Rapport BIPM-80/3

Report on the BIPM enquiry on error statements

March 1980

BUREAU INTERNATIONAL DES POIDS ET MESURES

F-92310 SEVRES

The BIPM enquiry on error statements

1. Introduction

At the suggestion of Dr. E. Ambler (NBS), a member of the Comité International des Poids et Mesures (CIPM), the Bureau International des Poids et Mesures (BIPM) prepared a "Questionnaire sur les incertitudes" which was sent to a number of laboratories and institutions which were thought to be particularly interested in these problems. They were asked to give replies which, to some extent, could be considered as reflecting the opinions of their respective countries. The English version of the distributed document is reproduced here as Appendix I.

We recall that the main aim of this enquiry was to enable the BIPM to arrive at a selection of those specific problems which, at a later stage, could be useful subjects for discussion by an appropriate Working Party. For details we refer to the introduction of Appendix I.

In view of the broad range of views held on these matters, such a preliminary sorting out of the most urgent problems was considered necessary, and the answers received have confirmed this opinion. Although, in fact, the laboratories were not asked to take a definite position with respect to a given question, but only to say whether, in their mind, this problem should be the subject of further discussions or not, it did not come as a surprise that many replies also contained a number of arguments or statements which must reflect, at least to some extent, the prevailing attitude a laboratory has at the present time towards some of the main problems in this field.

It is the purpose of the present document to summarize the relevant information on error statements as far as it can be extracted from the answers received. In doing so we should not overlook the fact that for the more difficult questions there is often a large percentage of answers which do not express a clear preference. One often gets the impression that this "silent majority", rather than having already made up its mind, is primarily looking for more clarification and guidance. It can be hoped, therefore, that the dissemination of the various opinions received, although they are at times difficult to reconcile, may give useful additional information. A broader and deeper understanding of the various aspects can possibly contribute to finding new ways of tackling and eventually solving current problems. The elaboration of guidelines which are useful and acceptable to a majority of users will not be a simple task, but a serious attempt to reach this goal is certainly worthwhile.

Table 1 lists the countries from which answers have been received (till the end of 1978); for the sake of convenience they will be referred to in what follows by their respective numbers (in parentheses).

Table 1 - Countries from which replies were received, in alphabetical order

(1) Africa, South
National Physical Research Laboratory, Pretoria

(2) America, United States of
National Bureau of Standards, Washington

(3) Australia
National Measurement Laboratory, Lindfield

(4) Brazil
assembled by Luiz Cintra do Prado, São Paulo (member of CIPM)

(5) Canada
National Research Council, Ottawa

(6) China
National Institute of Metrology, Peking

(7) Czechoslovakia
Czechoslovak Institute of Metrology, Bratislava

(8) Denmark
Technical University, Dept. of Physics, Lyngby

(9) France
Bureau National de Métrologie, Paris

(10) Germany, Democratic Republic of
Amt für Standardisierung, Messwesen und Warenprüfung, Berlin

(11) Germany, Federal Republic of
Physikalisch-Technische Bundesanstalt, Braunschweig

(12) Great Britain
National Physical Laboratory, Teddington

(13) Hungary
Office National des Mesures, Budapest

(14) Italy
Istituto di Metrologia "G. Colonnetti", Torino

(15) Japan
National Research Laboratory of Metrology, Takyo

(16) Korea
Korea Standards Research Institute, Seoul

(17) Netherlands
Van Swinden Laboratorium, 's-Gravenhage

(18) Poland
summary of answers given by four institutions

(19) Rumania
Institutul National de Metrologie, Bucuresti

(20) Sweden
Statens Provningsanstalt, Borås

(21) USSR
probably from Gosstandard, Moscow (answer received via OIML, Paris).

2. Replies to the questions

The questions to be answered will just be repeated. For the supplementary information on the problems which was included in the questionnaire, see Appendix I.

For some laboratories, the answers of which are particularly informative, the complete documents received are reproduced as Appendices. This concerns

(2)	NBS, Washington	see Appendix II
(9)	BNM, Paris	" " III
(11)	PTB, Braunschweig	" " IV
(12)	NPL, Teddington	" " V.

It should be mentioned that a simple classification of the answers, especially of those given with a substantial number of comments, into the simple "yes" and "no" categories was not always very easy to perform. By checking with the full answers given as appendices, the reader may get an idea of the amount of arbitrary judgment which may have been applied. In any case, no serious distortion of the results as a whole should come from this.

Let us now look separately at the different questions.

a) <u>Question 1</u>: "Should one recommend the use of the standard deviation to characterize the random uncertainty?"

This introductory question was a rather simple one and the answers can be grouped as follows.

- "Yes": (1) (5) (6) (7) (18) (19) (21).

- "Yes", but asking that the number of degrees of freedom be added: (2) (4) (8) (9) (16).

- Several laboratories hesitate to indicate a clear preference or leave the choice open: (10) (11) (12) (13) (14) (17) (20).

Finally (15) requests that "it should be clearly stated which of repeatability, replicatability or reproducibility the standard deviation is representing"; seven different kinds of random uncertainties are distinguished. (3) says "Yes for reporting in scientific journals, no for calibration reports and certificates".

For additional comments see (2) and (12).

b) <u>Question 2</u>: "Should one recommend the use of a conventional probability and the corresponding confidence limits (instead of the standard deviation)?"

This question is obviously linked with the previous one; our comments below will therefore concern both.

Among the answers no unrestricted "Yes" can be found. On the other hand, the following laboratories give a clear "No": (1) (2) (5) (8) (9) (16) (18) (21).

A majority of replies is rather unclear or votes for "no decision". Here we find (3) (4) (6) (7) (10) (11) (12) (13) (14) (15) (17) (19) (20). However, the reasons (when indicated) for the hesitations are quite varied. In several answers it was said that the two suggested solutions were not real alternatives, but had both got their useful fields of application. (15) recommends "the measuring process should be randomized so that random variation of data should conform to the normal distribution". Finally, (7) suggests using "les limites de l'étendue vraie des erreurs" since this quantity "est finie en comparaison avec l'étendue infinie d'une distribution normale". The pertinent remark of (9) may be noted.

