Expression of the Uncertainties of Final Results

Clear statements of the uncertainties of reported values are needed for their critical evaluation.

Churchill Eisenhart

Measurement of some property of a thing in practice always takes the form of a sequence of steps or operations that yield as an end result a number that serves to represent the amount or quantity of some particular property of a thing—a number that indicates how much of this property the thing has, for someone to use for a specific purpose. The end result may be the outcome of a single reading of an instrument, with or without corrections for departures from prescribed conditions. More often it is some kind of average, for example, the arithmetic mean of a number of independent determinations of the same magnitude, or the final result of a least squares "reduction" of measurements of a number of different magnitudes that bear known relations with one another in accordance with a definite experimental plan. In general, the purpose for which the answer is needed determines the precision or accuracy required and ordinarily also the method of measurement employed.

Although the accuracy required of a reported value depends primarily on the *intended* use, or uses, of the value, one should not ignore the requirements of other uses to which it is likely to be put. A reported value whose accuracy is entirely unknown is worthless.

Strictly speaking, the actual *error* of a reported value, that is the magnitude and sign of its deviation from the truth (*1*), is usually unknowable. Limits to this error, however, can usually be inferred—with some risk of being incorrect—from the precision of the measurement process by which the reported value was obtained, and from reasonable limits to the possible bias of the measurement process. The *bias*, or *systematic error*, of a measurement process is the magnitude and direction of its tendency to measure something other than what was intended; its *precision* refers to the typical closeness together of successive independent measurements of a single magnitude generated by repeated applications of the process under specified conditions; and its *accuracy* is determined by the closeness to the true value characteristic of such measurements.

Precision and accuracy are inherent characteristics of the measurement process employed and not of the particular end result obtained. From experience with a particular measurement process and knowledge of its sensitivity to uncontrolled factors, one can often place reasonable bounds on its likely systematic error (bias). It is also necessary to know how well the particular value in hand is likely to agree with other values that the same measurement process might have provided in this instance, or might yield on remeasurement of the same magnitude on another occasion. Such information is provided by the estimated *standard error* (*2*) of the reported value, which measures (or is an index of) the characteristic disagreement of repeated determinations of the same quantity by the same method, and thus serves to indicate the precision (strictly, the imprecision) of the reported value (*3*).

Four Distinct Forms of Expression Needed

The uncertainty of a reported value is indicated by stating credible limits to its likely inaccuracy. No single form of expression for these limits is universally satisfactory. In fact, different forms of expression are recommended, which will depend on the relative magnitudes of the imprecision and likely bias, and their relative importance in relation to the intended use of the reported value, as well as to other possible uses to which it may be put (*4*).

Four distinct cases need to be recognized: (i) both systematic error and imprecision negligible, in relation to the requirements of the intended and likely uses of the result; (ii) systematic error not negligible, imprecision negligible; (iii) neither systematic error nor imprecision negligible; and (iv) systematic error negligible, imprecision not negligible.

Specific recommendations with respect to each of these cases are made below. General guidelines upon which these specific recommendations are based are discussed in the following paragraphs.

Perils of Shorthand Expressions

Final results and their respective uncertainties should be reported in sentence form whenever possible. The shorthand form "$a \pm b$" should be avoided in abstracts and summaries; and never used without explicit explanation of its connotation. If no explanation is given, many persons will take $\pm b$ to signify bounds to the inaccuracy of a. Others may assume that b is the "standard error," or the "probable error," of a, and hence the uncertainty of a is at least $\pm 3b$, or $\pm 4b$, respectively. Still others may take b to be an indication merely of the imprecision of the individual measurements, that is, to be the "standard deviation," or the "average deviation," or the "probable error" of a single observation. Each of these interpretations reflects a practice of which instances can be found in current scientific literature. As a step in the direction of reducing this current confusion, it is recommended that the use of "$a \pm b$" in presenting results be limited to that sanctioned for the case of tabular results in the fourth recommendation of the section below headed "Systematic error not negligible, imprecision negligible."

The author is a senior research fellow and former chief of the Statistical Engineering Laboratory at the National Bureau of Standards, Washington, D.C. 20234. The recommendations presented in this paper have evolved at the Bureau over a period of many years and are made public here for general information, and to educe comments and suggestions.

