d_6 - (1/4)\ [d_2 - d_3 + d_4 - d_5 + 2d_6]\ .
\end{aligned}$$

The within standard deviation for the second series is

$$s_{w_2} = \left(\frac{1}{6} \sum_{i=1}^{11} \xi_i{}^2 \right)^{1/2} \tag{4.4.4}$$

with $\nu_2 = 6$ degrees of freedom.

The deviations needed to compute the within standard deviation s_{w_2} are defined as follows:

$$\begin{aligned}
\xi_1 &= d_1 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 - \hat{x}_1 + \hat{\Sigma}_2 \\
\xi_2 &= d_2 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 - \hat{\Sigma}_2 + c_2 \\
\xi_3 &= d_3 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 + \hat{x}_1 - c_2 \\
\xi_4 &= d_4 - \hat{x}_5 + \hat{x}_3 + \hat{x}_2 \\
\xi_5 &= d_5 - \hat{x}_5 + \hat{x}_2 + \hat{x}_1 + \hat{\Sigma}_2 + c_2 \\
\xi_6 &= d_6 - \hat{x}_3 + \hat{x}_2 - \hat{x}_1 + \hat{\Sigma}_2 + c_2 \\
\xi_7 &= d_7 - \hat{x}_3 + \hat{x}_2 + \hat{x}_1 - \hat{\Sigma}_2 + c_2 \\
\xi_8 &= d_8 - \hat{x}_3 + \hat{x}_2 + \hat{x}_1 + \hat{\Sigma}_2 - c_2 \\
\xi_9 &= d_9 - \hat{x}_2 + \hat{x}_1 + \hat{\Sigma}_2 \\
\xi_{10} &= d_{10} - \hat{x}_2 + \hat{x}_1 + c_2 \\
\xi_{11} &= d_{11} - \hat{x}_2 + \hat{\Sigma}_2 + c_2
\end{aligned} \tag{4.4.5}$$

where

$$\begin{aligned}
\hat{x}_5 &= (1/920)\ \{100(d_1 + d_2 + d_3 + d_4) + 60d_5 \\
&\qquad - 20(d_6 + d_7 + d_8 + d_9 + d_{10} + d_{11})\} \\
\hat{x}_3 &= (1/920)\ \{-68(d_1 + d_2 + d_3 + d_4) - 4d_5 + 124(d_6 + d_7 + d_8) \\
&\qquad - 60(d_9 + d_{10} + d_{11})\} \\
\hat{x}_2 &= (1/920)\ \{-32(d_1 + d_2 + d_3 + d_4) - 56d_5 - 104\ (d_6 + d_7 + d_8) \\
&\qquad + 80(d_9 + d_{10} + d_{11})\} \\
\hat{x}_1 &= (1/920)\ \{119d_1 + 4d_2 - 111d_3 + 4d_4 - 108d_5 + 128d_6 \\
&\qquad - 102(d_7 + d_8) - 125(d_9 + d_{10}) - 10d_{11}\} \\
\hat{\Sigma}_2 &= (1/920)\ \{-111d_1 + 119d_2 + 4(d_3 + d_4) - 108d_5 - 125d_6 + 128d_7 \\
&\qquad - 102d_8 - 125d_9 - 10d_{10} - 125d_{11}\}
\end{aligned} \tag{4.4.6}$$

Accepted values for the check standards, within standard deviations, and total standard deviations are obtained from n initial repetitions of the two series. Check standard values $c_{11},\cdots,c_{1n}$ and $c_{21},\cdots,c_{2n}$ from the respective series are averaged to obtain accepted values,

$$A_{c_1} = \frac{1}{n} \sum_{i=1}^{n} c_{1i}$$

and (4.4.7)

$$A_{c_2} = \frac{1}{n} \sum_{i=1}^{n} c_{2i} \quad .$$

Similarly, within standard deviations $s_{w_{11}},\cdots,s_{w_{1n}}$ from the first series and $s_{w_{21}},\cdots,s_{w_{2n}}$ from the second series are pooled to obtain accepted within standard deviations for the two series:

$$s_{p_1} = \left(\frac{1}{n} \sum_{i=1}^{n} s_{w_{1i}}^{2} \right)^{1/2}$$

and (4.4.8)

$$s_{p_2} = \left(\frac{1}{n} \sum_{i=1}^{n} s_{w_{2i}}^{2} \right)^{1/2} \quad .$$

The total standard deviations for the check standards for each series are respectively

$$s_{c_1} = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_{1i} - A_{c_1})^2 \right)^{1/2}$$

and (4.4.9)

$$s_{c_2} = \left(\frac{1}{n-1} \sum_{i=1}^{n} (c_{2i} - A_{c_2})^2 \right)^{1/2} \quad .$$

4.4.3 Control Procedure[ℓ]

Statistical control is maintained on the measurements by series. For the first series, test statistics computed from the current check standard value c_1, and the within standard deviation s_{w_1} are used to test for control. Let

$$t_{c_1} = \frac{|A_{c_1} - c_1|}{s_{c_1}} \quad . \tag{4.4.10}$$

If

$$t_{c_1} < 3 \tag{4.4.11a}$$

and if

$$s_{w_1} < s_{p_1} \sqrt{F_\alpha(4,4n)} \tag{4.4.11b}$$

for α chosen suitably small, the measurement process is in control, and the following values are assigned to the the test weight X_{10} and summation Σ_1:

$$X_{10}^* = -(1/8)\{3d_2 + d_3 + 3d_4 + d_5 - 2d_6\} + R^*$$

$$\Sigma_1^* = -(1/8)\{d_2 + 3d_3 + d_4 + 3d_5 + 2d_6\} + R^* \tag{4.4.12}$$

where $R^* = \frac{1}{2}(R_1^* + R_2^*)$ and R_1^* and R_2^* are the corrections to nominal size for the kilogram standards R_1 and R_2.

Statistical control for the second series depends upon the current check standard value c_2 and within standard deviation s_{w_2} for that series. Let

$$t_{c_2} = \frac{|A_{c_2} - c_2|}{s_{c_2}} \quad . \tag{4.4.13}$$

If

$$t_{c_2} < 3 \tag{4.4.14a}$$

and if

$$s_{w_2} < s_{p_2} \sqrt{F_\alpha(6,6n)} \tag{4.4.14b}$$

the measurement process is in control for that series.

Equations (4.4.10) and (4.4.13) are the simplest constructions for testing for offset using a t statistic. The technique for constructing these statistics follows the general method for t statistics; namely, the difference between

[ℓ] The factor 3 is used in this and all subsequent computations in place of the appropriate factor of the t distribution; namely, $t_{\alpha/2}(\nu)$.

the current value of the check standard and its accepted value divided by the standard deviation of the check standard. As such the construction is applicable to any design. In this case the statistic defined by (4.4.10) is precisely correct if the data base for check standard C_1 comes from identical designs with identical restraints, and similarly for the statistic deifined by (4.4.13). In practice a check standard, especially C_2, can be utilized in a variety of designs. This does not affect the interpretation of the accepted value of the check standard, but it does affect the interpretation of the total standard deviation. In such case the test statistics can be computed using the within standard deviations as follows:

$$t_{c_1} = \frac{\sqrt{2}\,|A_{c_1} - c_1|}{s_{w_1}} \tag{4.4.10a}$$

$$t_{c_2} = \frac{|A_{c_2} - c_2|}{\left(\frac{29}{230}\, s_{w_2}^{\,2} + \frac{1}{100} \cdot \frac{3}{8}\, s_{w_1}^{\,2}\right)^{1/2}} \tag{4.4.13a}$$

These equations are compatible with the documenation in reference [55] where the between component of variance is assumed to be zero --an assumption that is true for the NBS mass calibration process. Notice that the construction of the relevant t statistic becomes increasingly complicated as one moves through the series of weighings depending as it does on the within standard deviations from all prior designs. See Appendix A for the general construction for any design.

Given that (4.4.14a) and (4.4.14b) are satisfied, values are reported for test items and summation for the next series as follows:

Weights	Reported Values	
500g	$X_5^* = \hat{x}_5 + \frac{1}{2}\,\Sigma_1^*$	
300g	$X_3^* = \hat{x}_3 + \frac{276}{920}\,\Sigma_1^*$	
200g	$X_2^* = \hat{x}_2 + \frac{184}{920}\,\Sigma_1^*$	(4.4.15)
100g	$X_1^* = \hat{x}_1 + \frac{92}{920}\,\Sigma_1^*$	
Σ100g	$\Sigma_2^* = \hat{\Sigma}_2 + \frac{92}{920}\,\Sigma_1^*$	

where $\hat{x}_5$, $\hat{x}_3$, $\hat{x}_2$, $\hat{x}_1$ and $\hat{\Sigma}_2$ are defined in (4.4.6) and Σ_1^* is defined in (4.4.12). Whenever a series is out-of-control, the calibration results for the test weights in that series are invalid and must be repeated.

4.4.4 Transfer with NBS

For a mass measurement assurance program the laboratory's starting kilograms are calibrated at NBS and assigned values R_1^* and R_2^* and associated uncertainties U_{R1} and U_{R2}. The transfer is accomplished by relating all weighings to these standards as explained in section 4.4.1.

4.4.5 Uncertainty[m]

The uncertainty associated with the value assigned to any weight is a function of the design and the within standard deviations for that series and all prior series. It also includes as systematic error a proportional part of the uncertainty associated with the starting restraint. For example, the uncertainty for the value assigned to the one kilogram summation Σ_1^* which is the starting restraint for the second series is U_{1000} where

$$U_{1000} = 3\sqrt{k_1}\, s_{w_1} + \frac{1}{2}(U_{R1} + U_{R2}), \quad k_1 = \frac{3}{8} \quad . \qquad (4.4.16)$$

The uncertainties for the 500g, 300g, 200g, and 100g test weights are respectively:

$$U_{500} = 3\left(k_2\, s_{w_2}^2 + \frac{3}{8}\, m_2^{\,2} s_{w_1}^2\right)^{1/2} + m_2(U_{R1} + U_{R2}), \quad k_2 = \frac{50}{920}, \quad m_2 = \frac{1}{2}$$

$$U_{300} = 3\left(k_3\, s_{w_2}^2 + \frac{3}{8}\, m_3^{\,2} s_{w_1}^2\right)^{1/2} + m_3(U_{R1} + U_{R2}), \quad k_3 = \frac{82}{920}, \quad m_3 = \frac{3}{10}$$

$$U_{200} = 3\left(k_4\, s_{w_2}^2 + \frac{3}{8}\, m_4^{\,2} s_{w_1}^2\right)^{1/2} + m_4(U_{R1} + U_{R2}), \quad k_4 = \frac{64}{920}, \quad m_4 = \frac{1}{5}$$

$$U_{100} = 3\left(k_5\, s_{w_2}^2 + \frac{3}{8}\, m_5^{\,2} s_{w_1}^2\right)^{1/2} + m_5(U_{R1} + U_{R2}), \quad k_5 = \frac{116}{920}, \quad m_5 = \frac{1}{10}$$

[m] Uncertainties are computed assuming the between component of variance is zero. See reference [55] for the general construction.

4.5 Comparator Process for Four Reference Standards and Four Test Items

4.5.1 Measurement Sequence

This design for four reference standards and four test items involves the intercomparison of items two at a time where each test item is intercompared with each standard one time, and there is no direct intercomparison among standards or test items. The design is routinely used for voltage measurements where the laboratory's reference standards R_1, R_2, R_3 and R_4 in one temperature controlled box are intercompared with test items W, X, Y and Z or transfer standards in another box, and there are no intercomparisons within a box.

Schematically, the intercomparisons are as shown below where a plus (+) or a minus (-) indicates relative position in the circuit.

Ref Test	R_1	R_2	R_3	R_4
W	+	-	+	-
X	-	+	-	+
Y	+	-	+	-
Z	-	+	-	+

Measurements on the laboratory standards R_1, R_2, R_3, and R_4 and the test items W, X, Y and Z are designated by r_1, r_2, r_3 and r_4 and w, x, y, and z respectively. The design consists of the following sequence of difference measurements:

$$\begin{aligned}
d_1 &= r_1 - w \\
d_2 &= r_1 - y \\
d_3 &= r_3 - y \\
d_4 &= r_3 - w \\
d_5 &= r_2 - x \\
d_6 &= r_2 - z \\
d_7 &= r_4 - z \\
d_8 &= r_4 - x \\
d_9 &= x - r_1 \\
d_{10} &= z - r_1 \\
d_{11} &= z - r_3 \\
d_{12} &= x - r_3 \\
d_{13} &= w - r_2 \\
d_{14} &= y - r_2 \\
d_{15} &= y - r_4 \\
d_{16} &= w - r_4
\end{aligned} \qquad (4.5.1)$$

The design has several features that make it particularly suitable for intercomparing saturated standard cells. Let the observations d_i, ordered as in (4.5.1) so as to minimize the number of circuit connections, represent the differences in emf between two cells as measured by a potentiometer. The convention adhered to is, for example, that r_1-w represents the measured difference between R_1 and W with the cells reversed in the circuit relative to their positions for the difference w-r_1.

The design is balanced so as to cancel out any spurious emf that may be present in the circuit [56]. In the presence of such systematic error, called left-right effect, the measurements d_i are assumed to be related to the actual differences D_i in emf between two cells in the following way:

$$d_i = D_i + \zeta + \varepsilon_i \qquad i = 1,\cdots,16$$

where ζ is the left-right effect, and ε_i is random error. For a circuit with negligible left-right effect, one expects that the measurements would sum to zero except for the effect of random error. Any disparity between this expectation and the summation gives an estimate of the magnitude of left-right effect; namely,

$$\hat{\zeta} = \frac{1}{16} \sum_{i=1}^{16} d_i \; . \tag{4.5.2}$$

A measuring process such as the one described in the foregoing paragraph can be characterized by:

i) a short-term or within standard deviation which describes variability during the time necessary to make the sixteen measurements for one design.

ii) accepted values for check standards which have been specifically chosen for this measurement situation.

iii) a total standard deviation for the process based on the check standard measurements.

The difference of each test item from the average of the reference group is computed by:

$$\begin{aligned}
\hat{w} &= -\frac{1}{4}(d_1 + d_4 - d_{13} - d_{16})\\
\hat{x} &= -\frac{1}{4}(d_5 + d_8 - d_9 - d_{12})\\
\hat{y} &= -\frac{1}{4}(d_2 + d_3 - d_{14} - d_{15})\\
\hat{z} &= -\frac{1}{4}(d_6 + d_7 - d_{10} - d_{11})
\end{aligned} \tag{4.5.3}$$

The foregoing quantities in conjunction with the differences of the reference standards from their group average; namely,

$$\begin{aligned}
\hat{r}_1 &= \frac{1}{16}\,(3d_1+3d_2-d_3-d_4-d_5-d_6-d_7-d_8-3d_9-3d_{10}+d_{11}+d_{12}+d_{13}+d_{14}+d_{15}+d_{16}) \\
\hat{r}_2 &= \frac{1}{16}\,(-d_1-d_2-d_3-d_4+3d_5+3d_6-d_7-d_8+d_9+d_{10}+d_{11}+d_{12}-3d_{13}-3d_{14}+d_{15}+d_{16}) \\
\hat{r}_3 &= \frac{1}{16}\,(-d_1-d_2+3d_3+3d_4-d_5-d_6-d_7-d_8+d_9+d_{10}-3d_{11}-3d_{12}+d_{13}+d_{14}+d_{15}+d_{16}) \\
\hat{r}_4 &= \frac{1}{16}\,(-d_1-d_2-d_3-d_4-d_5-d_6+3d_7+3d_8+d_9+d_{10}+d_{11}+d_{12}+d_{13}+d_{14}-3d_{15}-3d_{16})
\end{aligned} \tag{4.5.4}$$

and the estimated left-right effect $\hat{\zeta}$ are used to estimate a within standard deviation s_w for each design; namely,

$$s_w = \left(\frac{1}{8} \sum_{i=1}^{16} \xi_i^{\,2} \right)^{1/2} \tag{4.5.5}$$

with $\nu=8$ degrees of freedom. The individual deviations ξ_i are given by:

$$\begin{aligned}
\xi_1 &= d_1 - \hat{r}_1 + \hat{w} - \hat{\zeta} \\
\xi_2 &= d_2 - \hat{r}_1 + \hat{y} - \hat{\zeta} \\
\xi_3 &= d_3 - \hat{r}_3 + \hat{y} - \hat{\zeta} \\
\xi_4 &= d_4 - \hat{r}_3 + \hat{w} - \hat{\zeta} \\
\xi_5 &= d_5 - \hat{r}_2 + \hat{x} - \hat{\zeta} \\
\xi_6 &= d_6 - \hat{r}_2 + \hat{z} - \hat{\zeta} \\
\xi_7 &= d_7 - \hat{r}_4 + \hat{z} - \hat{\zeta} \\
\xi_8 &= d_8 - \hat{r}_4 + \hat{x} - \hat{\zeta} \\
\xi_9 &= d_9 - \hat{x} + \hat{r}_1 - \hat{\zeta} \\
\xi_{10} &= d_{10} - \hat{z} + \hat{r}_1 - \hat{\zeta} \\
\xi_{11} &= d_{11} - \hat{z} + \hat{r}_3 - \hat{\zeta} \\
\xi_{12} &= d_{12} - \hat{x} + \hat{r}_3 - \hat{\zeta} \\
\xi_{13} &= d_{13} - \hat{w} + \hat{r}_2 - \hat{\zeta} \\
\xi_{14} &= d_{14} - \hat{y} + \hat{r}_2 - \hat{\zeta} \\
\xi_{15} &= d_{15} - \hat{y} + \hat{r}_4 - \hat{\zeta} \\
\xi_{16} &= d_{16} - \hat{w} + \hat{r}_4 - \hat{\zeta}
\end{aligned} \tag{4.5.6}$$

Check standards for electrical measurements are not easily defined because of the inherent nature of electrical quantities to drift over time. For this reason, three separate check standards are recommended for measurements on standard cells. The left-right effect reflects many of the sources of error in the measurement system and can be presumed to remain stable over time. For this reason it makes a suitable check standard for process control. Specifically, the value of the first check standard is defined for each design as ζ from (4.5.2).