<u>Comments of BIPM on questions 1 and 2</u> : From the bulk of the replies a clear preference for the standard deviation seems to emerge, provided that it be complemented with the degrees of freedom (or the number of measurements performed). Confidence limits have found few adherents and several replies recall rightly that such an interval does not contain more information than the estimated standard deviation, but in addition supposes a normal population. Its use should therefore be restricted to cases where a statistical decision is required.

c) Question 3a: "Is there an essential difference between random and systematic errors?"

This innocent-looking question has given rise to a large range of replies, as expected. A simple "No" has been given by (1) (3) (14) (19). The "Yes" replies are more numerous, but they differ quite a lot among themselves in the cases where explanations are added. "Yes" answers were given by (5) (8) (13) (17) (18). While for (21) a systematic error is equivalent to a correction, (16) makes the observation that a "random error at one level could be considered as a systematic error at the next level in the calibration hierarchy".

However, the largest group consists of those which give no clear (or simple enough) reply, to which we may add the hesitants. Such answers came from (2) (4) (6) (7) (9) (10) (11) (12) (15) (20). Among them (6) and (15) note explicitly that "they can change type" and (9) adds that "il y a aussi tous les cas intermédiaires". (12) suggests a new subdivision according to the method of derivation (cf. Appendix III), while (2) says that "the essential difference ... becomes apparent when the uncertainty ... is put into actual use" (cf. Appendix II). Finally, (20) finds that "there exists a more or less broad zone where it cannot be stated with confidence that a measurement uncertainty has to be regarded as being systematic or random".

d) Question 3b: "Should one recommend a practical rule which enables one to know with which type of error one is dealing?"

Again, this question is related to the previous one, as reflected by the answers. "No" as answer was given by the same labs as above.
"Yes" came from (7) (12) (13) (17) (18) (21).
No clear-cut position was expressed by (2) (4) (5) (6) (8) (9) (10) (11) (15) (16) (20). Note in particular the comments given by (2) and (9).

Comments of BIPM on questions 3a and 3b : To judge from the variety of replies this was a good question. However, since the opinions are still widely discordant it would be a dangerous subject for a discussion which is supposed to lead in a finite time to a practical conclusion. More thinking, involving the fundamental as well as the practical aspects, seems to be needed before there is real hope for a satisfactory solution to this difficult problem. This process is probably already under way at different places, but it will take some time since the problem is for many experimentalists quite recent. In addition, it may be at variance with their current practices.

e) Question 4 : "Should one recommend a practical rule for the expression of systematic errors?"

Since the explicit form of a possible "practical rule" was not asked for, it was perfectly legitimate to answer by a simple yes or no.

A "yes" without further comment was given by (3) (11) (13) (21), while (7) adds that there would be different recommendations, depending on the type of measurement, the way corrections for systematic effects are obtained and the aim of the measurement. We can also understand (20) which says "it seems extremely difficult to arrive at a rule which has the prospect of being applied and interpreted in a uniform manner".

Only one lab (8) votes for "no", commenting that "no rule is better than one which is ambiguous, ill-defined or even misleading".

In some answers the "maximum limit" concept is recommended, namely by (10) (17) and also (14), although with caution. Finally (18) favours the form of a standard deviation.

No clear-cut answer to the question as asked is given by (4) (5) (6) (9) (19). A grouping into two different categories (experimental data or personal judgment) is suggested by (2) and (16), whereas (12) would prefer to describe the uncertainty by means of a subjective phrase (cf. Appendix V).

Comments of BIPM : Here we are clearly facing one of the basic problems. While most laboratories would welcome having a "practical rule" available, there is little convergence in the proposals made so far. Confidence limits (for various probability levels) and maximum limits as well as standard deviations have found adherents. The problem of choosing a given form is obviously linked with questions 3a and 3b and has important practical implications. Clearly, any suggested solution would also have to be judged with respect to its possible usefulness for further data handling, and some of the problems implied are alluded to in questions 5 to 7.

f) Question 5: "Should one recommend a practical rule for combining systematic errors (with other systematic errors)?"

Since there can be more than one systematic error, it is obvious that some rule for combining them is needed indeed. Therefore nobody has given a simple "no" as answer. However, most laboratories have clearly realized that any answer is conditioned by what has been said in reply to question 4 and that a simple and general solution is far from obvious.

"Yes" as answer, usually without further comment, came from (3) (4) (10) (11) (13) (17) (19) (21). Positive, but more sceptical or restrictive replies were given by (5) (8) (9) (14) (20). In particular, (2) discusses various possible combinations (linear, quadratic), but without recommending a given rule (cf. Appendix II). Simple addition of the variances is suggested by (6). Most answers, however, are difficult to group. Thus (16) says that "a few practical rules, instead of one, can be suggested", which is somewhat less pessimistic than (7) which writes that "il est possible d'obtenir quelques dizaines de cas différents". While all this may be true, at least to some extent, it is certainly not the type of rule most people are looking for.

(12) lists five methods in common use and adds "What is required is a convincing theory ... from which a method for combining systematic uncertainties can be derived". Possibly a useful hint is given by (18) in stating that "il ne faudrait pas distinguer la règle de l'addition des incertitudes systématiques ou aléatoires et systématiques, mais appliquer la loi de propagation de l'écart-type connue en statistique mathématique. Alors on pourrait exprimer toutes les incertitudes par l'écart-type de la moyenne; il ne faudrait chercher aucune autre terminologie pour les incertitudes de mesure".

Comments of BIPM : This is obviously an important subject for discussion and a single general rule would be most welcome. The adherents of a linear addition seem to be diminishing in number and in particular nobody has suggested using one for maximum limits. This looks, after all, quite encouraging, although the goal is not yet at hand.

h) Question 6: "Should one recommend a practical rule for combining random and systematic errors?"

Here the various opinions differ greatly. Whereas (8) is very sceptical and fears "that such a rule will often be more misleading than elucidating", (13) and (17) are of the opinion that this would depend on the "level of metrology". Clear opposition also exists. Thus (20) says that they "should not be combined", and (16) notes quite similarly "we should avoid the combination of the two types of error in general".