Imprecision and Systematic Error Require Separate Treatment

Since imprecision and systematic error are distinctly different components of inaccuracy, and are subject to different treatments and interpretations in usage, two numerics respectively expressing the imprecision and bounds to the systematic error of the reported result should be used whenever both of these errors are factors requiring consideration. Such instances are discussed in the section below for the case of "Neither systematic error nor imprecision negligible."

In quoting a reported value and its associated uncertainty from the literature, the interpretation of the uncertainty quoted should be stated if given by the author. If the interpretation is not known, a remark to this effect is in order. This practice may induce authors to use more explicit formulations of their statements of uncertainty.

Standard Deviation and Standard Error

The terms *standard deviation* and *standard error* should be reserved to denote the canonical values for the measurement process, based on considerable recent experience with the measurement process or processes involved. When there is insufficient recent experience, an estimate of the standard error (standard deviation) must of necessity be computed by recognized statistical procedures from the same measurements as the reported value itself. To avoid possible misunderstanding, in such cases, the term "computed (or estimated) standard error" ("computed standard deviation") should be used. A formula for calculating this computed standard error is given in the section below for the case of "Neither systematic error nor imprecision negligible."

Uncertainties of Accepted Values of Fundamental Constants or Primary Standards

If the uncertainty in the accepted value of a national primary standard or of some fundamental constant of nature (for example, in the volt as maintained at the National Bureau of Standards, or in the acceleration of gravity *g* on the Potsdam basis) is an important source of systematic error affecting the measurement process, no allowance for possible systematic error from this source should be included ordinarily in evaluating overall bounds to the systematic error of the measurement process. Since the error concerned, whatever it is, affects all results obtained by the method of measurement involved, to include an allowance for this error would be to make everybody's results appear unduly inaccurate relative to each other. In such instances one should state: (i) that measurements obtained by the process concerned are expressed in terms of the volt (or the kilogram, or other unit) "as maintained at the National Bureau of Standards," or (ii) that the indicated bounds to the systematic error of the process are exclusive of the uncertainty of the stated value adopted for some particular constant or quantity. An example of the latter form of statement is:

. . . neglecting the uncertainty of the value 6.6256×10^{-34} joule seconds adopted for Planck's constant.

Systematic Error and Imprecision Both Negligible

In this case the reported result should be given, after rounding, to the number of significant figures consistent with the accuracy requirements of the situation, together with an explicit statement of its accuracy. An example is:

. . . the wavelengths of the principal visible lines of mercury-198 have been measured relative to the 6057.802106 Å (angstrom units) line of krypton-98, and their values in vacuum are

5792.2685 Å
5771.1984 Å
5462.2706 Å
4359.5625 Å
4047.7146 Å

correct to eight significant figures.

It needs to be emphasized that if no statement of accuracy or precision accompanies a reported number, then, in accordance with the usual conventions governing rounding, this number will ordinarily be interpreted as being accurate within $\pm\frac{1}{2}$ unit in the last significant figure given; that is, it will be understood that its inaccuracy before rounding was less than $\pm$ 5 units in the next place. The statement "correct to eight significant figures" is included explicitly in the foregoing example, rather than left to be understood in order to forestall any concern that an explicit statement of lesser accuracy was inadvertently omitted.

Systematic Error Not Negligible, Imprecision Negligible

When the imprecision of a result is negligible, but the inherent systematic error of the measurement process concerned is not negligible, then the following rules are recommended:

1) Qualification of a reported result should be limited to a single quasi-absolute type of statement that places bounds on its inaccuracy.

2) These bounds should be stated to no more than two significant figures.

3) The reported result itself should be given (that is, rounded) to the last place affected by the stated bounds (unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that it may possess for certain particular uses).

4) Accuracy statements should be given in sentence form in all cases, except when a number of results of different accuracies are presented, for example, in tabular arrangement. If it is necessary or desirable to indicate the respective accuracies of a number of results, the results should be given in the form $a \pm b$ (or $a\ {}^{+b}_{-c}$, if necessary) with an appropriate explanatory remark (as a footnote to the table, or incorporated in the accompanying text) to the effect that the $\pm b$, or ${}^{+b}_{-c}$, signify bounds to the systematic errors to which the *a*'s may be subject.