There is also a need to check on the stability of the reference standards, changes or instabilities in which may not be reflected in the left-right effect. The least-squares estimates for the reference standards from the design (4.5.4) cannot be used to check on the stability of the standards themselves because these estimates are in effect a consequence of the design, subject to the restraint, and are not meaningful separately. For example, if the restraint is changed to exclude one of the reference standards, the least-squares estimates for the remaining reference standards as computed from the same observed differences (4.5.1) can change appreciably.

The information in a design does, however, allow a way of monitoring the change in one reference standard relative to another reference standard. A measured difference between two reference standards that is not subject to the restraint can be computed from each design, and two check standards, each one involving the difference between two reference standards, are recommended for monitoring the stability of the four reference standards.

Check standard C_1 is defined for the difference between R_1 and R_3, and check standard C_2 is defined for the difference between R_2 and R_4. Their respective values c_1 and c_2 are computed for each design as follows:

$$\begin{aligned} c_1 &= \frac{1}{4}(d_1 + d_2 - d_3 - d_4 - d_9 - d_{10} + d_{11} + d_{12}) \\ c_2 &= \frac{1}{4}(d_5 + d_6 - d_7 - d_8 - d_{13} - d_{14} + d_{15} + d_{16}) \,. \end{aligned} \tag{4.5.7}$$

Because it is anticipated that the change in one reference standard relative to another may not be stable over time, the method for analyzing check standards C_1 and C_2 is a modified process control technique that allows for linear drift.

4.5.2 Process Parameters for Stable and Drifting Check Standards

Initial values for the process parameters are established from n repetitions of the design in which the four reference standards are compared to any four test items. The resulting check standard measurements are $\zeta_1,\cdots,\zeta_n$; $c_{11},\cdots,c_{1n}$; and $c_{21},\cdots,c_{2n}$. For the left-right effect the n values are averaged to obtain the accepted value

$$A_\zeta = \frac{1}{n} \sum_{i=1}^{n} \hat{\zeta}_i . \qquad (4.5.8)$$

A total standard deviation for the left-right effect is also computed from the initial check standard measurements by

$$s_\zeta = \left(\frac{1}{(n-1)} \sum_{i=1}^{n} (\hat{\zeta}_i - A_\zeta)^2 \right)^{1/2} \qquad (4.6.9)$$

with $\nu = (n-1)$ degrees of freedom.

The control limits[o] that are appropriate for future measurements on the left-right effect are:

$$\text{Upper Control Limit} = A_\zeta + 3s_\zeta$$

$$\text{Lower Control Limit} = A_\zeta - 3s_\zeta .$$

Similar calculations of accepted values and standard deviations are made for C_1 and C_2 where the check standard measurements $c_{11},\cdots,c_{1n}$ and $c_{21},\cdots,c_{2n}$ are stable over time. More often than not these quantities are not stable over time, and this fact must be taken into account in the analysis. If the check standard values show drift and if the drift is linear with time, check standard values $c_1,\cdots,c_n$ at times $t_1,\cdots,t_n$ can be characterized by

$$c_i = \alpha + \beta t_i \qquad i=1,\cdots,n$$

where the intercept α and the slope β are estimated by

$$\hat{\alpha} = \bar{c} - \hat{\beta}\,\bar{t}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (t_i-\bar{t})(c_i-\bar{c})}{\sum_{i=1}^{n} (t_i-\bar{t})^2}$$

[o] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

with $$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \quad \text{and} \quad \bar{c} = \frac{1}{n} \sum_{i=1}^{n} c_i.$$

In the linear case the accepted total standard deviation for each check standard is

$$s_c = \left(\frac{1}{n-2} \sum_{i=1}^{n} (c_i - \hat{\alpha} - \hat{\beta} t_i)^2 \right)^{1/2} \qquad (4.5.10)$$

with $\nu = n-2$ degrees of freedom. See reference [59] for analyses relating to linear regression models.

The parameters of the linear fit and associated standard deviations should be computed for C_1 and C_2 separately resulting in estimates α_1, β_1, s_{c_1} with $\nu_1 = n-2$ degrees of freedom for check standard C_1 and α_2, β_2, s_{c_2} with $\nu_2 = n-2$ degrees of freedom for check standard C_2. The value that a check standard is expected to take on at any given time is thus dependent on the linear fit. Therefore, for a future time t', provided t' is not too far removed from t_n, the accepted values for the check standards are defined by

$$A'_{c_1} = \hat{\alpha}_1 + \hat{\beta}_1 t'$$

and (4.5.11)

$$A'_{c_2} = \hat{\alpha}_2 + \hat{\beta}_2 t' .$$

A total standard deviation for the measurements on C_1 and C_2 can be pooled from s_{c_1} and s_{c_2} by the formula

$$s_c = \left(\frac{1}{2} \left(s_{c_1}^2 + s_{c_2}^2 \right) \right)^{1/2} \qquad (4.5.12)$$

with $\nu = 2(n-2)$ degrees of freedom.

The control procedure assumes that t' is close to t_n because the standard deviation of a predicted value from a linear fit increases dramatically as the linear fit is extrapolated beyond the check standard data. Thus the chance of detecting a real shift in the process diminishes as the tests for control are continued into the future. This fact necessitates frequent updating of the parameters of the linear fit based on recent check standard values.

Furthermore, the control procedure and the assumption of a linear model are interdependent. Because there is no way of separating these two elements, an out-of-control signal can be caused by either lack of process control or a breakdown in the linearity of the check standard measurements. One must recognize this as a short-coming in the control procedure and arrange for other independent checks on the stability of the reference standards.

The control procedure also makes use of the accepted within standard deviation s_p which is not dependent upon model assumptionsfor the check standards . It is computed from the within standard deviations $s_{w_1},\cdots,s_{w_n}$ for each design as follows:

$$s_p = \left(\frac{1}{n} \sum_{i=1}^{n} s_{w_i}^2 \right)^{1/2} \qquad (4.5.13)$$

with $\nu_3 = 8n$ degrees of freedom.

4.5.3 Process Control

Process control is maintained by monitoring the within standard deviation for each design and the performance of the three designated check standards. If check standard C_1 or C_2 repeatedly fails the test for control, it is likely that one of the two reference standards comprising the check standard has changed in value. In this case it will be necessary to replace one or both of the standards in question or reestablish their values relative to NBS.

Process control should be verified for the within standard deviation s_w as it is calculated for each design and for the current values of the check standards for that design; namely, ζ, c_1, and c_2. For the left-right effect ζ, the test statistic is:

$$t_\zeta = \frac{|\hat{\zeta} - A_\zeta|}{s_\zeta} . \qquad (4.5.14)$$

For check standards C_1 and C_2 that are drifting linearly over time the corresponding test statistics at time t' are:

$$t_{c_1} = \frac{|c_1 - A'_{c_1}|}{\tilde{s}}$$

and

$$t_{c_2} = \frac{|c_2 - A'_{c_2}|}{\tilde{s}} \qquad (4.5.15)$$

where

$$\tilde{s} = s_c \left(\frac{n+1}{n} + \frac{(t' - \bar{t})^2}{\sum_{i=1}^{n} (t_i - \bar{t})^2} \right)^{1/2} .$$

Then the following conditions can be imposed:

$$\text{If } t_\zeta \text{ and } t_{c_1} \text{ and } t_{c_2} \text{ are all } < 3 \qquad (4.5.16a)$$

and if

$$s_w < s_p\sqrt{F_\alpha(8,\nu_3)} \qquad (4.5.16b)$$

for α suitably small, the process is judged in control for that design.

The values of the test items are reported as

$$\begin{aligned} W^* &= \hat{w} + R^* \\ X^* &= \hat{x} + R^* \\ Y^* &= \hat{y} + R^* \\ Z^* &= \hat{z} + R^* \end{aligned} \qquad (4.5.17)$$

where the restraint $R^* = \frac{1}{4}(R_1^* + R_2^* + R_3^* + R_4^*)$, and R_1^*, R_2^*, R_3^*, and R_4^* are the values assigned to the laboratory's reference standards.

If the results of the control procedures along with other experimental evidence indicate instability or other anomalous behavior on the part of one of the reference standards, the entire experiment need not necessarily be discarded. It is possible to delete the reference standard in question from the restraint and obtain new values for the test items if the values of the remaining reference standards are known individually. For example, if one is involved in a transfer with NBS, and if reference standard R_1 shows signs of serious malfunction after several days of intercomparisons between the reference standards and the transfer standards, the values for the transfer standards can be recomputed for each design as follows:

$$\hat{w} = \frac{1}{48}\{-9d_1+3d_2-d_3-13d_4-d_5-d_6-d_7-d_8-3d_9-3d_{10}+d_{11}+d_{12}+13d_{13}+d_{14}+d_{15}+13d_{16}\}$$

$$\hat{x} = \frac{1}{48}\{3d_1+3d_2-d_3-d_4-13d_5-d_6-d_7-13d_8+9d_9-3d_{10}+d_{11}+13d_{12}+d_{13}+d_{14}+d_{15}+d_{16}\}$$

$$(4.5.18)$$

$$\hat{y} = \frac{1}{48}\{3d_1-9d_2-13d_3-d_4-d_5-d_6-d_7-d_8-3d_9-3d_{10}+d_{11}+d_{12}+d_{13}+13d_{14}+13d_{15}+d_{16}\}$$

$$\hat{z} = \frac{1}{48}\{3d_1+3d_2-d_3-d_4-d_5-13d_6-13d_7-d_8-3d_9+9d_{10}+13d_{11}+d_{12}+d_{13}+d_{14}+d_{15}+d_{16}\}$$

and W^*, X^*, Y^* and Z^* are computed according to (4.5.17) with the restraint R^* changed to:

$$R^* = \frac{1}{3}\,(R_2^* + R_3^* + R_4^*).$$

The differences of reference standards R_2, R_3 and R_4 from their average value are recomputed to be:

$$\begin{aligned}
\hat{r}_2 &= \frac{1}{12}\,\{-d_3-d_4+2d_5+2d_6-d_7-d_8+d_{11}+d_{12}-2d_{13}-2d_{14}+d_{15}+d_{16}\} \\
\hat{r}_3 &= \frac{1}{12}\,\{2d_3+2d_4-d_5-d_6-d_7-d_8-2d_{11}-2d_{12}+d_{13}+d_{14}+d_{15}+d_{16}\} \qquad (4.5.19) \\
\hat{r}_4 &= \frac{1}{12}\,\{-d_3-d_4-d_5-d_6+2d_7+2d_8+d_{11}+d_{12}+d_{13}+d_{14}-2d_{15}-2d_{16}\}
\end{aligned}$$

The within standard deviation for each design (see equations (4.5.5) and (4.5.6)) can be computed using either the original quantities in (4.5.3) and (4.5.4) or the adjusted quantities in (4.5.18) and (4.5.19) with identical results.

4.5.4 Transfer with NBS

Transfer with NBS is accomplished by means of p repetitions of the design in which four transfer standards T_1, T_2, T_3, and T_4 replace the four test items. If one of the tests for control defined by (4.5.16a) and (4.5.16b) is not satisfied, the design should be repeated or else that repetition should be deleted from the transfer.

Given p repetitions of the design in which T_1 replaces W, T_2 replaces X, T_3 replaces Y and T_4 replaces Z, the p values assigned to each transfer standard by the participant's process are computed from (4.5.17); namely,

$$\begin{gathered}
W_1^*,\cdots,W_p^* \\
X_1^*,\cdots,X_p^* \\
Y_1^*,\cdots,Y_p^* \\
Z_1^*,\cdots,Z_p^*.
\end{gathered}$$

NBS assigns values to electrical transfer standards that take into account their individual and collective behavior both before, during, and after their sojourn in the participant's laboratory. A transfer standard that displays unstable behavior during one of these periods may be excluded from the analysis. Normally the averages for the four transfer standards from the "before and after" NBS determinations are fit by least-squares to a linear function of time; then average values T_j^* are predicted for the times $t_j (j=1,\cdots,p)$ that the transfer standards were in the participant's laboratory by the equation

$$T_j^* = \hat{\alpha}_o + \hat{\beta}_o t_j \qquad\qquad j=1,\cdots,p$$

where $\hat{\alpha}_o$ and $\hat{\beta}_o$ are estimated from NBS measurements.

This makes it possible to compute daily offsets $\Delta_j (j=1,\cdots,p)$ for the reference group where

$$\Delta_j = \frac{1}{4} (W_j{}^* + X_j{}^* + Y_j{}^* + Z_j{}^*) - T_j{}^* \qquad j=1,\cdots,p \quad (4.5.20)$$

and assuming the reference group is stable, an average offset for the reference group is computed by

$$\bar{\Delta} = \frac{1}{p} \sum_{j=1}^{p} \Delta_j. \qquad (4.5.21)$$

The offset is judged significant if

$$\tilde{t} > 3$$

where $$\tilde{t} = \frac{4\sqrt{p}|\bar{\Delta}|}{(4s_c{}^2 - s_p{}^2)^{1/2}} .$$

In such case the value of the laboratory restraint is changed to $R^* - \Delta$.

Otherwise, the restraint is unchanged.

4.5.5 Uncertainty

The uncertainty of the transfer is

$$U_{tr} = \frac{3(4s_c{}^2 - s_p{}^2)^{1/2}}{4\sqrt{p}} + U_T \qquad (4.5.22)$$

where U_T is the uncertainty assigned to the transfer standards by NBS.

The uncertainty that is appropriate for the laboratory's process as it assigns a value to a test item based on a single design is

$$U = \frac{3}{4} (10s_c{}^2 - s_p{}^2)^{1/2} + U_{tr} . \qquad (4.5.23)$$

4.5.6 Example

An example is presented from the Volt Transfer Program where an environmentally controlled box of four standard cells was sent to an industrial participant to be intercompared with the participant's box of four standard cells. After the NBS cells had been in the participant's laboratory for two weeks, thereby giving them a chance to recover from the trip, the laboratory's reference cells were intercompared with the NBS cells each day for 16 days using the design described in sec 4.5.1. The data corrected for the temperature in each box are shown in exhibit 4.5.1.