Adherents of a practical rule, but sometimes with restrictions, are (4) (9) (10) (11) (21), while (19) thinks that "it is preferable to have a reduced number of practical rules".

Quadratic addition is advocated by (3) and (6), possibly also by (18). No clear-cut answer (or none at all) was received from (1) (2) (5) (7) (12) (14).

For further exposition of the problems see the appendices.

Comments of BIPM: The previously very popular belief that any combination of types must be avoided seems to be changing. The new and more realistic attitude is no doubt inspired by the practical need. While there now probably exists a majority which is willing to accept some rule for a combination of random and systematic uncertainties, no concensus on how this should be accomplished in detail is yet clearly visible.

i) Question 7: "Should one recommend a practical rule for expressing the final uncertainty?"

Again the close relation of this question with the previous one is obvious. Scepticism or opposition is expressed by (9) and (20) which finds it "most difficult ... to formulate one single rule". Similarly (21) thinks that "il faut recommander plusieurs règles ... en fonction des problèmes". While the laboratories (2) (3) (4) (7) (10) (11) (14) would find such a rule desirable, four others mention specifically that according to them such a rule should depend on the "level of metrology", namely (12) (13) (16) (17). No definite position has been taken by (1) (5) (6) (18).

Comments of BIPM: The usefulness of a "final uncertainty" will no doubt depend to a large extent on the intended application. Perhaps a subdivision of the problem should be taken seriously into account. Nevertheless, the elements or components needed to construct such a quantity would be necessarily those discussed previously. This limits the possibilities. However, the question in its general form might well be beyond the field for which the BIPM can or should try to suggest a solution. In particular, the responsibility for calibration certificates must remain fully with the issuing laboratory which, in turn, is bound in many cases by legal prescriptions. It seems difficult, therefore, to expect more than some general hints for "good and reasonable practice".

j) Question 8: "Do other questions appear essential to you? Does it seem to you that the preceding questions should be considered in a different way?"

Here the majority of laboratories (namely 8) express satisfaction or give no reply (5 in all), which can also be interpreted as an agreement.

On the other hand, three laboratories (7) (17) (18) insist strongly on the presumed importance of a well-defined terminology, which may also depend on the "level of metrology". An extensive "glossary of terms connected with measuring uncertainties" is presented in (11) (cf. Appendix IV for details), and in the same reply it is suggested that "one should discuss what information should be given in test certificates in different fields of work". However, it may be left open whether discussion on such detailed questions is possible or desirable in a working party dealing with the general problems connected with uncertainties. Some work will inevitably have to be left to more specialized bodies.

Some further suggestions concern the following points:

- (20) thinks it would be desirable to have a discussion on "the possible use of the range as an alternative measure of the scatter of individual readings, especially if their number is small", as well as on some other concepts, such as "repeatability, reproducibility or inaccuracy".

- (14) asks the question "faut-il recommander une règle pratique pour établir si deux ou plusieurs mesures fournies avec une certitude globale donnée sont non discordantes entre elles?"

- (21) tries to remind us that "il est très important de définir et de tenir compte des lois de distribution des erreurs composantes dans les calculs des erreurs de mesure".

- (17) insists that "much more publicity should be given to the results of the international comparisons held under the auspices of the BIPM".

Finally, a somewhat confusing remark was found in (8) which says "especially the Bureau has succeeded in hiding its own point of view - which it undoubtedly must possess". For lack of better information we take this as a compliment.

Comments of BIPM : The problems associated with terminology had not escaped our attention when the questionnaire was prepared. They are real and of practical importance, but are not necessarily at the heart of the problems involved. As long as the basic concepts are still disputed, questions of vocabulary are better deferred. Besides, experience shows that - since everybody is an expert on language - discussions on questions of terminology are usually lengthy and inefficient. The additional problems stemming from linguistic barriers together with the known attraction such discussions exert on purists seem to recommend avoidance of these matters as far as possible in a first round. Real specialists can tackle them later.

As for the better defined statistical questions, most of them can be answered in a straightforward manner.

Bibliography

In reply to our request for information on publications related to the statement of uncertainties we have received a number of documents which are listed below, in chronological order, together with some coming from our own files.

Considering the enormous quantity of documents and papers which have been published on questions related to error statements as well as their often quite modest scientific interest, completeness will not be attempted. In most documents further bibliographic references can be found.

The few documents which discuss general problems in a way similar to ours and which may therefore be of special interest to the reader are marked with an asterisk.

a) Regulations or guidelines ("recipies")

*- P.J. Campion, J.E. Burns, A. Williams: "A code of practice for the detailed statement of accuracy" (HMSO, London, 1973), 52 p.

- "Methods for treating the results of measurements", Proc. Institutes of Metrology USSR, No. 172 (232), (Energia, Leningrad, 1975), 72 p.

- "Glosario de terminos empleados en metrologia" (Comite de metrologia de la AECC, Madrid, 1976), 35 p.

- "The expression of uncertainty in electrical measurements" (No. 3003 British Calibration Service, 1977), 16 p.

- "The Australian Standards for the Measurement of Physical Quantities", National Measurement Laboratory (CSIRO, Melbourne, 1977), 22 p.

- O. Mathiesen: "How should the result of a measurement be reported?", Svensk Mätplatskalender 1977, 14 p. (in Swedish)

- "Expresión de los resultados de une calibración o medida (comite de metrologia de la AECC, Madrid, 1978), 16 p.

- "Metrology. Expression of result of measurement", Svensk standard SS 01 41 50 (draft, 1978), 12 p.

- "Measurement of fluid flow - Estimation of uncertainty of a flow-rate measurement", ISO 5168-1978 (E), 26 p.

- "Assessment of uncertainty in calibration and use of flow measurement devices", 79/31646 (British Standards Institution, London), 78 p.