5) The fact that the imprecision is negligible should be stated explicitly.

The particular form of the quasi-absolute type of statement employed in a given instance will depend ordinarily on personal taste, experience, current and past practice in the field of activity concerned, and so forth. Some examples of good practice are:

. . . is (are) not in error by more than 1 part in (*x*).
. . . is (are) accurate within $\pm$ (*x* units) [or $\pm$ (*x*) percent].
. . . is (are) believed accurate within (. . . . .).

Positive wording, as in the first two of these quasi-absolute statements, is appropriate only when the stated bounds to the possible inaccuracy of the reported value are themselves reliably established. However, when the indicated bounds are somewhat conjectural, it is desirable to signify this fact (and put the reader on guard) by inclusion of some modifying expression such as "believed," "considered," "estimated to be," "thought to be," and so forth, as exemplified by the third of the foregoing examples.

The term *uncertainty* may sometimes be used effectively to achieve a conciseness of expression otherwise difficult or impossible to attain. Thus, one might make a statement such as:

> The uncertainties in the above values are not more than ± 0.5°C in the range 0°C to 1100°C, and then increase to ± 2°C at 1450°C,

or

> The uncertainty in this value does not exceed . . . excluding (or, including) the uncertainty of . . . in the value . . . adopted for the (reference standard involved).

A statement giving numerical limits of uncertainty as in the above should be followed by a brief discussion telling how the limits were derived.

Finally, the following forms of quasi-absolute statements are considered poor practice, and are to be avoided:

> The accuracy of . . . is 5 percent.
> The accuracy of . . . is ± 2 percent.

These are presumably intended to mean that the result concerned is not inaccurate, that is, not in error, by more than 5 percent or 2 percent, respectively, but they explicitly state the opposite.

Neither Systematic Error Nor Imprecision Negligible

When neither the imprecision nor the systematic error of a result are negligible, then the following rules are recommended:

1) A reported result should be qualified by a quasi-absolute type of statement that places bounds on its systematic error, and a separate statement of its standard error or its probable error, or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise a computed value of the standard error, or, probable error, so designated, should be given together with a statement of the number of degrees of freedom on which it is based.

2) The bounds to its systematic error and the measure of its imprecision should be stated to no more than two significant figures.

3) The reported result itself should be stated at most to the last place affected by the finer of the two qualifying statements (unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that it may possess for certain particular uses).

4) The qualification of a reported result with respect to its imprecision and systematic error should be given in sentence form, except when results of different precision or with different bounds to their systematic errors are presented in tabular arrangement. If it is necessary or desirable to indicate their respective imprecisions or bounds to their respective systematic errors, such information may be given in a parallel column or columns, with appropriate identification.

Here, and in the next section, the term *standard error* is to be understood as signifying the standard deviation of the reported value itself, not as signifying the standard deviation of the single determination (unless, of course, the reported value is simply the result of a single determination).

The above recommendations should not be construed to exclude the presentation of a quasi-abolute type of statement placing bounds on the inaccuracy, that is, on the overall uncertainty, of a reported value, provided that separate statements of its imprecision and its possible systematic error are included also. To be in good taste, the bounds indicating the overall uncertainty should not be numerically less than the corresponding bounds placed on the systematic error outwardly increased by at least three times the standard error. The fourth of the following examples of good practice is an instance at point:

> The standard errors of these values do not exceed 0.000004 inch, and their systematic errors are not in excess of 0.00002 inch.

> The standard errors of these values are less than (*x* units), and their systematic errors are thought to be less than ± (*y* units). No additional uncertainty is assigned for the conversion to the chemical scale since the adopted conversion factor is taken as 1.000275 exactly.

> . . . with a standard error of (*x* units), and a systematic error of not more than ± (*y* units).

> . . . with an overall uncertainty of ± 3 percent based on a standard error of 0.5 percent and an allowance of ± 1.5 percent for systematic error.

When a reliably established value for the relevant standard error is available, and the dispersion of the present measurements is in keeping with this experience, then this canonical value of the standard error should be used (5). If such experience indicates that the standard error is subject to fluctuations greater than the intrinsic variation of such a measure, then an appropriate upper bound should be given, for example, as in the first two of the above examples, or by changing "a standard error . . ." in the third and fourth examples to "an upper bound to the standard error . . ."