Exhibit 4.5.1 - Intercomparison of laboratory standard cells with NBS cells
Value in microvolts

Day Obs	1	2	3	4	5	6	7	8
d_1	86.70	86.92	86.86	86.98	86.97	86.99	87.07	87.17
d_2	87.28	87.02	86.77	86.69	86.57	86.60	86.63	86.68
d_3	89.29	89.01	88.75	88.64	88.52	88.52	88.54	88.57
d_4	88.84	88.98	88.87	88.94	88.97	88.94	88.98	89.10
d_5	88.06	87.97	87.84	87.89	88.33	88.17	88.03	88.11
d_6	87.43	87.04	86.94	86.94	86.96	86.92	86.92	87.07
d_7	88.96	88.58	88.46	88.47	88.45	88.47	88.50	88.55
d_8	89.63	89.52	89.39	89.43	89.85	89.70	89.63	89.60
d_9	-86.47	-86.60	-86.32	-86.34	-86.71	-86.71	-86.61	-86.71
d_{10}	-85.81	-85.70	-85.40	-85.41	-85.39	-85.49	-85.54	-85.66
d_{11}	-87.82	-87.67	-87.34	-87.37	-87.34	-87.40	-87.42	-87.54
d_{12}	-88.49	-88.62	-88.25	-88.32	-88.72	-88.64	-88.54	-88.59
d_{13}	-88.80	-89.11	-88.84	-88.92	-88.90	-88.88	-88.92	-89.03
d_{14}	-89.15	-89.16	-88.73	-88.64	-88.48	-88.47	-88.48	-88.55
d_{15}	-90.90	-90.69	-90.24	-90.13	-89.93	-89.98	-90.02	-90.08
d_{16}	-90.38	-90.64	-90.33	-90.40	-90.35	-90.41	-90.49	-90.60

Day Obs	9	10	11	12	13	14	15	16
d_1	87.25	87.28	87.32	87.45	87.46	87.50	87.53	87.59
d_2	86.72	86.80	86.81	86.87	86.90	86.91	86.92	86.97
d_3	88.60	88.52	88.52	88.62	88.59	88.59	88.59	88.62
d_4	89.13	89.07	89.09	89.16	89.18	89.17	89.21	89.24
d_5	88.09	88.00	87.78	88.12	88.05	88.06	88.04	88.05
d_6	87.07	86.99	86.89	87.17	87.15	87.09	87.07	87.07
d_7	88.58	88.56	88.62	88.69	88.82	88.68	88.69	88.70
d_8	89.60	89.55	89.60	89.68	89.79	89.67	89.65	89.68
d_9	-86.66	-86.66	-86.74	-86.78	-86.78	-86.84	-86.92	-86.89
d_{10}	-85.63	-85.67	-85.79	-85.79	-85.80	-85.88	-85.96	-85.93
d_{11}	-87.52	-87.48	-87.55	-87.58	-87.58	-87.60	-87.59	-87.60
d_{12}	-88.57	-88.47	-88.51	-88.53	-88.57	-88.54	-88.57	-88.57
d_{13}	-89.04	-89.00	-89.01	-89.10	-89.06	-89.10	-89.10	-89.12
d_{14}	-88.53	-88.44	-88.47	-88.55	-88.47	-88.51	-88.55	-88.53
d_{15}	-90.07	-90.00	-90.05	-90.12	-90.14	-90.10	-90.12	-90.14
d_{16}	-90.60	-90.54	-90.58	-90.68	-90.69	-90.69	-90.75	-90.77

Exhibit 4.5.2 - Estimates for transfer standards and reference standards
Values in microvolts

	NBS Standard Cells				Laboratory Reference Cells				L-R
Day	T_1	T_2	T_3	T_4	R_1	R_2	R_3	R_4	Effect
	$\tilde{w}$	$\hat{x}$	$\hat{y}$	$\hat{z}$	$\hat{r}_1$	$\hat{r}_2$	$\hat{r}_3$	$\hat{r}_4$	$\hat{\zeta}$
1	-88.68	-88.16	-89.15	-87.50	-1.811	-0.016	0.234	1.592	-0.102
2	-88.91	-88.17	-88.96	-87.24	-1.767	-0.007	0.243	1.531	-0.197
3	-88.72	-87.94	-88.62	-87.03	-1.746	+0.004	0.219	1.522	-0.098
4	-88.80	-87.99	-88.50	-87.04	-1.739	+0.003	0.223	1.513	-0.097
5	-88.80	-88.40	-88.38	-87.04	-1.743	+0.015	0.235	1.492	-0.075
6	-88.81	-88.31	-88.40	-87.07	-1.696	-0.033	0.232	1.497	-0.104
7	-88.87	-88.20	-88.42	-87.10	-1.683	-0.058	0.225	1.515	-0.108
8	-88.97	-88.25	-88.46	-87.20	-1.671	-0.036	0.224	1.482	-0.119
9	-89.00	-88.23	-88.48	-87.20	-1.664	-0.047	0.226	1.486	-0.098
10	-88.98	-88.17	-88.44	-87.18	-1.587	-0.082	0.196	1.473	-0.093
11	-89.00	-88.16	-88.46	-87.22	-1.543	-0.171	0.209	1.504	-0.129
12	-89.10	-88.28	-88.54	-87.31	-1.583	-0.071	0.167	1.487	-0.086
13	-89.10	-88.30	-88.53	-87.34	-1.579	-0.132	0.166	1.546	-0.072
14	-89.12	-88.28	-88.53	-87.32	-1.526	-0.118	0.167	1.477	-0.099
15	-89.15	-88.30	-88.55	-87.30	-1.496	-0.139	0.161	1.474	-0.116
16	-89.18	-88.30	-88.57	-87.33	-1.497	-0.149	0.166	1.481	-0.102

Figures 4-7 show the individual behavior of the transfer standards, and figure figure 8 shows the behavior of the transfer group on the average. One might conclude based on these graphs that the cells were not sufficiently stabilized at the beginning of the experiment and that the first two measurements in the participant's laboratory should be deleted from the transfer data.

The differences for the transfer cells and the reference cells from their group means (See equations (4.5.3) and (4.5.4)) are listed in exhibit 4.5.2. The behavior of the reference cells during the transfer with NBS is of interest because the final assignment of offset depends on the assumption that the reference cells are stable. As was noted earlier in this section, the quantities listed in exhibit 4.5.2 do not describe the behavior of the individual reference cells because these quantities are constrained so that their sum is equal to zero.

The only way to observe the individual cells during the transfer is to reverse the way in which the assignments are currently made; i.e., to analyze the data from the intercomparisons using the reference cells as unknowns and the value of the transfer group from NBS as the restraint. This will give individual values for each reference cell and can be done after the fact if the transfer group proves sufficiently stable. The rationalization for computing an offset using the reference cells as the restraint is that one would expect the reference cells, if they are of the same quality as the transfer cells, to be more stable considering they have not recently been in transit.

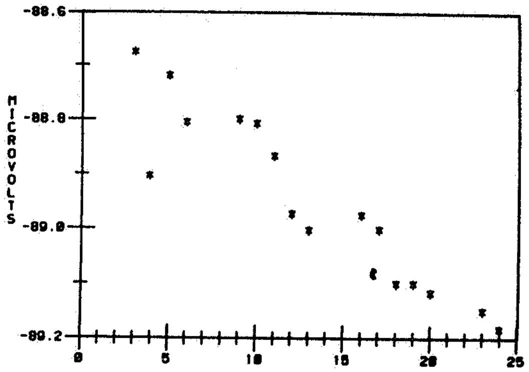


Figure 4
Values (μV) assigned to transfer standard T_1 versus time (days)

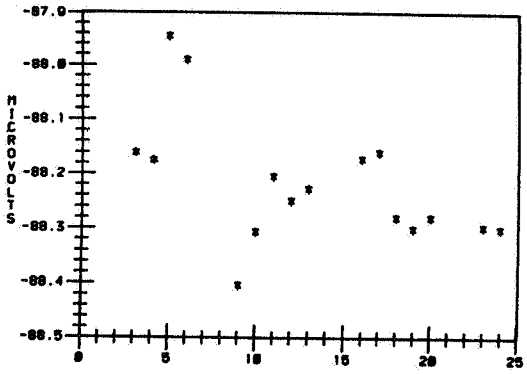


Figure 5
Values (μV) assigned to transfer standard T_2 versus time (days)

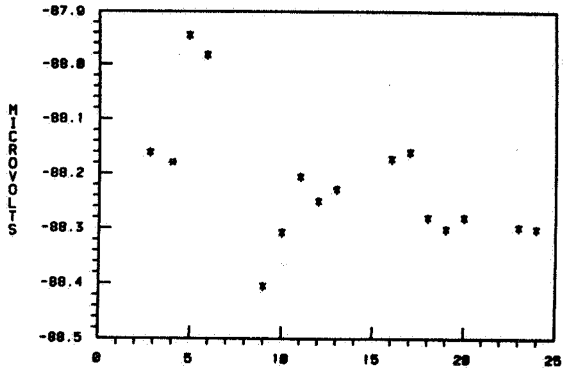


Figure 6
Values (μV) assigned to transfer standard T_3 versus time (days)

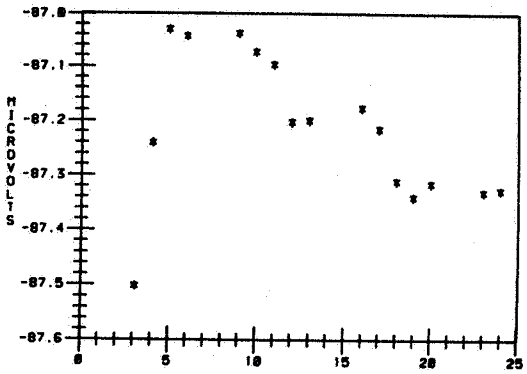


Figure 7
Values (μV) assigned to transfer standard T_4 versus time (days)

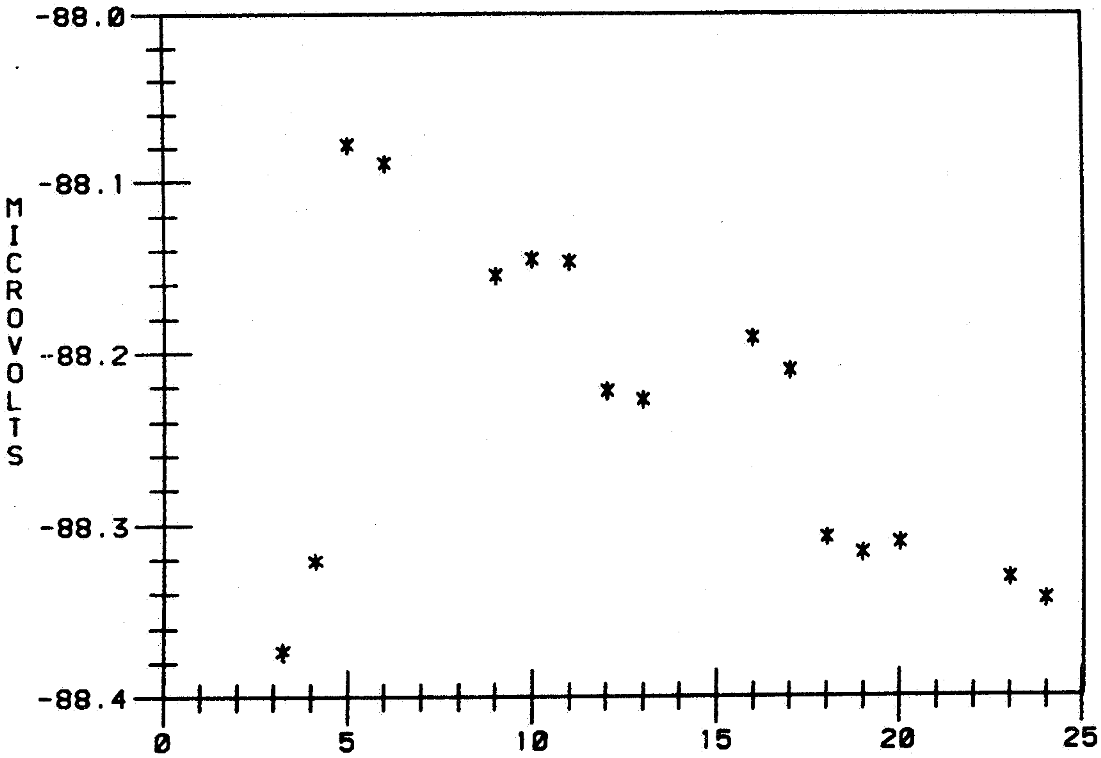


Figure 8
Average values (μV) assigned to four transfer standards versus time (days)

Each day's intercomparisons are analyzed for internal consistency via an F test on the within standard deviation for that day. The stability of the three designated check standards is also tested each day. Results of those designs which show evidence of lack of statistical control or anomalous behavior on the part of one of the check standards are excluded from the transfer experiment. Because we do not have prior history on this measurement process, we rely on hypothetical data to demonstrate to the reader the analysis that should be done for each design.

The left-right effects (4.5.2) are plotted in figure 9. Their respective test statistics (4.5.8) are listed in exhibit 4.5.3. Upper and lower control limits in figure 9 are indicated by dashed lines, and points that fall outside these control limits are equivalent to the corresponding test statistics being significant. These computations assume that prior data on the left-right effect established a standard deviation for the left-right effect of $s_\zeta = 0.02\mu V$ with $\nu_1 = 50$ degrees of freedom and that the accepted value of the left-right effect was established as $A_\zeta = 0.100\mu V$ from the same data.

Check standards C_1 and C_2 as constructed in (4.5.7) are observed differences between two reference cells and do not depend on the restraint or the design. Tracked over a period of time they show the way in which two cells are changing in respect to each other. Their values are listed in exhibit 4.5.3 and plotted as a function of time in figures 10-11.

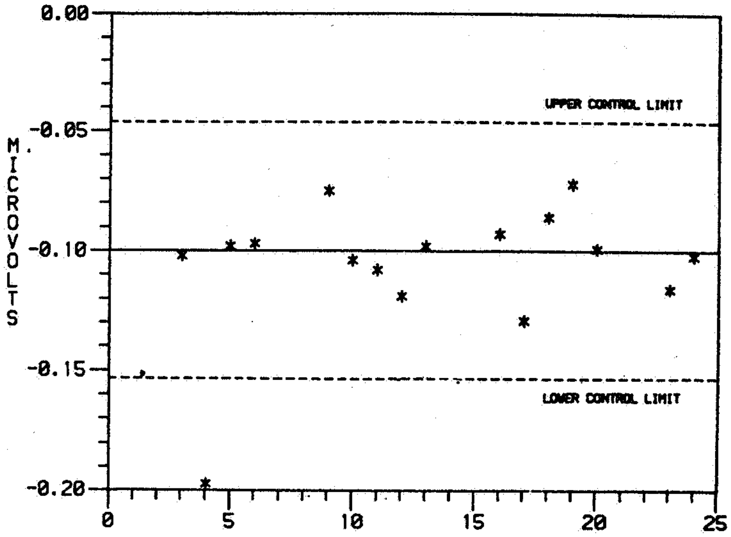


Figure 9
Left-right effect (µV) plotted against time (days) with dashed lines indicating upper and lower control limits at the 1% significance level

Exhibit 4.5.3 - check standards and test statistics[†]
Values in microvolts

Run	Date	Left-right Effect	Test Stat	Check Std	Test Stat	Check Std	Test Stat
#	t'	$\hat{\zeta}$	t_ζ	c_1	t_{c_1}	c_2	t_{c_2}
1	3	-0.102	0.1	-2.0450	0.22	-1.6075[¶]	2.90[¶]
2	4	-0.197[§]	4.8[§]	-2.0100	0.29	-1.5375	0.51
3	5	-0.098	0.1	-1.9650	1.10	-1.5175	0.29
4	6	-0.097	0.2	-1.9625	0.58	-1.5100	0.68
5	9	-0.075	1.2	-1.9775	1.66	-1.4775	2.16
6	10	-0.104	0.2	-1.9275	0.69	-1.5300	0.69
7	11	-0.108	0.4	-1.9075	0.66	-1.5725	0.46
8	12	-0.119	1.0	-1.8950	0.85	-1.5175	1.37
9	13	-0.098	0.1	-1.8900	1.27	-1.5300	1.14
10	16	-0.093	0.4	-1.7825	0.25	-1.5550	0.84
11	17	-0.129	1.4	-1.7525	0.58	-1.6675	2.35
12	18	-0.086	0.7	-1.7500	0.09	-1.5575	1.05
13	19	-0.072	1.4	-1.7450	0.32	-1.6775	2.31
14	20	-0.099	0.0	-1.6925	0.65	-1.5950	0.25
15	23	-0.116	0.8	-1.6575	0.01	-1.6125	0.18
16	24	-0.102	0.1	-1.6625	0.66	-1.6300	0.17

[†] We choose to illustrate the control procedure at the 1% significance level.

[§] Failed test for control at 1% level of significance based on a critical value $t_{.005}(50)$ = 2.678 from Table I.

[¶] Failed test for control at 1% level of significance based on a critical value $t_{.005}(100)$ = 2.626 from Table I.

For this analysis, it was assumed that data from fifty-one initial designs established a linear relationship with time for each check standard as follows:

$$\begin{aligned} c_1 &= -2.095 + 0.0190t \\ c_2 &= -1.501 - 0.00513t \end{aligned} \qquad (4.5.25)$$

and that standard deviations, s_{c_1} for C_1 and s_{c_2} for C_2, were pooled to form a process standard deviation $s_c = 0.030\mu V$ with $\nu = 100$ degrees of freedom.

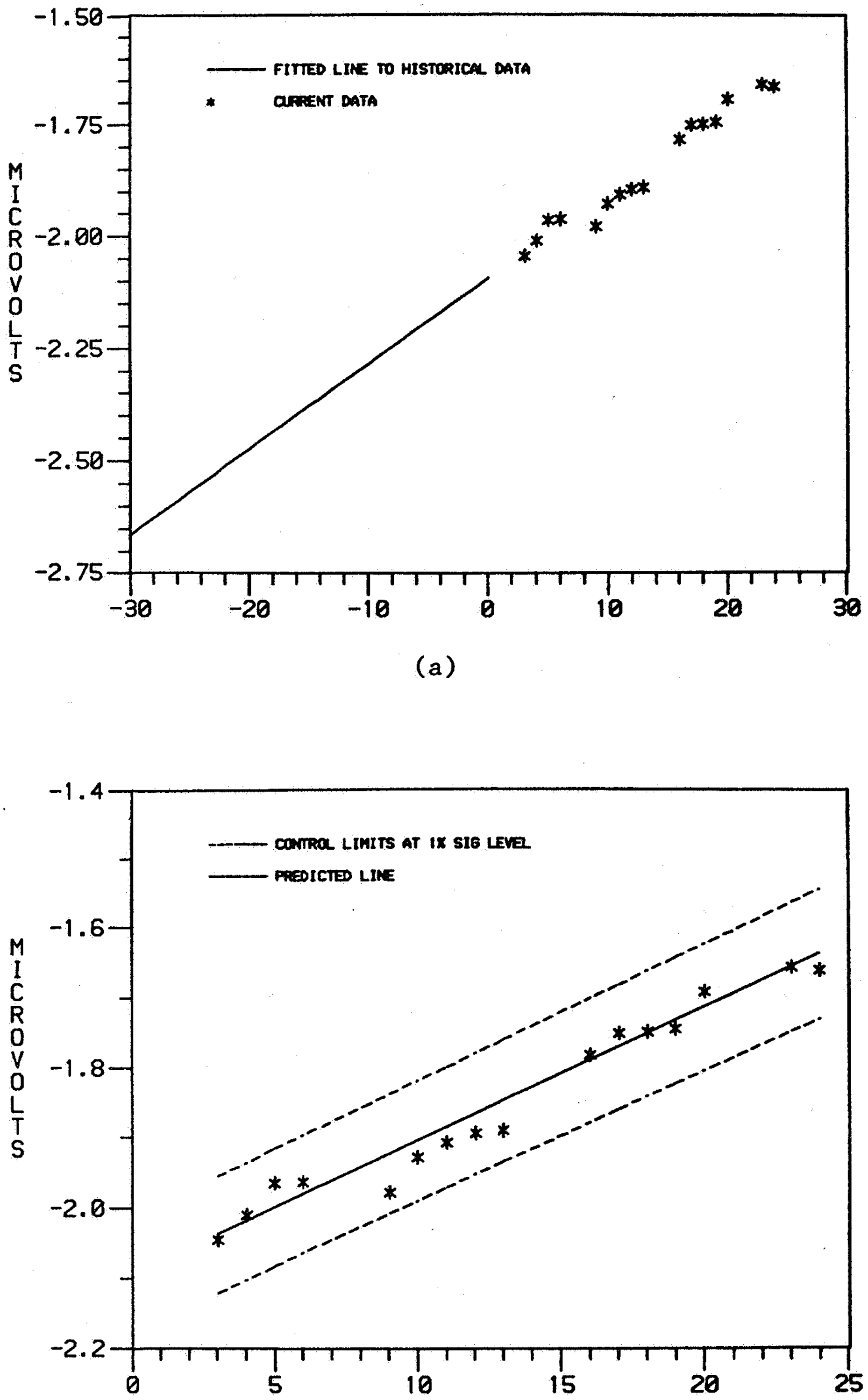
Based on the foregoing assumption, predicted values (4.5.11) for the check standards were computed for each time t' that the transfer standards were measured in the participant's laboratory. Given this information, the check standard measurements on each day were tested for agreement with the extrapolated line by the test statistics listed in exhibit 4.5.3. The test statistics for C_1 and C_2 that are shown in exhibit 4.5.3 were computed from (4.5.15) with $n = 31$ and values of $t_i(i=1,\cdots,31) = -30(1)0$.