- "Grundbegriffe der Messtechnik; Begriffe für Unsicherheit beim Messen und Fehler, Korrektion, Fehlergrenzen bei Messgeräten", DIN 1319, part 3 (draft), (Deutsches Institut für Normung, Berlin, ca. 1979), 17 p.

- "Guidelines for estimation and statement of overall uncertainty in measurement results", CSC (80) MS-9, National Physical Laboratory of India, Standards and Industrial Research Institute of Malaysia (Commonwealth Science Council, London, 1980), 13 p.

b) Special studies on error statements

- C. Eisenhart: "Realistic evaluation of the precision and accuracy of instrument calibration systems", J. Res. NBS 67C, 161-187 (1963)

- H.O.W. Diettrich: "Ueber die Konsistenz von Messergebnissen", Report BTB VI C/N-4 (PTB, Braunschweig, 1965), 19 p.

- H.H. Ku: "Notes on the use of propagation of error formulas", J. Res. NBS 70C, 263-273 (1966)

- C. Eisenhart: "Expression of the uncertainties of final results", Science 160, 1201-1204 (1968)

- K. Weichselberger: "Ueber die Addition zufälliger und systematischer Fehler", Operations Research Verfahren 5, 423-444 (1968)

- C.F. Dietrich: "Uncertainty, Calibration and Probability" (Hilger, London, 1973), 411 p.

- F.E. Grubbs: "Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments", Technometrics 15, 53-66 (1973)

- H.H. Ku: "Statistical methods applicable to counting experiments and evaluation of experimental data", Nucl. Instr. and Meth. 112, 377-383 (1973)

- A. Williams: "Statement of results of experiments and their accuracy", Nucl. Instr. and Meth. 112, 373-376 (1973)

- L. Gonella: "Proposal for a revision of the measure theory and terminology", Alta Frequenza 44, 622-628 (1975)

- G.C. Martarelli, A. Zanini: "Expression in unitary form of combinations of measurement error sources by means of normal distributions", Alta Frequenza 44, 629-633 (1975)

- W. Masin: "Ueber die Informationsfähigkeit eines Messgerätes", Metrologia 11, 169-177 (1975)

- H. Reich: "Die Unsicherheit von Messungen mit Ionisationskammer-Dosimetern", PTB-Mitteilungen 86, 421-426 (1976)

- G. Becker: "Ueber die Unsicherheit von primären Cäsiumstrahlnormalen für Zeit und Frequenz", PTB-Mitteilungen 87, 131-137 (1977)

- P.J. Campion: "Problems in the statement of uncertainties" (NPL, Teddington, 1977), 4 p.

- F.G. Perey: "Covariance matrices of experimental data", Proc. Int. Conf. on Neutron Physics and Nuclear Data, Harwell 1978, p. 104-115

- L.M. Schwartz: "Statistical uncertainties of analyses by calibration of counting measurements", Analytical Chemistry 50, 980-985 (1978)

- W. Wöger: "Remarks on the confidence limits resulting from two models for the randomization of systematic uncertainties" (PTB Braunschweig, 1978; draft), 16 p.

*- J.W. Müller: "Some second thoughts on error statements", Nucl. Instr. and Meth. 163, 241-251 (1979)

*- S.R. Wagner: "On the quantitative characterization of the uncertainty of experimental results in metrology", PTB-Mitteilungen 89, 83-89 (1979)

- Liu Chih Min: "A method for the combination of measuring errors"(NIM Peking, no date), 20 p. (in Chinese)

- Hsiao Ming Yao: "The calculating methods frequently used for estimating experimental errors" (NIM Peking, no date), 23 p.

- Hsiao Ming Yao: "The precise calculation of the confidence factor in combination of errors" (NIM Peking, no date), 34 p.

Conclusion

For most of the specific problems raised in this enquiry it is difficult, if not impossible, to draw definite conclusions on the basis of the answers received. After all, this was not the object of sending out the questionnaire. Nevertheless, a few words could perhaps be said on the general situation at present and on the trends. Since opinions are still changing, due caution is necessary in any case.

There is little doubt that the general subject of how to express experimental uncertainties has come into focus during the last few years. This renewal of interest is striking for those who remember the state of affairs say ten years ago when all seemed to have been settled for ever.

The continuing need for improved accuracy in science and technology together with the powerful methods for data evaluation by computers, now readily available to practically everybody, have brought to light the practical importance of error statements. At present, most data compilations suffer greatly from inconsistency, incompleteness or plain absence of information on the respective experimental uncertainties.

The traditional subdivision of errors into random and systematic parts, although often useful and simple to perform for the experimenter, in most cases only shifts the real problem to the evaluator who rarely knows exactly what to do with this information in his own work.

This leads us to the first and certainly one of the basic problems, namely to the question of the error types. If one arrives at the conclusion that their distinction is necessary, one is immediately faced with the problems of their further use.

As the questions asked in the enquiry and the answers received clearly show, the various problems raised are strongly linked with each other. Hence, none of them can be really solved without taking into due account the influence that a certain choice will have on the other questions. A further condition is imposed by the requirement that the suggestions or guidelines put forward should be useful in practice and easy to apply. A simple rule which can be readily understood or justified will therefore be preferable to some elaborate system lacking transparency.

Quite independent of the class terminology is the more practical question as to the way in which one should express those contributions to the overall uncertainty that are usually estimated by methods depending more on personal judgment than on statistical theory. Here, too, one has to bear in mind the possible usefulness of such a quantity for further processing, and in particular there arises the question of how it should be combined with the other error components. Let us not be unduly discouraged by the apparent fact that it seems to be easier to say what must be avoided than what should be chosen. There is still hope that the pursuit of some general and sound principles can help us to find a solution which will bring the actual situation of disarray to an end.

The diversity of the replies received shows clearly that a uniformity of opinion is not yet reached. On the other hand, the probing questions may have stirred up the minds of some participants and led them to question things which are usually considered as well established. Indeed, the process of thinking over some of the basic problems seems to be well under way. Perhaps the most remarkable outcome of the questionnaire lies in the simple fact that the majority of the participants seem to have no final opinion on most of the problems raised: they are realizing the difficulties involved and are, one has the impression, waiting for sound proposals. This should be a favorable situation for coming to some agreement which, we hope, will then be acceptable to a large majority. Inevitably, it will imply for some (perhaps for many) experimentalists the change of previous habits, and we fully realize that this is a difficult process. Otherwise, however, the goal of uniformity can never be attained.