When there is insufficient recent experience with the measurement processes involved, an estimate of the standard error must of necessity be computed by recognized statistical procedures from the same measurements as the reported value itself. It is essential that such computations be carried out according to an agreed-upon standard procedure, and the results thereof presented in sufficient detail to enable the reader to form his own judgment, and make his own allowances for their inherent uncertainties. To avoid possible misunderstanding, in such cases, first, the term *computed standard error* should be used; second, the estimate of the standard error employed should be that obtained from

$$\text{estimate of standard error} = \left(\frac{\text{sum of squared residuals}}{n\nu}\right)^{1/2}$$

where n is the (effective) number of completely independent determinations of which a is the arithmetic mean (or other appropriate least-squares adjusted value) and ν is the number of degrees of freedom involved in the sum of squared residuals (that is, the number of residuals minus the number of fitted constants or other independent constraints on the residuals); and third, the number of degrees of freedom should be explicitly stated. If the reported value a is the arithmetic mean, then:

$$\text{estimate of standard error} = (s^2/n)^{1/2}$$

where

$$s^2 = \sum_{i=1}^{n} (x_i - a)^2/(n-1)$$

and n is the number of completely independent determinations of which a is the arithmetic mean. For example:

> . . . which is the arithmetic mean of (*n*) independent determinations and has a standard error of . . .

> . . . with an overall uncertainty of ± 5.2 km/sec based on a standard error of 1.5 km/sec and estimated bounds of ± 0.7 km/sec on the systematic error. (The figure 5.2 is equal to 0.7 plus 3 times 1.5.)

or, if based on a computed standard error,

> The computed probable error (or, standard error) of these values is (*x* units),

based on (ν) degrees of freedom, and the systematic error is estimated to be less than $\pm$ (y units).

. . . with an overall uncertainty of $\pm$ 7 km/sec derived from bounds of $\pm$ 0.7 km/sec on the systematic error and a computed standard error of 1.5 km/sec based on 9 degrees of freedom. [The number 7 is approximately equal to $0.7 + (4.3 \times 1.5)$, where 4.3 is the value of Student's t for 9 degrees of freedom exceeded in absolute value with 0.002 probability. As $\nu \rightarrow \infty$, $t_{.002}(\nu) \rightarrow 3.090$.]

When the reported value is the result of a complex measurement process and is obtained as a function of several quantities whose standard errors have been computed, these several quantities and their standard errors should usually be reported, together with a description of the method of computation by which the standard errors were combined to provide an overall estimate of imprecision for the reported value.

Systematic Error Negligible, Imprecision Not Negligible

When the systematic error of a result is negligible but its imprecision is not, the following rules are recommended:

1) Qualification of a reported value should be limited to a statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise a computed value of the standard error, so designated, should be given together with a statement of the number of degrees of freedom on which it is based.

2) The standard error or upper bound thereto, should be stated to not more than two significant figures.

3) The reported result itself should be stated at most to the last place affected by the stated value or bound to its imprecision (unless it is desired to indicate and preserve such relative precision of a higher order that it may possess for certain particular uses).

4) The qualification of a reported result with respect to its imprecision should be given in sentence form, except when results of different precision are presented in tabular arrangement and it is necessary or desirable to indicate their respective imprecisions in which event such information may be given in a parallel column or columns, with appropriate identification.

5) The fact that the systematic error is negligible should be stated explicitly.

The above recommendations should not be construed to exclude the presentation of a quasi-absolute type of statement placing bounds on its possible inaccuracy, provided that a separate statement of its imprecision is included also. To be in good taste, such bounds to its inaccuracy should be numerically equal to at least three times the stated standard error. The fourth of the following examples of good practice is an instance at point.

The standard errors of these values are less than (x units).

. . . with a standard error of (x units).

. . . with a computed standard error of (x units) based on (ν) degrees of freedom.

. . . with an overall uncertainty of $\pm$ 4.5 km/sec derived from a standard error of 1.5 km/sec. (The figure 4.5 is equal to 3×1.5.)

or, if based on a computed standard error,

. . . with an overall uncertainty of $\pm$ 6.5 km/sec derived from a computed standard error of 1.5 km/sec (based on 9 degrees of freedom). (The number 6.5 is equal to 4.3×1.5, where 4.3 is the value of Student's t for 9 degrees of freedom exceeded in absolute value with 0.002 probability. As $\nu \rightarrow \infty$, $t_{.002}(\nu) \rightarrow 3.090$.)