The same analysis is shown graphically in figures 10-11. The upper portion of figure 10 shows the linear fit to the historical data as a solid line, and the values of check standard C_1 for the transfer experience are shown as discrete points, (*) with the convention that the transfer experiment starts at time $t = 0$.

The lower portion of figure 10 shows the analysis of the check standard measurements. The solid line is an extrapolation of the linear fit from the upper portion of the same figure to the time of the transfer experiment. The dashed lines are upper and lower control limits that show the range within which the check standard measurements are expected to deviate from the extrapolated line. A point being outside these control limits is exactly analogous to the corresponding test statistic being significant in exhibit 4.5.3. Although it is not readily apparent from the graph, the control limits become wider as the check standard measurements are further removed in time from their historical data base. Thus, there is a smaller chance of detecting anomalous behavior as the experiments are continued into the future if the data base is not updated frequently.

Figure 11 shows the same analysis for the values of check standard C_2 from the transfer experiment with check standard C_2 out-of-control on the first day.

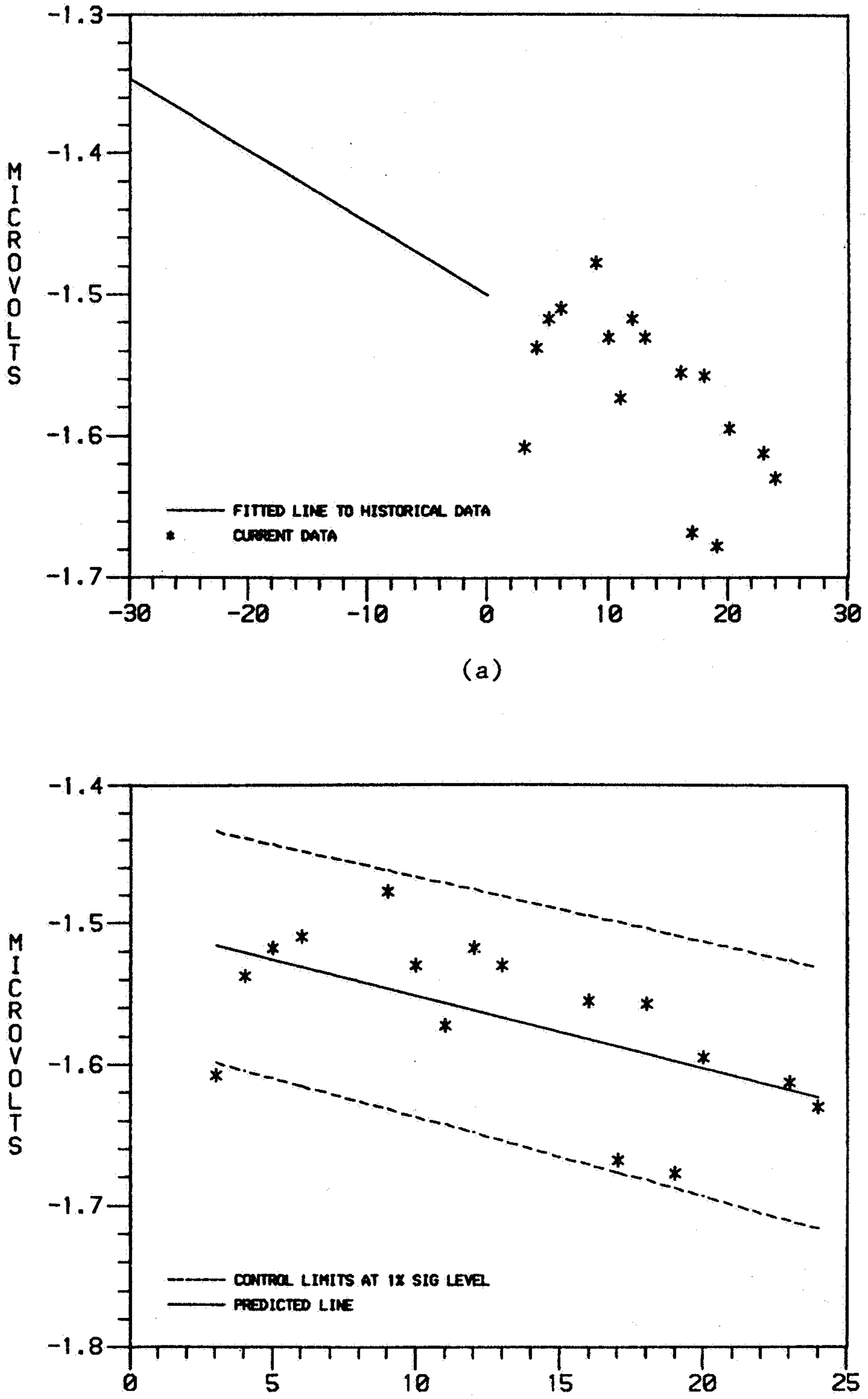
The within standard deviations are listed in exhibit 4.5.4 and plotted in figure 12. An F test based on an accepted standard deviation $s_p = 0.02\mu V$ with $\nu_3 = 408$ degrees of freedom indicates that there are measurement problems on the first and eleventh days. It is interesting to note that check standard C_2 is low on the eleventh day although it is not actually out-of-control and that the left-right effect is very close to being out-of-control on that same day. Given the responses of the check standards and the transfer standards and the information garnered from the control procedure, it would seem reasonable to delete three measurements from the transfer data; namely, the first, second and eleventh days' measurements.



(a)

(b)

Figure 10
Check standard C_1 (μV) plotted against time (days)
with a solid line indicating a predicted linear fit and dashed lines
indicating upper and lower control limits at the 1% level of significance



(a)

(b)

Figure 11
Check standard C_2 (µV) plotted against time (days)
with a solid line indicating a predicted linear fit and dashed lines
indicating upper and lower control limits at the 1% level of significance.

Exhibit 4.5.4 - Within sandard deviations and test statistics
Values in microvolts

Run	Date	Within SD	DF
#	t	s_W	ν_3
1	3	0.054§	8
2	4	0.018	8
3	5	0.019	8
4	6	0.015	8
5	9	0.022	8
6	10	0.011	8
7	11	0.016	8
8	12	0.021	8
9	13	0.013	8
10	16	0.021	8
11	17	0.054§	8
12	18	0.018	8
13	19	0.031	8
14	20	0.013	8
15	23	0.018	8
16	24	0.011	8

§Failure to satisfy the inequality $s_w < s_p\sqrt{F_{.01}(8,\infty)}$ at the 1% significance level based on $s_p = 0.02\mu V$ and a critical value $F_{.01}(8,\infty) = 2.5$ from Table II.

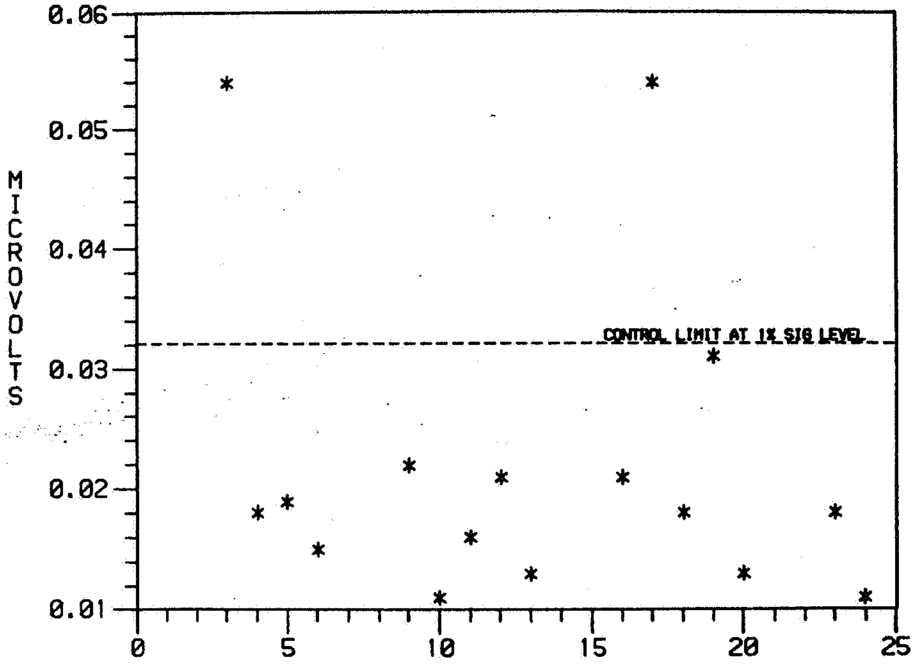


Figure 12
Within standard deviations (μV) plotted against time (days)
with dashed line indicating control limit at 1% level of significance.

4.6 Direct Reading of the Test Item with an Instrument Standard

4.6.1 Measurement Sequence

In this mode of operation a value is directly assigned to a test item X by a calibrated instrument. Observations on a stable artifact that takes on the role of the check standard C are used to establish a base line for the instrument and to maintain and control its variability in what amounts to a surveillance type test. An observation on the test item is denoted by x, and an observation on the check standard is denoted by c.

4.6.2 Process Parameters

Initial values of the process parameters are obtained from n independent measurements on the check standard $c_1,\cdots,c_n$. The accepted value of the check standard is defined by the mean of the check standard measurements; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i \; . \tag{4.6.1}$$

The total standard deviation of the instrument is

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n}(c_i - A_c)^2\right)^{1/2} \tag{4.6.2}$$

with $\nu = n-1$ degress of freedom. The control limits[n] that are appropriate for future observations on the check standard are given by

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \; .$$

4.6.3 Control Procedure

The primary purpose of the control procedure is to monitor instrumental drift, and observations on the check standard should be taken frequently enough to ensure that such drift is being contained. A test statistic t_c computed from the most recent check standard measurement c is given by

$$t_c = \frac{|c - A_c|}{s_c} \; .$$

The process is in control at the time of the check standard measurement c if

$$t_c < 3 \; . \tag{4.6.3}$$

[n] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

If

$$t_c > 3 ,$$

the process is not in control at the time of the check standard measurement, and measurements should be discontinued until the problem with the instrument is rectified.

4.6.4 Transfer with NBS

Determination of systematic error can be made by making p measurements $r_1,\cdots,r_p$ on a calibrated artifact or transfer standard which has an assigned value T^* and associated uncertainty U_T. Instrumental offset ψ defined by

$$\psi = \frac{1}{p}\sum_{i=1}^{p} r_i - T^* \qquad (4.6.4)$$

is not significant if

$$\frac{\sqrt{p}\,|\psi|}{s_c} < 3 . \qquad (4.6.5)$$

It is extremely important to recognize that this approach makes two important assumptions that must be verified experimentally; namely, that the instrument has a constant offset from the NBS process over the regime of interest as in (1.4.2) and that the precision of the instrument is constant over this same regime. The question of constant offset is considered first. A single point is not sufficient for the determination, and the system must be checked using several calibrated artifacts that span the regime of interest. Assume that m transfer standards are sufficient to verify the points of interest and that the transfer standards have assigned values $T_1^*,\cdots,T_m^*$ and associated uncertainties $U_{T1},\cdots,U_{Tm}$. Assume also that m offsets $\psi_1,\cdots,\psi_m$ computed according to (4.6.4) have been determined from measurements made on the transfer standards.

If all ψ_j $(j=1,\cdots,m)$ are insignificant as judged by (4.6.5), no adjustment to the instrument is needed. If the offsets are of varying magnitudes, and if it can be shown that these offsets are functionally related to the assigned values of the transfer standards, it may be possible to calibrate the instrument using a calibration curve based on the offsets (see section 2.3.3). Finally, if the offsets are significant and of the same magnitude, either the instrument is adjusted for the average offset

$$\bar{\psi} = \frac{\sum_{j=1}^{m} p_j \cdot \psi_j}{\sum_{j=1}^{m} p_j} \qquad (4.6.6)$$

where p_j $(j=1,\cdots,m)$ represents the number of measurements on the jth transfer standard or a reading x on a test item x is reported as

$$X^* = x - \bar{\psi} .$$

The uncertainty of the transfer is

$$U_{tr} = \frac{3s_c}{\sqrt{p_1+\cdots+p_m}} + \frac{1}{m}\left(U_{T1}^{2} + \cdots + U_{Tm}^{2}\right)^{1/2} . \qquad (4.6.7)$$

4.6.5 Uncertainty

The total uncertainty that is appropriate for one measurement made on a test item using the calibrated instrument is

$$U = U_{tr} + 3s_c. \qquad (4.6.8)$$

4.6.6 Process Precision.

The question concerning whether or not the precision of an instrument remains constant over a given regime can be addressed by comparing standard deviations from several levels in the regime. A familiar example is an electronic balance that is used over a large range of loads where the precision of the instrument may be load dependent. This assumption can be checked either with calibrated or uncalibrated artifacts.

Standard deviations with their associated degrees of freedom should be tabulated by load and inspected for consistency. It is possible to quote one uncertainty over the entire regime only if the precision is constant over all load levels; i.e., if these standard deviations are all of the same magnitude.

A visual inspection of the values may be sufficient for determining whether or not the standard deviations are of roughly the same magnitude in which case the standard deviations should be pooled using (2.2.3) and the uncertainty computed by replacing s_c in equations (4.6.7) and (4.6.8) with the pooled standard deviation.

If there is some question about the propriety of combining all the standard deviations, the largest standard deviation can be checked for agreement with the others using a test developed by Cochran [57]. A description of the test statistic and tables for deciding whether or not the largest standard deviation in a group is significantly different from the group are tabulated by Eisenhart [58].

If it is logical to assume that the precision of the instrument will vary with the magnitude of the quantity of interest, then a series of check standards should be established, one at each level of interest, with the estimate of process precision (4.6.2), the test for statistical control (4.6.3), and computation of uncertainty (4.6.8) being made at each level independently, thus begging the question of constant variability.

4.7 Simultaneous Measurement of a Group of Test Items and a Group of Reference Standards

4.7.1 Measurement Sequence

This scheme is appropriate for assigning values to individual test items or instruments relative to the average of a bank or group of reference standards, called the restraint R*, when all items including the standards are simultaneously subjected to the same stimuli such as a power source or a vacuum chamber. Assume there are m reference standards $R_1,\cdots,R_m$, and ℓ test items $X_1,\cdots,X_\ell$. One position in the configuration of test items should be reserved for a check standard Y, an artifact similar to the test items, where a reading on Y is always recorded along with the other readings.

Assume that a measurement sequence produces readings $r_1,\cdots,r_m$ on the standards, $x_1,\cdots,x_\ell$ on the test items and y on the check standards. The value that is recorded as the check standard measurement for one sequence is

$$c = y - \frac{1}{m}\sum_{i=1}^{m} r_i \ . \qquad (4.7.1)$$

In other words the measured difference between the artifact check standard and the average of the reference standards is the check standard measurement. In the remainder of this section, the term check standard refers to this recorded difference rather than the measured value y.

4.7.2 Process Parameters

Initial values of the process parameters are obtained from n such measurement sequences where $c_1,\cdots,c_n$ are the check standard measurements.