Some of the problems involved are difficult and far-reaching. A full presentation of the various opinions held is therefore of basic importance, and we hope that the variety of the answers given in this report can be of some help for mutual understanding. Nobody would expect the task of aligning divergent positions to be an easy or quick matter. Rather, we are faced with a difficult and possibly lengthy process, but the prospect of arriving finally at something which is clearer, more useful and generally agreed upon makes the attempt worthwhile and challenging.

March 1980

APPENDIX I

QUESTIONNAIRE ON UNCERTAINTIES*

Deadline date for response : May 15, 1978

A. Introduction

In the course of its last meeting, from the 20th to the 23rd of September 1977, the International Committee of Weights and Measures (CIPM) decided, at the suggestion of Dr. E. Ambler (NBS), to create a study group to examine the problems associated with the evaluation and the presentation of the uncertainties of the results of measurement.

These problems have already brought forth much controversy, generally of an unproductive nature. The principal difficulties, it seems, are associated with the confused nature of what one usually calls "systematic errors". The confusion obviously persists when one wishes to combine these "systematic" errors with errors of a "random" origin, for example when one wishes to establish, globally, tolerance limits.

It seems that agreement can be reached with regard to very high precision measurements, for which it is always desirable to give as much information as possible about the estimation of the uncertainties. On the other hand, one often needs, in common usage, to characterize the uncertainty or the tolerance by a single parameter (or if need be by a very few parameters). It is this problem that we hope to advance.

Different rules have been proposed by professional groups as well as by national laboratories ; they are often contradictory. It seems urgent that we clarify where we stand in this matter and that we attempt to present a consensus at the international level ; the situation may otherwise rapidly become inextricable.

Given the multiplicity of opinions, often divergent, one must choose carefully the subjects to be discussed by the proposed study group. The principal aim of the following questionnaire is therefore to elucidate the essential problems and the points on

* Translated from the French original by L. Maximon, NBS.

which the discussion has a chance of leading to a practical result. In particular, it would be advisable to avoid purely philosophical or mathematical discussions which have little influence on applied matters, and one should find a compromise between the opposing trends of excessive rigor and excessive freedom.

In order to be productive, the study group should be a limited group ; it should be composed of representatives of national laboratories and of international groups interested in these questions. The answers to this questionnaire will serve it as a guide. It is therefore desirable that these answers indicate by brief but explicit comments, the main points to be remembered ; they will be transmitted through the BIPM to all the groups which have sent an answer.

B. Possible topics for discussion

Before stating the questions themselves, it seems useful to us to recall briefly their context in order that the proposed subject be made more precise and to avoid misunderstandings.

1. Standard deviation

Among the different parameters which enable one to characterize the uncertainty associated with a random variable $\underline{x}$, the standard deviation $\underline{\sigma}_{\underline{x}}$ occupies a privileged position. It is defined by

$$\underline{\sigma}_{\underline{x}}^2 = E\{[\underline{x} - E(\underline{x})]^2\},$$ ($E(\underline{x})$ is the expectation value of $\underline{x}$, $\underline{\sigma}_{\underline{x}}^2$ is the variance).

One generally estimates the standard deviation, in a sample of size $\underline{n}$, by the "experimental standard deviation" $\underline{s}_{\underline{x}}$, deduced from

$$\underline{s}_{\underline{x}}^2 = \frac{1}{\underline{n} - 1} \sum_{\underline{i}=1}^{\underline{n}} (\underline{x}_{\underline{i}} - \overline{\underline{x}})^2, \quad \text{where } \overline{\underline{x}} = \frac{1}{\underline{n}} \sum_{\underline{i}} \underline{x}_{\underline{i}}.$$

In analogous fashion, the experimental standard deviation of the mean value $\overline{\underline{x}}$ is estimated by $\underline{s}_{\overline{\underline{x}}} = \underline{s}_{\underline{x}}/\sqrt{\underline{n}}$.

Let us recall that the evaluation of these quantities does not assume a particular distribution ; the independence of the values of $\underline{x}$ and the existence of the variance are sufficient.

Question 1 : Should one recommend the use of the standard deviation to characterize the random uncertainty ?

2. Limits of confidence

In order to make statistical decisions at a given level of probability, one must know the corresponding limits for the variables, called confidence limits (for a probability p of 95 %, for example). Their determination assumes a certain distribution (one most often assumes a normal distribution) for the population, and depends on the size n of the sample used to calculate $s_{\bar{x}}$. The confidence limits are then given by an expression of the form $\bar{x} \pm s_{\bar{x}} \cdot t(p,n)$, where t is the Student factor (for a normal distribution). The hypothesis of a normal distribution becomes less critical for n sufficiently large and t is then practically a function of p alone.

It should be remembered, in any event, that for confidence limits, the simple quadratic addition is only justified if several conditions are fulfilled, such as normal populations, the same confidence level and the same degree of freedom for all the components, whereas the addition of the variances is independent of these conditions.

Question 2 : Should one recommend the use of a conventional probability and the corresponding confidence limits (instead of the standard deviation) ?

3. The nature of the systematic errors

Much has already been written about the estimation of errors denoted as "systematic", their expression and their practical utilization. In fact, even the notion of systematic error leads to arguments, and the attempts at a precise definition are rare. Moreover, it seems difficult to establish rules to apply to a quantity which can cover - at times simultaneously - several different and poorly defined notions. In particular, one can question whether the traditional distinction between random errors and systematic errors is of such a nature that one can always justify their separation and (when it occurs) a different treatment concerning the rules of propagation of errors.

For example, it often happens that the mean value and the standard deviation of a quantity may be estimated (as in the case of the fundamental constants), but one hesitates to attribute a random or a systematic character to the uncertainty.

Question 3a : Is there an essential difference between random and systematic errors ?