The remarks with regard to a computed standard error in the preceding section apply with equal force to the last two examples above.

Conclusion

The foregoing recommendations call for fuller and sharper detail than is general in common pactice. They should be regarded as minimum standards of good practice. Of course, many instances require fuller treatment than that recommended here.

Thus, in the case of determinations of the "fundamental physical constants" and other basic properties of nature, the author or authors should give a detailed account of the various components of imprecision and systematic error, and list their respective individual magnitudes in tabular form, so that (i) the state of the art will be more clearly revealed, (ii) each individual user of the final result may decide for himself which of the indicated components of imprecision or systematic error are, or are not, relevant to his use of the final result, and (iii)—most important—the final result itself or its uncertainty can be modified appropriately in the light of later advances. This is, and has long been, the practice followed in the best reports of fundamental studies, but current efforts to prepare critically evaluated standard reference data have revealed that far too great a fraction of the data in the scientific literature "cannot be critically evaluated because the minimum of essential information is not present" (*6*).

References and Notes

1. The true value defined conceptually by an exemplar measurement process, or the target value intended in a practical measurement process.
2. The standard error is the standard deviation of the probability distribution of estimates (that is, reported values) of the quantity that is being measured. See M. G. Kendall and W. R. Buckland, *A Dictionary of Statistical Terms* (Hafner, New York, 1957).
3. For a comprehensive discussion on precision and accuracy, and a selected bibliography of 80 references, see C. Eisenhart, "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems," *J. Res. Nat. Bur. Std.* **67C**, No. 2, 161–187 (1963). (Reprints are available upon request.)
4. The essential elements of the present recommendations first appeared in a 1955 National Bureau of Standards task group report prepared principally by Malcolm W. Jensen (Office of Weights and Measures), Leroy W. Tilton (Optics and Metrology Division), and Churchill Eisenhart (Applied Mathematics Division), which was based for the most part on detailed recommendations developed some years earlier by Dr. Tilton for the internal guidance of the Optics and Metrology Division. In September 1961, new introductory material was added to the recommendations of the 1955 task group; a few minor changes were made in the illustrative examples, and the resulting revised version was circulated as a working paper of the Subcommittee on Accuracy Statements of the NBS Testing and Calibration Committee. This 1961 version was incorporated without essential change as chapter 23, "Expression of the Uncertainties of Final Results," of NBS Handbook 91, *Experimental Statistics* (U.S. Government Printing Office, Washington, 1963), reprinted with corrections in 1966. (This handbook brought together in a single volume the material on experimental statistics prepared at the National Bureau of Standards for the U.S. Army *Ordnance Engineering Design Handbook*, and printed in 1962 for limited distribution as U.S. Army Ordnance Corps Pamphlets ORDP 20-110 through 20-114. Subsequently, when these five pamphlets became parts of the *AMC Engineering Design Handbook*, they were designated Army Materiel Command Pamphlets AMCP 706-110 through 706-114.)
   In the present version, the content of chapter 23 has been rearranged and, in order to be more appropriate to calibration work, more explicit consideration has been given to the case where the value of the standard deviation σ of the measurement process involved has been well established by recent past experience. A terse summary of the principal recommendations of the present paper in the form of a text figure (Fig. 1) is contained in H. H. Ku, "Expressions of Imprecision, Systematic Error, and Uncertainty Associated with a Reported Value," to be published in *Measurements and Data*. The earlier versions were addressed primarily to the case of isolated experiments or tests, where the relevant value of σ is usually unknown in advance, and the statistical uncertainty of the final results must therefore be expressed entirely in terms of quantities derived from the data of the experiment itself.
5. The control chart is an invaluable tool in providing justification for the use of a canonical value of the standard error. See, for example, ASTM Manual on *Quality Control of Materials* (American Society for Testing and Materials, Philadephia, 1951).
6. L. M. Branscomb, "The misinformation explosion: Is the literature worth reviewing?," a talk presented to the Philosophical Society of Washington, 17 November 1967, and to be published in *Scientific Research*.