The accepted value of the check standard is the mean of these values; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i \ . \qquad (4.7.2)$$

The total standard deviation of the check standard is

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n}(c_i - A_c)^2\right)^{1/2} \ . \qquad (4.7.3)$$

Control limits[q] that are appropriate for future check standard observations are given by

$$\text{Upper Control Limit} = A_c + 3s_c$$

$$\text{Lower Control Limit} = A_c - 3s_c \ .$$

[q] The factor 3 is used in this and all subsequent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

The control procedure applied to each calibration depends on a test statistic t_c computed from the value of the check standard c for that measurement sequence by

$$t_c = \frac{|c - A_c|}{s_c} . \qquad (4.7.4)$$

If

$$t_c < 3 \qquad (4.7.5)$$

the process in control, and the value of a test item is reported as

$$X_j^* = x_j - \frac{1}{m}\sum_{i=1}^{m} r_i + R^* \qquad j=1,\cdots,\ell \quad (4.7.6)$$

where $R^* = \frac{1}{m}(R_1^* + \cdots + R_m^*)$ and $R_1^*,\cdots,R_m^*$ are the values assigned to the reference standards. If

$$t_c > 3$$

the calibration of the test items is invalid and must be repeated.

4.7.3 Transfer with NBS

The transfer with NBS is accomplished by p repetitions of the measurement sequence during which a group of ℓ transfer standards $T_1,\cdots,T_\ell$ replaces the group of test items. Process control as defined by (4.7.5) should be confirmed for each repetition. Any sequence that is out-of-control should be repeated until control is restored or else that repetition is deleted from the transfer. The values assigned the transfer standards are $T_1^*,\cdots,T_\ell^*$ with uncertainties $U_{T1},\cdots,U_{T\ell}$.

The offset Δ_i $(i=1,\cdots,p)$ of the laboratory process from NBS for the ith repetition is based on the values assigned to the ℓ transfer standards by (4.7.6); namely, $X_1^*,\cdots,X_\ell^*$ where

$$\Delta_i = \frac{1}{\ell}\sum_{j=1}^{\ell}(X_j^* - T_j^*) \qquad i=1,\cdots,p$$

and the average offset computed for the p repetitions is

$$\overline{\Delta} = \frac{1}{p}\sum_{i=1}^{p}\Delta_i . \qquad (4.7.7)$$

The uncertainty of the transfer is

$$U_{tr} = \frac{3s_c}{\sqrt{p\ell}} + \frac{1}{\ell}\left(U_{T1}^2 + \cdots + U_{T\ell}^2\right)^{1/2} . \qquad (4.7.8)$$

The offset is judged significant if

$$\frac{\sqrt{p\ell}\,|\Delta|}{s_c} > 3 \qquad (4.7.8)$$

and in such case the assigned value of the restraint is changed to $R^* - \Delta$. The restraint is unchanged if

$$\frac{\sqrt{p\ell}\,|\Delta|}{s_c} < 3.$$

4.7.5 Uncertainty

The total uncertainty that is appropriate for a value assigned to a test item by (4.7.6) from one calibration is

$$U = U_{tr} + 3s_c. \qquad (4.7.9)$$

4.8 Ratio Technique for One or More Test Items and One or Two Reference Standards

4.8.1 Measurement Scheme

In this section we describe calibration of a test item X by an instrument such as a scanning electron microscope which has only short-term stability. Consider the case where the test item X and the reference standard R are related by (1.4.9) and the instrument response is of the form (1.4.10). One reference standard R is sufficient to provide a calibrated value X^* for the test item given a single reading x on the test item and a single reading r on the reference standard. The calibrated value is

$$X^* = x \cdot R^*/r \qquad (4.8.1)$$

where R^* is the value assigned to the reference standard.

Where the test item and reference standard are related by (1.4.1) and the instrument response is of the form (1.4.6), two reference standards R_1 and R_2 are needed to calibrate a test item X (Cameron [60]). The artifacts should be measured in the sequence R_1, X, R_2 with the corresponding measurements denoted by r_1, x, r_2. The calibrated value for the test item is

$$X^* = R_1{}^* + \frac{(R_2{}^* - R_1{}^*)\cdot(x - r_1)}{(r_2 - r_1)} \qquad (4.8.2.)$$

where $R_1{}^*$ and $R_2{}^*$ are the values assigned to R_1 and R_2 respectively.

If before and after readings are taken on the test item in the sequence X, R_1, R_2, X with the measurements denoted by x_1, r_1, r_2, x_2 respectively, then the calibrated value for the test item is

$$X^* = \frac{1}{2}\left\{ (R_1^* + R_2^*) + \frac{(R_2^* - R_1^*)\cdot(x_1 - r_1 - r_2 + x_2)}{(r_2 - r_1)} \right\} . \qquad (4.8.3)$$

More than one unknown can be calibrated from the same pair of readings on R_1 and R_2 only if the sequence of measurements can be arranged so that no test item is too far removed from R_1 and R_2 in the measurement scheme. For example, for test items X, Y, and Z, the sequence X, R_1, Y, R_2, Z minimizes the separation between unknowns and standards, and the calibrated value for each unknown is calculated according to (4.8.2).

In practice, it may be necessary to have several artifact standards that cover the operating range of the instrument. In addition to artifact standards for every level, it is necessary to have one artifact check standard Y for every level. A measurement y on the check standard should be included in the calibration program on a regular basis, and if feasible, with every calibration scheme. The check standard value that is used for controlling the process and for estimating random error is computed in exactly the same way as X^*. For example, for the measurement sequence described by (4.8.2), the check standard value from one calibration is

$$c = R_1^* + \frac{(R_2^* - R_1^*)\cdot(y - r_1)}{(r_2 - r_1)} . \qquad (4.8.4)$$

4.8.2 Process Parameters

Initial values of the process parameters are obtained from n such calibration sequences yielding check standard values $c_1,\cdots,c_n$. The accepted value of the check standard is defined as the mean of the check standard values; namely,

$$A_c = \frac{1}{n}\sum_{i=1}^{n} c_i . \qquad (4.8.5)$$

The total standard deviation of the check standard is defined by

$$s_c = \left(\frac{1}{n-1}\sum_{i=1}^{n}(c_i - A_c)^2\right)^{1/2} \qquad (4.8.6)$$

with $\nu = n-1$ degrees of freedom.

In this case s_c is the standard deviation of a calibrated value X^* and will reflect not only the imprecision in the measurements x, r_1, and r_2 but also any changes in the response curve for the instrument that are not accounted for by the ratioing device.

The control limits[r] that are appropriate for future check standard values are:

$$\text{Upper control limit} = A_c + 3s_c$$

$$\text{Lower control limit} = A_c - 3s_c \; .$$

4.8.3 Control Procedure

A control procedure is applied to each calibration sequence which includes a check standard measurement. The control procedure is based on a test statistic t_c computed from the check standard value c for that sequence; namely,

$$t_c = \frac{|c - A_c|}{s_c} \; .$$

If

$$t_c < 3 \qquad (4.8.7)$$

the process is in control, and the value of a test item X is reported as X^*.

If

$$t_c > 3,$$

the process is out-of-control, and the calibration of the test item is invalid and must be repeated.

4.8.4 Transfer with NBS

The tie to NBS is via the reference standards which are either standard reference materials from NBS or secondary calibrated artifacts.

4.8.5 Uncertainty

The uncertainty for an artifact calibrated according to (4.8.1) is

$$U = 3s_c + U_R \qquad (4.8.8)$$

where U_R is the uncertainty for R^*. The uncertainty for an artifact calibrated according to (4.8.2) or (4.8.3) is

$$U = 3s_c + \frac{1}{2}\left(U_{R1}{}^2 + U_{R2}{}^2\right)^{1/2} \qquad (4.8.9)$$

where U_{R1} and U_{R2} are the uncertainties for $R_1{}^*$ and $R_2{}^*$ respectively.

[r] The factor 3 is used in this and all subsquent computations in place of the appropriate percent point of the t distribution; namely, $t_{\alpha/2}(\nu)$.

5. Control Charts

5.1 Introduction

The industrial application of control charts involves a production process that yields product that is assumed to be homogeneous with respect to a particular property that is measurable. The control chart is devised to detect any variation in the production process that is not random in nature and which, therefore, can be assigned a cause. Guaranteeing that all variation in the production process is random in nature guarantees that the process is operating in an optimal fashion, and if, given these circumstances, the product is not within specifications, major adjustments to the process are required in order to substantially affect its output.

Once a base line and control limits have been defined for the process, based on prior data from the same process, the control chart is set up with a solid horizontal line representing the base line and dashed lines above and below the base line representing the control limits. Samples drawn at random from the production process are measured for the property of interest, and the resulting values are plotted on the control chart as a function of time. Values that fall within the control limits are referred to as being "in statistical control" and values that fall outside the control limits are referred to as being "out of control". Values outside the control limits are a sufficient indication that the "process should be investigated and corrected" (Bicking & Gryna [61]).

The Shewhart control chart discussed above is appropriate for individual measurements or averages of "natural" groups. This type of control chart, used in conjunction with a control chart for standard deviations, is a powerful means of detecting changes in the measurement process. Other types of control procedures include a cusum chart (Duncan[62]) which is particularly useful for detecting gradual drifts in a continuous process as compared with abrupt shifts. Methods for detecting changes in both the base line of the process and in the variability of the process on a single control chart are discussed by Reynolds and Ghosh in reference [63].

Statistical control as originated by Shewhart [64] assumes that repeated measurements of a reproducible property are available and that these measurements constitute a random sample of all such possible measurements from a known distribution such as the normal distribution. The term random sample implies two important properties of the measurements; namely, that they are independent and that they all come from the same distribution. The average value and standard deviation calculated from a random sample in conjunction with known properties of the distribution are used to calculate limits within which a certain percentage of all measurements should fall. In other words, a series of initial measurements are made to characterize the distribution of all possible measurements, and future measurements are checked for conformity with this distribution.

Notice that one is not concerned with whether or not the product is within certain specification limits, but rather with whether or not the production process is behaving properly. The control procedure for a measurement process is similar in many respects to industrial control. In the measurement assurance context the measurement algorithm including instrumentation, reference standards and operator interactions is the process that is to be controlled, and its direct product is measurement per se. The measurements are assumed to be valid if the measurement algorithm is operating in a state of control; i.e., if the variations in that process are due to random causes which can be quantified, thus assuring that a value reported by the process will have negligible offset from national standards within predictable limits. This will be the case if the control chart shows that the base line for the process is not changing.

Statistical control in the measurement assurance context can conversely be predicated on the assumption that the measurement process is stable and that lack of control indicates a change in the artifact being measured. There are circumstances where this type of control is needed--that is, when it is necessary to know whether or not an artifact has changed with respect to the property being measured. For example, a transfer standard that is being circulated to several laboratories must be checked periodically at NBS. Similarly, intercomparisons between working standards and primary standards can be subjected to a control procedure to ensure that the working standards have not changed appreciably. In these instances, lack of control will result in either replacing the artifact in question or in reassigning its accepted value.

Calibration control is perhaps dissimilar to industrial control in that although artifacts submitted for measurement are of the same general type, their properties must be quantified individually. Thus, there is an inherent problem in controlling the values assigned to individual artifacts or instruments because the measurement is rarely repeated, let alone repeated sufficiently often to characterize the distribution of possible values. Without a historical data base there is no way of determining whether or not the current calibration is in control or is, in fact, a proper assignment for the item. For this reason a check standard is introduced into the measurement sequence in such a way that it can be assumed that the measurement algorithm acts on the check standard in much the same way as it acts on the item being calibrated. The redundant measurements on the check standard are the basis for both characterizing the distribution of measurements and deciding if the measurement process is in control on a given occasion.

The control limits are chosen so that the probability is 100α percent that future measurements will fall outside the control limits strictly by chance. Therefore, α if always chosen small, say $\alpha = .01$ or $\alpha = .05$ so that very few measurements will be discarded unnecessarily. Smaller values of α correspond to wider control limits which result in the measurement almost always being accepted unless there is a serious shift in the process. The converse is also true--larger values of α correspond to narrower control limits which result in tighter control of the measurement process with more frequent remeasurement. Obviously, the success that can be expected in detecting changes in the process which is referred to as the power of the control procedure is linked to the choice of α.

The reader may have already noted that the procedure for determining control or lack thereof is exactly analogous to a statistical t-test for deciding whether or not a single observation comes from a process with known mean and unknown standard deviation.

5.2 Control Charts for Single Measurements

The measurements for initiating the control chart must be collected over a sufficiently wide range of operating conditions to ensure a correct characterization of the distribution and over a sufficently long period of time to ensure independence. Grant and Leavenworth state that ideally twenty-five measurements should be spread over several months time [65]. As few as ten or fifteen measurements can suffice if this data base is updated when more measurements are available. The measurements are plotted as a function of time without imposing a base line or control limits on the plot in order to track the measurement process and verify that it produces stable measurements whose variability is random in nature. Such a plot also allows one to check specification limits, but specification limits do not constitute statistical control because they do not have a probabilistic interpretation.

When one is satisfied that the initial measurements are adequate for representing the distribution and that process variability is tolerable, a base line and control limits are computed from this data base.

For single measurements the base line is taken to be the average of initial measurements $x_1, \cdots, x_n$; namely

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (5.2.1)$$

and the control limits are taken to be

$$\begin{aligned} &\bar{x} + s \cdot t_{\alpha/2}(\nu) \\ &\bar{x} - s \cdot t_{\alpha/2}(\nu) \end{aligned} \qquad (5.2.2)$$

where s, the total standard deviation computed from the initial measurements is

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2} \qquad (5.2.3)$$

with $\nu = n-1$ degrees of freedom. The number $t_{\alpha/2}(\nu)$ is the $\alpha/2$ percentage point of Student's t distribution with ν degrees of freedom.

Once the average value and the control limits have been established, future measurements are tested for control. One concludes that measurements that fall within the control limits come from the hypothesized distribution, and that, therefore, the measurement process is acting in an acceptable and predictable manner. The converse is also true. Measurements that fall outside the control limits infer a significant change in the process. Where such a change is noted, one must determine whether the change is permanent or transitory.

In a measurement assurance context, every violation of the control limits requires a remedial action. It may be sufficient to simply repeat the offending measurement in order to reestablish control, but all measurements since the last successful test for control are discarded once an out-of-control condition occurs.

As an example, consider how repeated measurements on a calibrated weight can be used to demonstrate that an electronic balance is, indeed, weighing accurately at all times. Accuracy in this context means that values delivered by the balance are in agreement with national standards (prototype kilogram) as maintained by NBS within the stated uncertainty. Parobeck et al [66] describe a measurement assurance program for large volume weighings on electronic balances where redundancy and control are achieved by repeating weighings of selected test items on different days.

A program to control a weighing process is begun by making n initial measurements on the calibrated weight, being sure to allow enough time between successive measurements to cover a range of operating conditions in the laboratory, and using these initial measurements as a historical base for computing the average $\bar{x}$ and the standard deviation s of the balance.

Given a calibrated value A with uncertainty U_A for the weight, the balance is accurate within the uncertainty $U_A \pm s \cdot t_{\alpha/2}(n-1)$ if

$$A - \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) - U_A < \bar{x} < A + \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) + U_A.$$

Notice that this test takes into account both the limits to random error for the measurement process, $\pm s \cdot t_{\alpha/2}(n-1)/\sqrt{n}$, and the uncertainty associated with the calibrated value of the weight, U_A.

Once the accuracy has been verified, the control phase of the program is pursued by remeasuring the weight from time to time. The resulting values are plotted on a control chart having base line and control limits as defined in equations (5.2.1) and (5.2.3), and it is presumed that the balance continues to be accurate as long as

$$\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) < y_i < \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1)$$

for all future measurements y_i.

There is always a question, in this type of application, of how often one should check for control. It seems obvious, particularly if one is dealing with electronic instrumentation, that there should always be a check for control as part of any start-up procedures. After that, the frequency is dictated by the past performance of the system and by the amount of inconvenience and expense that is generated when an out-of-control condition is encountered--keeping in mind that when the balance is found to be out-of-control, it is necessary to recall all the measurements that were made on that balance since the previous successful check for control.

5.3 Control Charts for Averages or Predicted Values

Thus far, the discussion has centered on control charts for individual measurements, and it is easily extended to include control charts for averages that are completely analagous to the control charts for individual measurements. When the reported value of a measurement sequence, be it an average or a predicted value from a least-squares analysis, is computed from k intercomparisons that were made over a relatively short period of time, the "measurement of interest" is the corresponding average or predicted value of the check standard. This quantity is treated analogously to a single measurement with base line and control limits for the control chart determined from n such initial quantities. That is, given check standard values $x_1,\cdots,x_n$ each of which is an average or predicted value from k intercomparisons, the grand mean $\bar{x}$ computed from (5.2.1) represents the base line of the process and control limits as in (5.2.2) can be calculated using the total standard deviation s from (5.2.3). In this case the quantity s is the standard deviation of an average or predicted value and not the standard deviation of a single measurement from the process.

5.4 Control Charts for Within Standard Deviations

For a measurement scheme involving k intercomparisons, it is possible to generate a control chart for what is called the "within" or short-term variability of the process.

Assume that each check standard value x_i $(i=1,\cdots,n)$ is the result of k intercomparisons; namely, $x_{i1},\cdots,x_{ik}$ where the quantity x_i is the average of these intercomparisons,

$$x_i = \frac{1}{k}\sum_{j=1}^{k} x_{ij} \quad .$$

The within standard deviations are estimated by

$$s_{w_i} = \left(\frac{1}{k-1}\sum_{j=1}^{k}(x_{ij} - x_i)^2\right)^{1/2} \qquad (5.4.1)$$

with degrees of freedom $\nu_i = k-1$. Where the intercomparisons form a statistical design, the quantity x_i and the within standard deviation are computed from a least-squares analysis.