Question 3b : Should one recommend a practical rule which enables one to know with which type of error one is dealing ?

4. The presentation of systematic errors

The realistic estimation of an error of a systematic nature (one must obviously not confuse this with a known correction which we assume already applied) is always a delicate matter. It is most often based on uncertain data and requires much judgment on the part of the experimenter. A solid knowledge of the technical details and the theoretical bases of the method of measurement is indispensable. In spite of all precautions, important systematic effects can even elude the most cautious physicist.

A simple means of protecting oneself against such a possibility is to estimate the systematic uncertainties in a "generous" manner, for example on the basis of "maximum" errors. However, such a procedure is inconvenient in at least two aspects : it poses serious problems with regard to the propagation of errors (the usual formulas assume that the variances are known), and it risks hiding the presence of poorly understood factors the detection of which is a prerequisite to further progress in the methods of measurement.

The use of systematic uncertainties which are as close as possible to the standard deviations avoids these inconveniences, but their common use still seems to encounter psychological barriers.

Question 4 : Should one recommend a practical rule for the expression of systematic errors ?

5. The propagation of systematic errors

When a quantity is a function of several variables having random uncertainties, the general law for the propagation of errors enables one to combine the different components and to deduce from them the resulting error. It utilizes the variances and the covariances of the variables in question, as well as the corresponding partial derivatives, and we will assume that it is known in what follows.

While the propagation of random errors can be considered as being well established, the combining of systematic errors continues to engender various proposals.

The use of maximum errors has sometimes led to the habit of describing systematic errors by rectangular densities and by associating with them special rules of calculation. Constructions of this type are not very convincing. They amount to associating the absence of precise knowledge with a uniform probability density in a limited region, which hardly seems realistic. The logical conclusion would, moreover, lead us to associate the linear addition rule to the maximum limits, to retain the property of a maximum. However, the excessively large limits which result from this are of little practical use. More flexible recipes which replace the linear addition by a quadratic addition, according to the number of components, appear arbitrary.

A more convincing rule, which clearly should owe something to the known relations for random errors, has not yet been given.

Question 5 : Should one recommend a practical rule for combining systematic errors (with other systematic errors) ?

6. The combination of random and systematic errors

If one considers the distinction between the two types of errors as merely reflecting a common practice of using certain words, often practical, but without major consequences, their combination poses no real problem. This is provided, of course, that they are all expressed in a form that enables one to estimate their standard deviation. However, difficulties do arise if one considers that these errors are of two fundamentally different types and that combining them has no physical meaning. In this case, a separate treatment must be carried out for each type of error. Problems also exist if the systematic errors are expressed as a "maximum value", for in this case one must assign to them (at least implicitly) a probability which allows one to return to a level of confidence and a standard deviation. The estimation of maximum errors appears thus as a detour which should be avoidable.

Other combinations can obviously be thought of (a mixture of sums of linear and quadratic contributions with more or less arbitrary factors, etc.), but such procedures have little chance of permitting generalization.

Question 6 : Should one recommend a practical rule for combining random and systematic errors ?

7. The form of the final uncertainties

It is generally recognized that, for the establishment of a calibration certificate for example, uncertainties which correspond to a standard deviation, and thus to a confidence level of about 70 %, would be too small, since in almost one third of the cases, the "true" value would lie outside. For such situations, probabilities of the order of 95 % are commonly used. The passage from one confidence level to another should not pose large problems if one starts from an estimation of the corresponding standard deviation. On the other hand, in most scientific applications such a passage does not necessarily occur. Should one, then, try to recommend the general use of a given confidence level ? The same problem occurs for the components in the case in which one neither can nor desires to evaluate a "total" uncertainty.

Besides, for measurements at the highest level of precision, a simple subdivision of the different contributions to the final uncertainty into two groups would, without doubt, be too simplistic and cause the loss of too much information. In this case, a detailed list giving all the causes of error and estimating them is necessary. It should contain all the information that one might need for a possible new evaluation, as, for example, in the framework of a data compilation.

Question 7 : Should one recommend a practical rule for expressing the final uncertainty ?

8. Other questions

The BIPM has tried to clarify these questions on the basis of the information at hand. It may well be that it has neglected some points which appear essential to you, or that it has chosen a questionable order of presentation.

Question 8 : Do other questions appear essential to you ? Does it seem to you that the preceding questions should be considered in a different order ?

C. Supplementary information

- Please describe the practices currently used in your laboratory or that you wish to suggest. Give, if possible, one (or a few) typical examples of their applications (certifications, references to the literature, etc.), mainly for difficult cases.

- Can you recommend references to the literature dealing with the problem of experimental uncertainties ?

APPENDIX II

BIPM Questionnaire on Uncertainties

Response submitted by
U.S. National Bureau of Standards
May 15, 1978

FOREWORD

We support the view that a CIPM ad hoc Committee on Statements of Uncertainty should prepare a report on principles to guide the presentation of information about uncertainties, to be used throughout the international measurement system and at various levels within that system including national standards laboratories, secondary standards laboratories, and calibration certificates for the final user of the instrument. For this purpose it is important that the report contain guidelines for statements of uncertainty, but avoid insofar as possible unresolvable philosophical discussions on statistical theory. The committee should consider existing policies of the various national standards laboratories on statements of uncertainty and seek a suitable international consensus, being careful to avoid procedures which lead to loss of scientific information or make post-analysis of experiments and calibration impossible. It may be possible to provide calibration certificates which include listing of systematic and random uncertainties, give the method of combination, and yield a single number, when appropriate, based on an arbitrary method of addition of random and systematic errors.

Question 1: Should one recommend the use of standard deviation to characterize the random uncertainty?

Answer: If the residuals, $(\underline{x}_i-\underline{\bar{x}})$, in a sequence of measurement results do behave approximately as a "random variable $\underline{X}$," then the use of standard deviations to characterize the random uncertainty is entirely proper.

Comment 1A: Since "randomness" is a concept that cannot be tested in its totality, the degree of approximation can only be tested for selected hypotheses of non-randomness. A minimum requirement is that plots of data, or residuals, against factors of interest (the sequence in which the numbers were obtained, days, temperature, humidity, operators, instruments, etc.) do not show clustering, trend, or pattern to visual inspection.

Comment 1B: The number of degrees of freedom, $\underline{\nu}$, is an integral part of the experimental standard deviation and should also be reported. When the result is not a simple average, $\underline{\nu}$ may be substantially smaller than the total number of individual observations used in computations.

Comment 1C: For routine calibration or certification, it is a poor practice to use the "experimental standard deviation, $\underline{s}$," to characterize the random uncertainty. Rather, the standard deviation $\underline{\sigma}$ as a parameter of the measurement process is preferred. The standard deviation $\underline{\sigma}$ is a stable quantity in contrast to $\underline{s}$ which itself has large amount of variability for small $\underline{n}$ in each set. By accumulating values of $\underline{s}^2$ over many sets of data, the value of $\underline{\sigma}^2$ can be determined for the measurement process under consideration.

Comment 1D: In more complicated situations, the experimental standard deviation could be computed from a number of components of variance representing different sources of variability. These sources of variability should be listed.

Question 2: Should one recommend the use of a conventional probability and the corresponding confidence limits (instead of the standard deviation)?

Answer: It has been our experience that confidence limits are often misinterpreted. Since the confidence limits can be computed easily from the basic quantities (experimental standard deviation and the degrees of freedom, and an added assumption of normal distribution of the residuals), we do not recommend its use instead of the standard deviation.

Comment 2A: Obviously the use of confidence limits (or intervals) is entirely proper for purposes they are designed for. In the statement of uncertainty, however, the assumption of normality and the selection of probability level add to the problem.

Comment 2B: We note that the expression for confidence limit can be generalized to take the form:

$$\bar{x} \pm \left(\frac{s}{\sqrt{n}}\right)t(p, \nu).$$

where n is the number of data points used for the calculation of $\bar{x}$, and ν is the degrees of freedom corresponding to s. Two important special cases can be mentionèd. (1) For k sets of data with n measurements each, s is calculated by pooling and $\nu = k(n-1)$, assuming all the experimental standard deviations from each set estimate the same σ. With an appropriate k (not too small), a reasonable width of confidence interval can be constructed for n as small as 2. (2) If there are differences between sets, then s should be calculated from k averages and $\nu = k-1$.

Comment 2C: It has been shown [1] that for symmetrical distributions, the use of $p = 0.95$ for confidence intervals is more stable (robust) and more nearly correct than the use of $p = 0.99$. If a probability level is to be selected, we recommend the use of $p = 0.95$.

---

[1] John E. Walsh. "Validity of Approximate Normality Values for $\mu \pm k\sigma$ Areas of Practical Type Continuous Populations," Annals of the Institute of Statistical Mathematics, Vol. 8, No. 2, June 1956, pp. 79-86.

Question 3a: Is there an essential difference between random and systematic errors?

Answer: Whether an error is considered to be random or systematic depends principally on the frame of reference. The essential difference between random and systematic errors becomes apparent when the uncertainty of the result is put into actual use.

Comment 3A: The difficulty in agreeing on a definition of systematic error, we believe, stems from the fact that different metrologists look at the problem from a different viewpoint. The standard error of the correction to kilogram N KG-1, based on repeated calibrations against other kilograms maintained at the National Bureau of Standards, is a measure of random error as far as NBS is concerned, but the assigned uncertainty of this same correction, say three standard errors, is a component of systematic error common to all kilogram standards calibrated against it for all laboratories within the United States.

Comment 3B: If objects B and C are both calibrated against standard A which has an assigned uncertainty of ΔA, then in comparing the difference, B-C, the systematic error ΔA cancels out, but the standard errors pertaining to B and C are combined in "quadrature." The sensitivity in detecting a difference between B and C, therefore, is improved when the uncertainty of B and C is reported in two parts, the systematic part which could be common to both, and the random parts which are individual.

Comment 3C: In industry and commerce, within a country, the uncertainty of a national standard is of little consequence since the same error is inherent in all artifacts and instruments through the calibration chain.

Question 3b: Should one recommend a practical rule which enables one to know with which type of error one is dealing?

Answer: Some guidelines to assist one to distinguish one type of error from another would be extremely useful. Examples should be provided for typical situations common in the field of metrology.

<u>Comment 3D</u>: Guidelines and examples are useful tools, but do not substitute for clear and thorough thinking through for the particular problem on hand.

Question 4: Should one recommend a practical rule for the expression of systematic error?

Answer: Limits for systematic errors may be grouped into two distinct categories: those based on experimental data (such as the examples used in comment 3A above), and those based entirely on the judgment of the experimenter. We recommend that separate practical rules be devised for each category.

Comment 4A: A practical rule for the expressions of systematic error, to be useful, needs to be tailored to correspond to the purposes for which the experiments are conducted.

Comment 4B: In many instances limits for systematic errors assume an asymmetric form, i.e., of the form (-a, +b).

Comment 4C: Systematic error in the result may also arise when a biased estimation procedure or computation method is used, e.g., the ratio of the averages versus the average of the ratios. This source of systematic error, however, can be identified and corrected, and its magnitude and direction can be estimated and allowance made for it.

Question 5: Should one recommend a practical rule for combining systematic errors (with other systematic errors)?

Answer: We believe some practical rules for combining systematic error limits with other systematic error limits could be devised and will be useful.

Comment 5A: For a "large" number of systematic errors of approximately equal magnitude, combination by quadrature is reasonable, in the sense some cancellation can be expected. For one or two systematic errors which are two to three times larger than the rest, the linear addition rule appears to be realistic [2]. The choice here, again, depends to a large extent on the purpose to which the result is going to be used.

Comment 5B: Whether one decides on the combination of systematic errors by quadrature, or by linear addition, the possibility remains that two sources of systematic errors may interact and give rise to a systematic error considerably greater than the sum of the two when considered individually. When such possibility exists, the validity of the assigned magnitudes can be checked only with experimental data.