The base line and limits for controlling short-term process variability make use of the same intercomparisons that were used to establish the control chart for averages. The base line is the pooled within standard deviation

$$s_p = \left(\frac{\nu_1 s_{w_1}^2 + \cdots + \nu_n s_{w_n}^2}{\nu_1 + \cdots + \nu_n}\right)^{1/2} \qquad (5.4.2)$$

The degrees of freedom $\nu = \nu_1 + \cdots + \nu_n$ allow for a different number of degrees of freedom in each estimate of the within standard deviation in (5.4.1). If all measurement schemes contain the same number of intercomparisons, say k, then $\nu = n(k-1)$.

Because a standard deviation is a positive quantity, it is only necessary to test against an upper limit in order to test the short-term variability. Thus for any future measurement sequence involving k intercomparisons, the within standard deviation s_w is computed as in (5.4.1) and is said to be in-control if

$$s_w < s_p \sqrt{F_\alpha(k-1,\nu)} \qquad (5.4.3)$$

where $F_\alpha(k-1,\nu)$ is the upper α percent point of the F distribution with k-1 degrees of freedom in the numerator and ν degrees of freedom in the denominator.

The control chart for averages used in conjunction with the control chart for within standard deviations is a powerful means of detecting changes in the process. The two control procedures are evoked simultaneously, and if an out-of-control condition is encountered for either test, the process is assumed to be out-of-control and the measurement sequence is repeated.

5.5 Alternative Control Limits

The reader may be familiar with control charts with control limits computed as the product of the total standard deviation and a fixed multiplicative factor, such as two or three, instead of the appropriate percentage point of the F or t-distribution. Control charts for within standard deviations should always be based on the F distribution because the critical values of the F distribution change rapidly with changes in degrees of freedom.

The consideration of whether a control chart for averages should be based on the percentage points of Student's t distribution or on a fixed multiplicative factor, such as three or two, is really a matter of choice depending on the level of control that one is hoping to achieve and on the type of measurements that are in question. The use of Student's t distribution is the most rigorous test if the measurements truly represent a random sample from a normal distribution. It allows a strict probability interpretation of the control procedure.

It cannot always be shown, and indeed is not always the case, that measurements come from an idealized distribution such as the normal distribution. If one looked at a large number of measurements on the same item, they might come from a distribution that is slightly skewed; i.e., for example, extreme large values may be more likely than extreme small values.

> The problem of deciding whether to use limits based on the normal distribution, those based on some other distribution, or those which involve no assumption about the form of the distribution is one which, though of a kind common in applied statistics, has no satisfactory solution. Limits based on the normal distribution are substantially shorter for a fixed sample size than those based on no assumption about the distribution, but they may be irrelevant if the distribution is too far from normal. (Bowker [67]).

For this reason it is customary in the United States to use plus or minus three standard deviations as the control limits (Duncan [68]). The factor three guarantees that a large proportion of the distribution is covered for measurements coming from any distribution that is close to the normal distribution in character. These limits are robust, and should be used when the intent is to identify measurements that are clearly out-of-control. Because these limits are so wide, an out-of-control finding is almost certainly an indication of a serious malfunction in the measurement process. If a somewhat tighter control is desired, two standard deviation limits can be considered. Very few values will fall between the two and three standard deviation limits, and the price of remeasuring for those few may be worth the added degree of control.

5.6 Control Charts for Drifting Check Standards

Another consideration concerns the problem of drifting check standards and whether or not they can be used for control purposes. The assumption is made in most measurement control programs that the check standard is stable and that any change that is noted by the control procedure is caused by changes in the measurement process itself. Obviously if the check standard is not completely stable, the ability to detect a change in the process is confounded with any possible drift in the check standard.

Unfortunately the situation in reality is that artifacts may not be completely stable, and this instability will be detected when it is large compared to the process precision. Changes in check standards over time can be expected. Of the forty or more check standards that are in continual use in the NBS mass calibration program, only about half of those standards are completely stable or do not show any drift over time. The question is, "Can a drifting check standard be used for control purposes?" Sometimes it can, but a drifting check standard causes complications in the analysis when, depending on the rate of change, the control limits pick up this change.

There are a few ad hoc procedures that can be used in lieu of a rigorous approach to this problem. Probably the simplest approach is to determine the time interval over which the check standard is stable by studying historical data and to enforce the control procedure over this interval. When this time interval has elapsed or when numerous values have been flagged as being out-of-control, the base line and control limits can be adjusted based on more recent measurements on the check standard.

It the check standard is changing steadily, as is the case for many artifact standards at NBS, it is sometimes possible to model the rate of drift and to predict from this model a value for the check standard at a future time that is not too far removed from the present. This involves fitting a regression equation to the measurements as a function of time by the method of least-squares and computing the values of the check standard for future times. Then the control procedure is time dependent; the base value is the predicted value from the regression equation at that time, and the control limits which depend on the standard deviation of this predicted value become wider with time. This approach has been used at NBS for check standards with linear drift rate as a function of time. It can work reasonably well as long as the drift remains linear, but the cause of a breakdown in the linearity assumption cannot be easily identified because it is never really possible to separate the change in the artifact from the change in the process. In such a situation it is imperative that the process be checked frequently for offset by comparison to a national standard or to other stable laboratory standards.

5.7 Synopsis and Examples of Control Charts

Four important ideas that are pertinent to calibration programs should emerge from the disucssion thus far. First, when dealing with statistical control of the properties of an artifact or statistical control of a measurement process, the control parameters are not imposed upon the process externally but are characteristic of the measurement process itself as described by historical data.

Secondly, if the check standard measurement is outside the established control limits, the calibration sequence is presumed to be out-of-control, and the calibrations of the test items are considered invalid. When such a condition is initially encountered, the instrumentation can be checked and the measurement sequence repeated--testing again for control. Any intervening results should be discarded. If control cannot be restored, a significant change has occurred in the process, and this change must be investigated. If a process is repeatedly out-of-control, the base line and control limits should be reestablished based on more recent data.

If the check standard measurement is in-control, this is taken as evidence that the process is behaving as expected in relation to the item submitted for calibration, and its assignment is assumed to be correct. Lack of control is certainly grounds for rejecting the calibration of the test item, but the complimentary argument is not as strong. The relationship between the measurement on the test item and the measurement on the check standard must be interrelated or executed very close together in time in order to be satisfied that the assignment ot the check standard has, indeed, been done properly.

Thirdly, the process precision is very well characterized by a total standard deviation calculated from measurements on the check standard. In some cases, such measurements provide the only way of obtaining a realistic estimate of this source of uncertainty. Fourthly, even though the tests for control can be automated, it is not only advantageous to visually examine the control charts in order to detect anomalies or slight shifts in the process and possible drifting of the check standard over time, but it is essential for understanding the long term behavior of the measurement process.

In order to demonstrate the value of such critical examinations, four examples that have been encountered in NBS measurement assurance programs are discussed.

The National Bureau of Standards maintains control charts on about forty check standards that are used in the mass calibration program. The control chart shown in figure 13 depicts values of the one kilogram check standard as it has been estimated from the measurement sequence used in the calibration workload for one kilogram weights. The three standard deviation limits shown by the dashed lines are the control limits that are used for this program, and if one compares these limits with the two standard deviation limits shown by the dotted lines, it is apparent that very few points fall between the two sets of limits. It can also be noted that the two standard deviation control limits are almost identical with control limits based on student's t distribution at significance level α = 0.01 when the number of points is large as in this case.

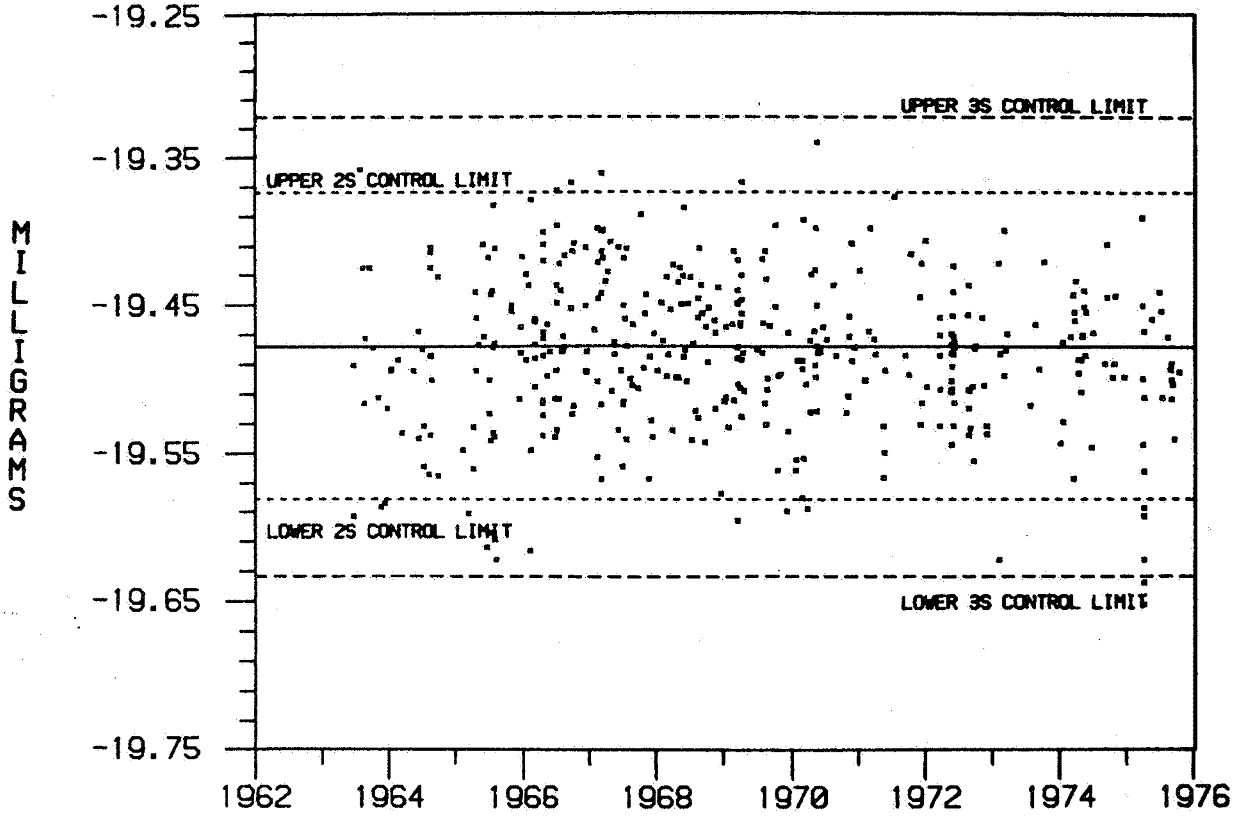


Figure 13
Check standard #41 (mg) as measured on NBS balance #4
plotted against time (years)

At this point the reader should be sufficiently sensitized to this approach to be aware of one shortcoming in this control chart. The chart implies that the process, which is demonstrably in-control, has never been out-of-control. A few points should fall outside of the control limits merely by chance, and as it happens other out-of-control situations have occurred in this program over the years. In fact, the control procedure would serve no useful purpose in the calibration program if there were no out-of-control situations to be detected. Actually this graph represents only the successful tests for control that were made with the one kilogram check standard because the calibration results and the check standard values were automatically discarded whenever the control limits were violated. The software for the NBS mass calibration program has been changed so that all values of the check standard are retained, and each value is flagged as to whether or not it was in control on that occasion. One should know when and how often control limits have been violated, and control charts should contain all findings.

The short-term or within variability of the same process is charted in figure 14 which shows within standard deviations for calibration sequences involving all weights calibrated on NBS balance #4. A calibration sequence typically requires between three and fifteen measurements, and the within standard deviation that is calculated from each sequence reflects the inherent variability of the balance and the effect of any environmental changes that occur during the time needed to make the requisite measurements. The base line for this control procedure, shown by the solid line, is the pooled within standard deviation in (5.4.2). Because the number of degrees of freedom varies with the design, it is not possible to establish a single upper control limit for this process; the control limit for each point is calculated separately, and the control procedure is automated using the control limit based on the F distribution as shown in (5.4.3). Once a year the within standard deviations are plotted to see if any degradation has occurred in the balance over the year.

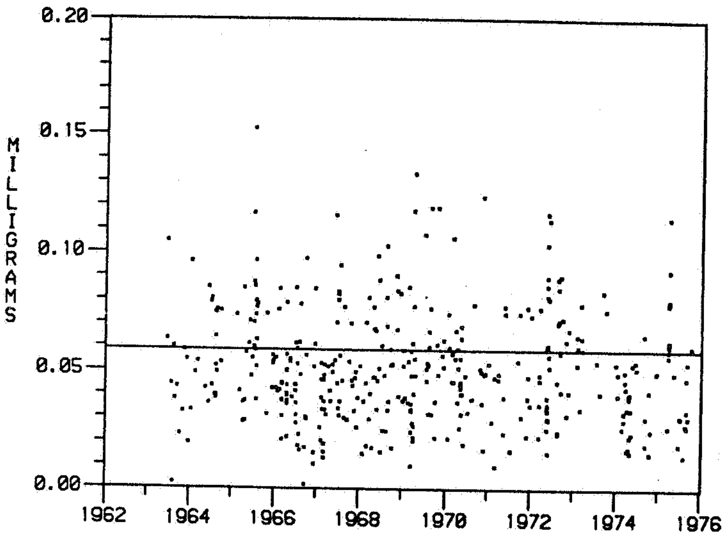


Figure 14
Within standard deviations (mg) for NBS balance #4
plotted against time (years)

The examples cited in figures 13 and 14 are for a process, as was said before, that has been in existence for a long time and that is demonstrably in-control. It may be instructive to examine a few processes, or at least the data from those processes, that have not been carefully monitored and that are not necessarily in-control.

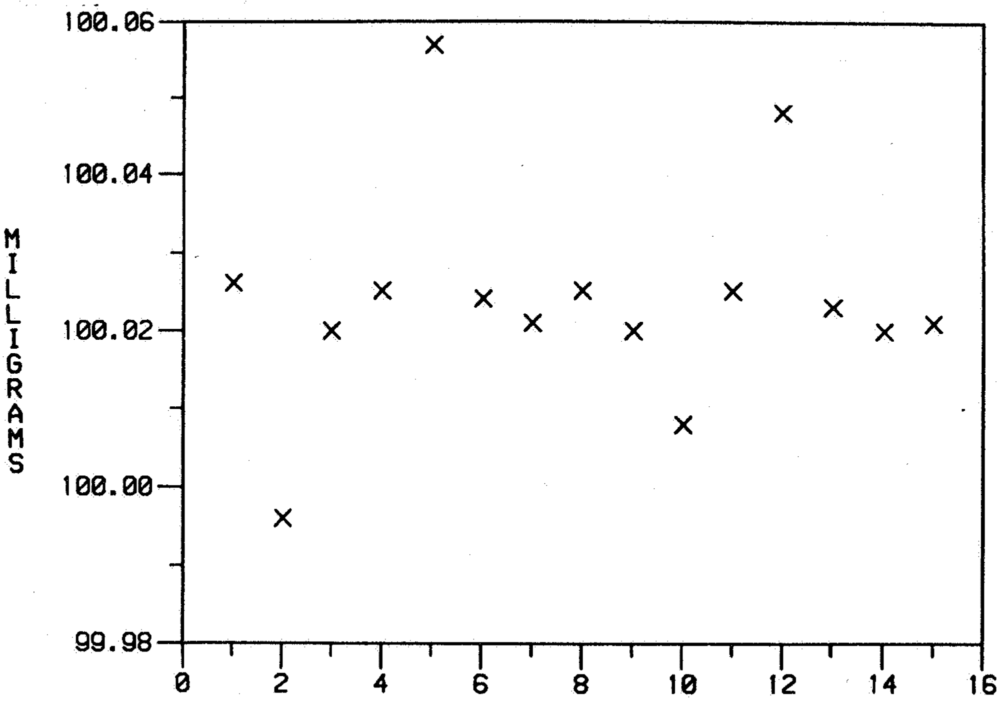


Figure 15
Measurements (mg) on a 100g weight plotted against time (months)

Take, for example, the data in figure 15 which represent repeated weighings made over a fifteen month period on a calibrated weight. Notice that the majority of the values are clustering close together but that there are a relatively large number of extremely discordant values. It is not sensible in this case to ask, "What base line and control limits are appropriate for this process?" In fact, at this point in time, a measurement process does not exist because it is not possible to predict a future value of the process, or in other words, the data as plotted in figure 15 do not represent a random sample from a single error distribution. In this case, a critical deficiency in the measurement process was tracked down; namely, that the elapsed time between two weighings being made on the balance in succession was not sufficient for the balance to come to proper equilibrium.

A control procedure involving a power instrument standard is shown in figure 16. The graph shows assignments made to the power standard as it was intercompared with its primary power source over a two-week period. The sixteen resulting measurements define the base line and control limits for the process.

The results of sixteen additional measurements taken a year later are shown in figure 17, and although they are clearly out-of-control with respect to the initial measurements, they are consistent among themselves raising a question as to whether the power standard itself is changing radically, whether the initial measurements were, in fact, out-of-control and should be discounted, or whether the process is not properly characterized by either set of measurements. Really only one thing is clear at this point -- that the assignment cannot be made with any degree of confidence and that the power standard should not be the basis for a calibration program until the process of assigning a value to the power standard is adequately characterized.

This was accomplished by repeating the intercomparison at three month intervals taking only two or three measurements each time instead of sixteen. The results are shown in figure 18. A large component of variance that did not show up in the initial two-week interval affects the measurement process, and the standard deviation computed from the short-term measurements under-estimates the process variability as it exists over, say, a year's time.

This example demonstrates an extremely important principle of measurement assurance; namely that in general there is little value in closely spaced repetitions. These should be kept to a minimum, and measurements should be taken frequently over a long period of time in order to correctly characterize a process. This practice should be continued until the process parameters are well established and only then should the intervals between intercomparisons be lengthened.

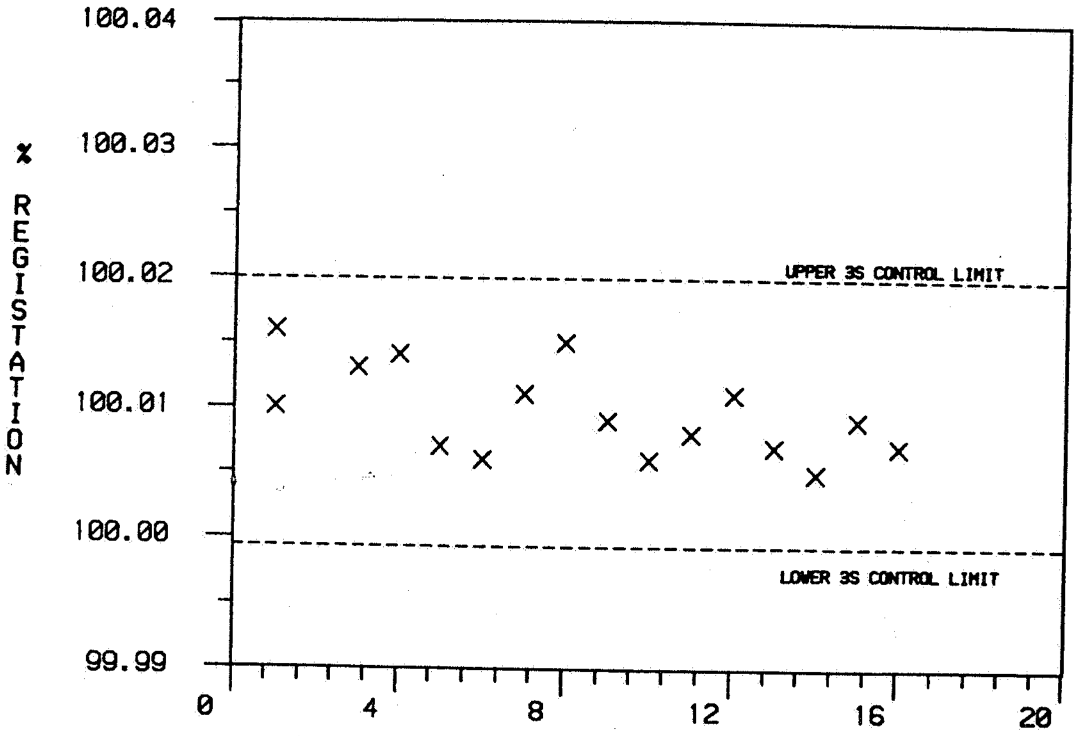


Figure 16
Measurements (% reg) on a power standard plotted against run sequence showing upper and lower three standard deviation limits

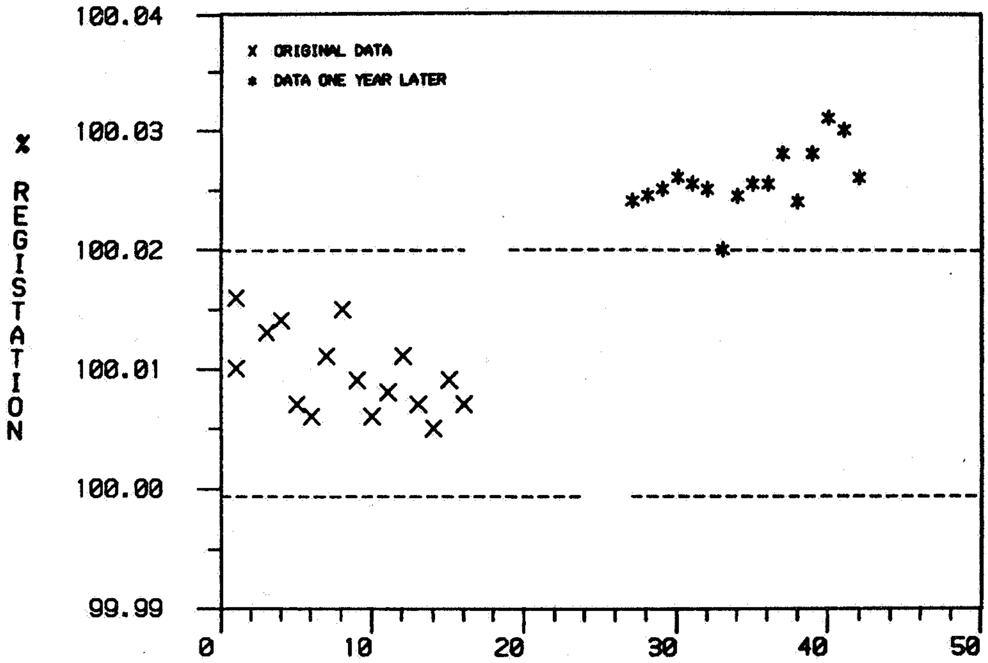


Figure 17
Original measurements (% reg) on power standard and measurements on the same standard a year later with original control limits

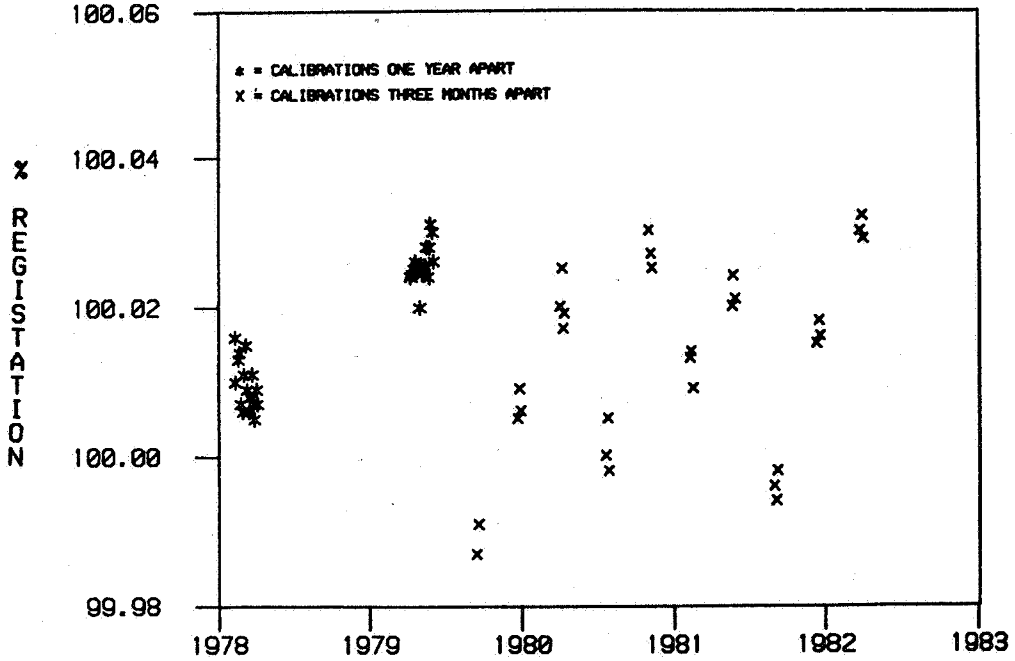


Figure 18
Measurements (% reg) on the power standard at three month intervals over three years

Table I

Critical Values $t_{\alpha/2}(\nu)$ of Student's t Distribution

ν	α=0.05	α=0.01	ν	α=0.05	α=0.01
2	4.303	9.925	62	1.999	2.657
4	2.776	4.604	64	1.998	2.655
6	2.447	3.707	66	1.997	2.652
8	2.306	3.355	68	1.995	2.650
10	2.228	3.169	70	1.994	2.648
12	2.179	3.055	72	1.993	2.646
14	2.145	2.977	74	1.993	2.644
16	2.120	2.921	76	1.992	2.642
18	2.101	2.878	78	1.991	2.640
20	2.086	2.845	80	1.990	2.639
22	2.074	2.819	82	1.989	2.637
24	2.064	2.797	84	1.989	2.636
26	2.056	2.779	86	1.988	2.634
28	2.048	2.763	88	1.987	2.633
30	2.042	2.750	90	1.987	2.632
32	2.037	2.738	92	1.986	2.630
34	2.032	2.728	94	1.985	2.629
36	2.028	2.719	96	1.984	2.628
38	2.024	2.712	98	1.983	2.627
40	2.021	2.704	100	1.983	2.626
42	2.018	2.698	102	1.983	2.625
44	2.015	2.692	104	1.982	2.624
46	2.013	2.687	106	1.982	2.623
48	2.011	2.682	108	1.981	2.622
50	2.009	2.678	110	1.981	2.621
52	2.007	2.674	112	1.981	2.620
54	2.005	2.670	114	1.981	2.620
56	2.003	2.667	116	1.981	2.619
58	2.002	2.663	118	1.980	2.618
60	2.000	2.660	120	1.980	2.617
			∞	1.960	2.576

ν = number of degrees of freedom in the total standard deviation.

Table II

Critical values $F_\alpha(\nu_1, \nu_2)$ of the F Distribution
α=0.01

DF	Degrees of freedom ν_1									
ν_2	1	2	3	4	5	6	7	8	9	10
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57
80	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55
85	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
95	6.91	4.84	3.99	3.52	3.22	3.00	2.83	2.70	2.60	2.51
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
105	6.88	4.81	3.97	3.50	3.20	2.98	2.81	2.69	2.58	2.49
110	6.87	4.80	3.96	3.49	3.19	2.97	2.81	2.68	2.57	2.49
115	6.86	4.79	3.96	3.49	3.18	2.96	2.80	2.67	2.57	2.48
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Table II continued

Critical Values $F_{\alpha}(\nu_1,\nu_2)$ of the F Distribution
$\alpha = 0.01$

DF	Degrees of freedom ν_1									
ν_2	12	14	16	18	20	22	24	26	28	30
10	4.71	4.60	4.52	4.46	4.41	4.36	4.33	4.30	4.27	4.25
11	4.40	4.29	4.21	4.15	4.10	4.06	4.02	3.99	3.96	3.94
12	4.16	4.05	3.97	3.91	3.86	3.82	3.78	3.75	3.72	3.70
13	3.96	3.86	3.78	3.72	3.66	3.62	3.59	3.56	3.53	3.51
14	3.80	3.70	3.62	3.56	3.51	3.46	3.43	3.40	3.37	3.35
15	3.67	3.56	3.49	3.42	3.37	3.33	3.29	3.26	3.24	3.21
16	3.55	3.45	3.37	3.31	3.26	3.22	3.18	3.15	3.12	3.10
17	3.46	3.35	3.27	3.21	3.16	3.12	3.08	3.05	3.03	3.00
18	3.37	3.27	3.19	3.13	3.08	3.03	3.00	2.97	2.94	2.92
19	3.30	3.19	3.12	3.05	3.00	2.96	2.92	2.89	2.87	2.84
20	3.23	3.13	3.05	2.99	2.94	2.90	2.86	2.83	2.80	2.78
22	3.12	3.02	2.94	2.88	2.83	2.78	2.75	2.72	2.69	2.67
24	3.03	2.93	2.85	2.79	2.74	2.70	2.66	2.63	2.60	2.58
26	2.96	2.86	2.78	2.72	2.66	2.62	2.58	2.55	2.53	2.50
28	2.90	2.79	2.72	2.65	2.60	2.56	2.52	2.49	2.46	2.44
30	2.84	2.74	2.66	2.60	2.55	2.51	2.47	2.44	2.41	2.39
35	2.74	2.64	2.56	2.50	2.44	2.40	2.36	2.33	2.30	2.28
40	2.66	2.56	2.48	2.42	2.37	2.33	2.29	2.26	2.23	2.20
45	2.61	2.51	2.43	2.36	2.31	2.27	2.23	2.20	2.17	2.14
50	2.56	2.46	2.38	2.32	2.27	2.22	2.18	2.15	2.12	2.10
55	2.53	2.42	2.34	2.28	2.23	2.18	2.15	2.11	2.08	2.06
60	2.50	2.39	2.31	2.25	2.20	2.15	2.12	2.08	2.05	2.03
65	2.47	2.37	2.29	2.23	2.17	2.13	2.09	2.06	2.03	2.00
70	2.45	2.35	2.27	2.20	2.15	2.11	2.07	2.03	2.01	1.98
75	2.43	2.33	2.25	2.18	2.13	2.09	2.05	2.02	1.99	1.96
80	2.42	2.31	2.23	2.17	2.12	2.07	2.03	2.00	1.97	1.94
85	2.40	2.30	2.22	2.15	2.10	2.06	2.02	1.98	1.95	1.93
90	2.39	2.29	2.21	2.14	2.09	2.04	2.00	1.97	1.94	1.92
95	2.38	2.28	2.20	2.13	2.08	2.03	1.99	1.96	1.93	1.90
100	2.37	2.27	2.19	2.12	2.07	2.02	1.98	1.95	1.92	1.89
105	2.36	2.26	2.18	2.11	2.06	2.01	1.97	1.94	1.91	1.88
110	2.35	2.25	2.17	2.10	2.05	2.00	1.96	1.93	1.90	1.88
115	2.34	2.24	2.16	2.10	2.04	2.00	1.96	1.92	1.89	1.87
120	2.34	2.23	2.15	2.09	2.03	1.99	1.95	1.92	1.89	1.86
∞	2.19	2.09	2.00	1.94	1.88	1.84	1.79	1.76	1.73	1.70

Table II continued

Critical Values $F_\alpha(\nu_1,\nu_2)$ of the F Distribution
$\alpha=0.01$

DF	Degrees of freedom ν_1									
ν_2	40	50	60	70	80	90	100	110	120	∞
10	4.17	4.12	4.08	4.06	4.04	4.03	4.01	4.00	4.00	3.91
11	3.86	3.81	3.78	3.75	3.73	3.72	3.71	3.70	3.69	3.60
12	3.62	3.57	3.54	3.51	3.49	3.48	3.47	3.46	3.45	3.36
13	3.43	3.38	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.17
14	3.27	3.22	3.18	3.16	3.14	3.12	3.11	3.10	3.09	3.01
15	3.13	3.08	3.05	3.02	3.00	2.99	2.98	2.97	2.96	2.87
16	3.02	2.97	2.93	2.91	2.89	2.87	2.86	2.85	2.84	2.76
17	2.92	2.87	2.83	2.81	2.79	2.78	2.76	2.75	2.75	2.65
18	2.84	2.78	2.75	2.72	2.70	2.69	2.68	2.67	2.66	2.57
19	2.76	2.71	2.67	2.65	2.63	2.61	2.60	2.59	2.58	2.49
20	2.69	2.64	2.61	2.58	2.56	2.55	2.54	2.53	2.52	2.42
22	2.58	2.53	2.50	2.47	2.45	2.43	2.42	2.41	2.40	2.31
24	2.49	2.44	2.40	2.38	2.36	2.34	2.33	2.32	2.31	2.21
26	2.42	2.36	2.33	2.30	2.28	2.26	2.25	2.24	2.23	2.13
28	2.35	2.30	2.26	2.24	2.22	2.20	2.19	2.18	2.17	2.07
30	2.30	2.24	2.21	2.18	2.16	2.14	2.13	2.12	2.11	2.01
35	2.19	2.14	2.10	2.07	2.05	2.03	2.02	2.01	2.00	1.89
40	2.11	2.06	2.02	1.99	1.97	1.95	1.94	1.93	1.92	1.81
45	2.05	2.00	1.96	1.93	1.91	1.89	1.88	1.86	1.85	1.74
50	2.01	1.95	1.91	1.88	1.86	1.84	1.82	1.81	1.80	1.69
55	1.97	1.91	1.87	1.84	1.82	1.80	1.78	1.77	1.76	1.64
60	1.94	1.88	1.84	1.81	1.78	1.76	1.75	1.74	1.73	1.60
65	1.91	1.85	1.81	1.78	1.75	1.74	1.72	1.71	1.70	1.57
70	1.89	1.83	1.78	1.75	1.73	1.71	1.70	1.68	1.67	1.54
75	1.87	1.81	1.76	1.73	1.71	1.69	1.67	1.66	1.65	1.52
80	1.85	1.79	1.75	1.71	1.69	1.67	1.65	1.64	1.63	1.50
85	1.83	1.77	1.73	1.70	1.67	1.65	1.64	1.62	1.61	1.48
90	1.82	1.76	1.72	1.68	1.66	1.64	1.62	1.61	1.60	1.46
95	1.81	1.75	1.70	1.67	1.65	1.63	1.61	1.60	1.58	1.45
100	1.80	1.74	1.69	1.66	1.63	1.61	1.60	1.58	1.57	1.43
105	1.79	1.73	1.68	1.65	1.62	1.60	1.59	1.57	1.56	1.42
110	1.78	1.72	1.67	1.64	1.61	1.59	1.58	1.56	1.55	1.41
115	1.77	1.71	1.66	1.63	1.60	1.58	1.57	1.55	1.54	1.40
120	1.76	1.70	1.66	1.62	1.60	1.58	1.56	1.54	1.53	1.39
∞	1.60	1.53	1.48	1.44	1.41	1.38	1.36	1.35	1.33	

REFERENCES

[1] Dorsey, N. E. The velocity of light. Trans. Am. Phil. Soc. XXXIV.; 1944, pp.1-110.

[2] Hayford, J. A. On the least square adjustment of weighings, U.S. Coast and Geodetic Survey Appendix 10, Report for 1892; 1893.

[3] Benoit, M. J. R. L'Etalonnage des Series de Poids Travaux et Memoirs du Bureau International des Poids et Mesures 13(1); 1907.

[4] Pienkowsky, A. T. Short tests for sets of laboratory weights. Scientific Papers of the Bureau of Standards, 21(527); 1926.

[5] Cameron, J. M.; Croarkin, M. C.; Raybold, R. C. Designs for the Calibration of Standards of Mass. Nat. Bur. Stand. (U.S.) Tech Note 952; 1977.

[6] §Eisenhart, C. Realistic evaluation of the precision and accuracy of instrument calibration systems. J. Res. Nat. Bur. Stand. (U.S.) 67C(2); 1962. pp. 161-187.

[7] §Youden, W. J. Experimental design and ASTM committees. Materials Research and Standards, 1(11); 1961. pp. 862-867.

[8] §Youden, W. J. Physical measurements and experiment design. Colloques Internationaux du Centre National de la Recherche Scientifique No. 110, le Plan d'Experiences; 1961. pp. 115-128.

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

[10] Croarkin, M. C.; Beers, J. S.; Tucker, C. Measurement Assurance for Gage Blocks. Nat. Bur. Stand. (U.S.) Monogr. 163; 1979.

[11] Croarkin, M. C.; Varner, R. N. Measurement Assurance for Dimensional Measurements on Integrated Circuit Photomasks. Nat. Bur. Stand. (U.S.) Tech. Note 1164; 1982.

[12] American National Standard ANSI N15.18-1975. Mass calibration techniques for nuclear materials control. Available from ANSI, Inc., 1430 Broadway, New York, NY 10018.

[13] Pontius, P. E. and Doher, L. W. The joint ANSI-INMM 8.1-Nuclear Regulatory Commission study of uranium hexafluoride cylinder material accountability bulk measurements. Proc. 18th Ann. Mtg. INMM, VI(III); 1977. p. 480.

§Reprinted in Nat. Bur. Stand. (U.S.) Spec. Publ. 300, Vol I. Precision Methods and Calibration: Statistical Concepts and Procedures. H. H. Ku, editor. 1969.

[14] Beers, J. S. A Gage Block Measurement Process Using Single Wavelength Interferometry. Nat. Bur. Stand. (U.S.) Monogr. 152; 1975.

[15] Cameron, J. M. Measurement Assurance. Nat. Bur. Stand. (U.S.) NBSIR 77-1240; 1977.

[16] Kieffer, L. J., ed. Calibration and Related Measurement Services of the National Bureau of Standards. Nat. Bur. Stand. (U.S.) Spec. Publ. 250; 1982.

[17] Pipkin, F. R., Ritter, R. C. Precision measurements and fundamental constants. Science, 219 (4587); 1983. p. 917.

[18] Pontius, P. E. The Measurement Assurance Program - A Case Study: Length Measurements Part I. Long Gage Blocks (5 in to 20 in). Nat. Bur. Stand. (U.S.) Monogr. 149; 1975.

[19] Beers, J. S., Tucker, C. D. Intercomparison Procedures for Gage Blocks Using Electromechanical Comparators. Nat. Bur. Stand. (U.S.) NBSIR 76-979; 1976. p. 9.

[20] Simpson, J. A. Foundations of metrology. J. Res. Nat. Bur. Stand. (U.S.). 86(3), 1981. p. 282.

[21] Nyyssonen, D. Linewidth measurement with an optical microscope: The effect of operating conditions on the image profile. Appl. Opt. 16(8); Ausust 1977. pp. 2223-2230.

[22] Jerke, J. M., ed. Semiconductor Measurement Technology: Accurate Linewidth Measurements on Integrated Circuit Photomasks. Nat. Bur. Stand. (U.S.) Spec. Publ. 400-43; 1980. pp. 7-15.

[23] Jerke, J. M.; Croarkin, M. C.; Varner, R. N. Semiconductor Measurement Technology: Interlaboratory Study on Linewidth Measurements for Antireflective Chromium Photomasks. Nat. Bur. Stand. (U.S.) Spec. Publ. 400-74; 1982.

[24] See reference 20, p. 283.

[25] See reference 22, p. 50.

[26] See reference 20, p. 283.

[27] Cameron, J. M. Encyclopedia of Statistical Sciences, Vol 1. S. Kotz and N. L. Johnson, ed. New York: John Wiley & Sons, Inc.; 1982. pp. 341-347.

[28] Ku, H. H. Statistical concepts of a measurement process. Precision Methods and Calibration: Statistical Concepts and Procedures. Nat. Bur. Stand. (U.S.) Spec. Publ. 300, Vol I. H. H. Ku, ed. 1969. pp 296-20 to 330-54.

[29] Mandel, J. The Statistical Analysis of Experimental Data. New York: Interscience Publ; 1964. pp. 278-279.

[30] Snedecor, G. W. and Cochran, W. G. Statistical Methods, Sixth ed. Ames, Iowa: The Iowa State University Press; 1976. pp. 279-280.

[31] See reference [7]. pp. 862-863.

[32] Mattingly, G. H.; Pontius, P. E.; Allion, H. H.; Moore, E. F. A laboratory study of turbine meter uncertainty. Proc. Symp. on Flow in Open Channels and Closed Circuits; Nat. Bur. Stand. (U.S.) Spec. Publ. 484; 1977.

[33] See reference [10].

[34] Duncan, A.J. Quality Control and Industrial Statistics, Fourth ed. Homewood: Richard D. Irwin, Inc; 1974. p. 381.

[35] See reference [8], p. 21-22.

[36] See reference [28], p. 299.

[37] See reference [28], p. 305-308.

[38] See reference [23].

[39] Hockersmith, T. E.; Ku, H. H. Uncertainties associated with proving ring calibration. Precision Methods and Calibration: Statistical Concepts and Procedures. Nat. Bur. Stand. (U.S.) Spec. Publ. 300, Vol. 1. H. H. Ku, ed.; 1969. pp. 257-1 to 264-8.

[40] See reference [11], p. 30-33.

[41] Youden, W. J. Uncertainties in calibration. IRE Transactions on Instrumentation, I-11, (3,4); 1962. p. 137.

[42] Eisenhart, C., Ku, H. H. and Colle, R. Expression of the Uncertainties of Final Measurement Results; Reprints. Nat. Bur. Stand. (U.S.) Spec. Pub. 644; 1983.

[43] Giacomo, P. News from BIPM. Metrologia, 17; 1981. pp. 73-74.

[44] See reference [28]. pp. 322-323.

[45] Raghavarao, D. Construction and Combinatorial Problems in Design of Experiments. New York: John Wiley & Sons, Inc; 1971. p. 315.

[46] Mood, A. M. On Hotelling's Weighing Problem. Annals of Mathematical Statistics, 17; 1946. pp.432-446.

[47] See reference [5].

[48] Cameron, J. M.; 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; 1974. p. 1.

[49] Jaegar, K. B. and Davis, R. S. A Primer for Mass Metrology. Nat. Bur. Stand. (U.S.) Special Publication: Industrial Measurement Series 700-1; 1984.

[50] See reference [10], pp. 13-25.

[51] See reference [10], pp. 27-39.

[52] Cameron, J. M.; Eicke, W. G. Designs for Surveillance of the Volt Maintained by a Small Group of Saturated Standard Cells. Nat. Bur. Stand. (U.S.) Tech Note 430; 1967.

[53] Cameron, J. M. The Use of the Method of Least Squares in Calibration. Nat. Bur. Stand. (U.S.) NBSIR 74-587; 1974.

[54] See reference [49].

[55] Varner, R. N. Mass Calibration Computer Software. Nat. Bur. Stand. (U.S.) Tech. Note 1127; 1980.

[56] See reference [52]. pp. 1-2.

[57] Cochran, W. J. The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Eugenics, 11; 1941. pp. 47-52.

[58] Eisenhart, C. Significance of the largest of a set of sample estimates of variance. Chapter 15 of Selected Techniques of Statistical Analysis. Eisenhart, C., Hastay, M. W., Wallis, W. A., editors. New York: McGraw Hill Book Co., Inc.; 1947. pp. 383-394.

[59] Draper, N. R.; Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, Inc.; 1966. p. 1-32.

[60] See reference [22], p. 350.

[61] Bicking, C. A.; Gryna, F. M. Jr. Process Control by Statistical Methods. Section 23 of Quality Control Handbook, Third Edition, J. M. Juran, ed. New York: McGraw-Hill Book Co.; 1974. p. 23-2, 23-3.

[62] See reference [34]. pp. 464-484.

[63] Reynolds, J. R. Jr.; Ghosh, B. K. Designing control charts for means and variances. 1081 ASQC Quality Congress Transactions: San Francisco; 1981.

[64] Shewhart, W. A. Statistical Method from the Viewpoint of Quality Control. The Graduate School, U.S. Department of Agriculture, Washington, DC; 1939.

[65] Grant, E. L.; Leavenworth, R. S. Statistical Quality Control, 4th Edition. New York: McGraw Hill Book Co.; 1976. p. 129.

[66] Parobeck, P.; Tomb, T.; Ku, H. H.; Cameron, J. M. Measurement assurance program for weighings of respirable coal mine dust samples. J. Quality Tech, 13(3); 1981. pp. 157-165.

[67] Bowker, A. H. Tolerance limits for normal distributions. Chapter 2 of reference [58]. p. 99.

[68] See reference [63]. p. 381.

APPENDIX A

The purpose of this appendix is to define the matrix manipulations[‡] that produce the least-squares solution to a weighing design along with the propagation of associated standard deviations and uncertainties.[†] The theory is explained by Cameron et al. in reference [5]. It is assumed that a series of weighing designs is required in order to calibrate an entire weight set and that assignments to individual weights depend upon a starting restraint with known value that is invoked in the first design. The starting restraint is usually the known sum of two reference kilograms. It is also assumed that the designs are interconnected in such a way that a value assigned to an individual weight or sum of weights from one design constitutes the restraint for the next design in the series.

Each design in the series involves n intercomparisons among p weights where the p weights include the reference standards composing the restraint, the test weights, and check standard.

The model for the measurement process assumes that these observations are related to the values of the weights by

$$D = AX^* + \varepsilon \qquad (A.1)$$

where D is the (nxl) vector of observations; A is an (nxp) design matrix of zeroes and ones such that a plus or minus one in the ijth position indicates that the jth weight is measured by the ith observation, and a zero indicates the converse; X^* is the (pxl) vector of unknown values for the p weights; and ε is the (nxl) vector of random errors.

Define

$$D' = (d_1 \cdot \cdot \cdot d_n) \qquad (A.2)$$

$$A = \begin{pmatrix} a_{11} & \cdot \cdot \cdot & a_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{n1} & \cdot \cdot \cdot & a_{np} \end{pmatrix} \qquad (A.3)$$

$$(X^*)' = (X_1{}^* \cdot \cdot \cdot X_p{}^*) \qquad (A.4)$$

and

$$\varepsilon' = (\varepsilon_1 \cdot \cdot \cdot \varepsilon_n) \qquad (A.5)$$

[‡] The matrix notation that is used in this appendix denotes the transpose of the matrix M by M' and the inverse of the matrix M by M^{-1}.

[†] Assuming that there is no significant between component of variance in the measurement process.

In order to define various linear combinations of the weights, we will also define several vectors of size (px1) which have the general form

$$\ell' = (\ell_1 \cdot \cdot \cdot \ell_p)$$

where each element ℓ_i $(i=1,\ldots,p)$ is either zero, plus one or minus one.

The least-squares estimate for (A.4) depends upon the inverse of the normal equations A'A. The usual case for calibration experiments is that A'A has rank p-1. Where A'A has rank less than p, the inverse does not exist and a solution can be obtained only by imposing a restraint upon the system of equations. Therefore, we let R^* be a scalar with known value called the restraint; and ℓ_R be a (px1) vector of zeroes and ones such that a one in the jth position indicates that the jth weight is in the restraint, and a zero indicates the converse. For example,

$$\ell_R' = (1 \; 1 \; 0 \cdot \cdot \cdot 0)$$

indicates that the restraint is over the first two weights.

One approach to finding the least-squares estimate for X^* is via an augmented matrix B where

$$B = \begin{pmatrix} A'A & \ell_R & A'D \\ \ell_R' & 0 & R^* \\ 0 & 0 & -1 \end{pmatrix} \qquad \text{(A.6)}$$

is a (p+2)x(p+2) matrix whose inverse

$$B^{-1} = \begin{pmatrix} Q & h & \hat{X}^* \\ h' & 0 & \cdot \\ \cdot\;\cdot\;\cdot & & \cdot \end{pmatrix} \qquad \text{(A.7)}$$

can be partitioned as shown above. The (pxp) matrix Q in the upper left hand corner of B^{-1} contains information relating to the variances of the estimates, and the (px1) matrix X^* in the upper right hand corner of B^{-1} contains the least-squares estimates for the p weights. The other quantities in B^{-1} are not of interest for this application. Notice that once the inverse of B has been computed, the estimates are immediately available without further matrix multiplications.[¶]

The individual deviations of the observations from their fitted values are given by the (px1) vector ξ where

$$\xi' = (D - AX^*)', \qquad \text{(A.8)}$$

[¶] The caret (^) indicating a least-squares estimate from the data is dropped in future references to X^*.

and the within standard deviation for the design is

$$s_W = \left(\frac{\xi'\xi}{n-p+1} \right)^{1/2} \qquad (A.9)$$

with n-p+1 degrees of freedom.

The restraint for the next design in the series can be written in the form

$$\Sigma^* = \ell_\Sigma ' X^* \qquad (A.10)$$

where ℓ_Σ is a (px1) vector of zeroes and ones where a one in the ith position indicates that the ith weight is to be included in the restraint for the next design, and a zero indicates the converse. The standard deviation for the outgoing restraint is given by

$$s_\Sigma = \left(\ell_\Sigma ' Q \ell_\Sigma \, s_W{}^2 + \left(\frac{\ell_\Sigma ' W}{\ell_R ' W} \right)^2 s_R{}^2 \right)^{1/2} \qquad (A.11)$$

where s_R is the standard deviation of the incoming restraint R^* as computed from the previous design, and

$$W' = (W_1 \; . \; . \; . \; W_p)$$

where W is a (px1) vector of nominal values for the p weights. If the current design is the first design in the series, then s_R is zero.

Notice that the computation of the standard deviation associated with the check standard as defined in (A.14) and the computation of the standard deviation associated with the values of the test weights as defined in (A.16) are also dependent on s_R. Thus, the standard deviations for each series are dependent on all prior series as they are propagated starting with the first series.

The current value for the check standard from the design can be written in the form

$$c = \ell_c ' X^* \qquad (A.13)$$

where ℓ_c is a (px1) vector of zeroes and ones such that a plus or minus one in the ith position indicates that the ith weight is in the check standard, and a zero indicates the converse.

The standard deviation of the check standard value is given by

$$s_c = \left(\ell_c ' Q \ell_c \, s_W{}^2 + \left(\frac{\ell_c ' W}{\ell_R ' W} \right)^2 s_R{}^2 \right)^{1/2} . \qquad (A.14)$$

Then given that the accepted value for the check standard is known from previous experiments to be A_c, a test for control is made by computing the test statistic

$$t_c = \frac{|A_c - c|}{s_c} \tag{A.15}$$

and comparing it to a critical value.

Finally, we are interested in the uncertainty of the value assigned to a single weight or to a collection of weights. For each summation or difference of weights that is of interest, we define a (px1) vector ℓ_S of zeroes, plus ones and minus ones such that a one in the ith position indicates that the ith weight is involved in the summation or difference, and a zero indicates the converse. The reported value for the summation S is S^* where

$$S^* = \ell_S{}'X^*.$$

The standard deviation for the summation, designated by s_S is

$$s_S = \left(\ell_S{}'Q\ell_S \, s_w{}^2 + \left(\frac{\ell_S{}'W}{\ell_R{}'W} \right)^2 s_R{}^2 \right)^{1/2} \tag{A.16}$$

and the uncertainty associated with the summation is

$$U = 3s_S + \frac{\ell_S{}'W}{\ell_{SR}{}'W} U_{SR} \tag{A.17}$$

where U_{SR} is the uncertainty assigned to the starting restraint in the series, and similarly ℓ_{SR} is the (px1) vector of zeroes, plus ones and minus ones such that a plus or minus one in the ith position indicates that the ith weight is in the starting restraint.

Notice that if we are talking about a single weight whose value is $X_j{}^*$, then the quantity

$$\ell_S{}'Q\ell_S = q_{jj}$$

where q_{jj} is the jth diagonal element in Q.

For the next design in the series, let the restraint be $R^* = \Sigma^*$ with standard deviation $s_R = s_\Sigma$ and proceed with the calculation starting with equation (A.1).