Comment 5C: The combination is actually performed on allowances or limits, for systematic errors (or random errors), not on these errors themselves. Perhaps "random uncertainty" and "systematic uncertainty" should be adopted in lieu of "random error" and "systematic error." To some people the word "error" has the connotation of a known mistake, and as such always raises the question of what prevents it from being eliminated entirely. This is especially true of the term "systematic error."

[2] Churchill Eisenhart. "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems," Journal of Research of the National Bureau of Standards, U.S., Vol. 67C, No. 2, April-June 1963, pp. 161-187. In particular, Table 1 on page 184.

Question 6: Should one recommend a practical rule for combining random and systematic errors?

Answer: Unless one type of error is negligible in comparison to the other type, a combination of the two necessarily results in some loss of information. Following comments we give for Question 3, we feel that the user of the results is in a better position to formulate his own rule.

Comment 6A: In the highest echelon of measurements, say determination of the fundamental physical constants, systematic errors most likely dominate the random errors; in the market place, the reverse is usually true. Between these two extremes, the magnitudes of the two sources of errors are likely to be the same, since the cause of an unusually large systematic error can and should be detected and removed. Calibration and certification fall essentially into this range.

Comment 6B: While practical rules for stating a final uncertainty may be useful (see question 7), it is difficult to rationalize any method for combining limits for two essentially different kinds of errors. At least limits for random and systematic errors should be stated separately; this could be made a rule. One could then discuss how these are combined to yield a single number.

Question 7: Should one recommend a practical rule for expressing the final uncertainty?

Answer: A practical rule for expressing the final uncertainty is useful in the sense that the final uncertainty gives the accepted inaccuracy of the result.

Comment 7A: Since the final uncertainty will include a number of components some of which are not based on experimental data, it is improper to attach to it an assigned level of probability. A 2-sigma or 3-sigma limit for the random component of error (or their equivalent in case of small samples [3]) is preferred. These conventional limits do not have precise probabilistic interpretations unless there is substantial evidence for the shape of the error distribution.

Comment 7B: The rule may need to be different for different circumstances. A few typical examples should be given.

---

[3] Brian L. Joiner, "Student-t Deviate Corresponding to a Given Normal Deviate," Journal of Research of the National Bureau of Standards, U.S., Vol. 73C, Nos. 1 and 2, June 1969, pp. 15-16.

Question 8: Do other questions appear essential to you? Does it seem to you that the preceding question should be considered in a different order?

Answer: The seven questions listed above are the essential ones to be considered by the group. These questions are listed in their natural order.

<u>Comment 8A</u>: Since we feel strongly that the statement of uncertainty, both in form and content, depends on the purpose to which the result is to be used, we have classified these purposes roughly into four categories for consideration by the study group.

Scientific Research

Standard reference data

Standard reference materials

Calibration and Standards certificates

Supplementary Information

A. Examples

We present four examples on the types of statements of uncertainties used or recommended for:

Scientific Research

"Microcalorimetric Determination of Glucose in Reference Samples of Serum," by Robert N. Goldberg, Clinical Chemistry, Vol. 22, No. 10, 1976, pp. 1685-1691. (Appendix 1.)

Standard Reference Data

"Guide for the Presentation in the Primary Literature of Numerical Data Derived From Experiments," Unesco-UNISIST Guide, Report of the CODATA Task Group on Publication of Data in the Primary Literature, September 1973. Reprinted as National Standard Reference Data System (NSRDS) News, February 1974. (Appendix 2.)

Standard Reference Materials

Certificate for Standard Reference Material 993, Uranium-235 Spike Assay and Isotopic Solution Standard, National Bureau of Standards, March 13, 1975. (Appendix 3.)

Calibration Certificate

Report of Length Values, Test Number L7051-1, Dimensional Technology Section, National Bureau of Standards. [Abridged version] (Appendix 4.)

B. References on Experimental Uncertainties

In addition to the NSRDS NEWS mentioned under Standard Reference Data above, we would like to cite the following publications representing the general guidelines for NPL and NBS, respectively.

NPL--A Code of Practice for the Detailed Statement of Accuracy, P.J. Campion, J.E. Barnes, and A. Williams, National Physical Laboratory, Her Majesty's Stationary Office, 1973.

NBS--"Expression of the Uncertainties of Final Results," Churchill Eisenhart; and "Expressions of Imprecision, Systematic Error, and Uncertainty Associated with a Reported Value," Harry H. Ku. Both were reprinted in NBS Special Publication 300, Vol. 1, Precision Measurement and Calibration: Statistical Concepts and Procedures, U.S. Government Printing Office, 1969.

The International Organization of Legal Metrology, in their International Recommendation "Vocabulary of Legal Metrology," also gives guidelines to the expression of uncertainty. See in particular Chapter 8 "Errors in the Results of Measurements and Errors of Measuring Instruments," and Chapter 9 "Conditions of Use and Metrological Properties of Measuring Instruments."

APPENDIX III

Réponses du BUREAU NATIONAL DE METROLOGIE, établies en liaison avec :

- l'E. T. C. A.
- l'I. N. M. (C. N. A. M.)
- le L. C. I. E.
- le L. M. R. I.

QUESTION 1 :

- Faut-il recommander l'usage de l'écart-type pour caractériser une incertitude aléatoire ?

REPONSE

- L'écart type présente l'avantage de caractériser une incertitude aléatoire quelle que soit la loi de distribution. "Néanmoins, comme on ne peut en connaître qu'une estimation au moyen de l'écart-type expérimental, il convient de donner le nombre de valeur utilisées. Dans certains cas, on peut également indiquer l'étendue de la dispersion des résultats.

QUESTION 2 :

- Faut-il recommander l'usage d'une probabilité conventionnelle et des limites de confiance correspondantes pour caractériser une incertitude aléatoire ?

REPONSE

- Il n'existe pas à proprement parler de choix à effectuer entre l'usage de limites de confiance ou de l'écart-type. En effet lorsque la loi de distribution est inconnue seul l'écart type peut-être estimé.

  L'utilisation de la limite de confiance s'avère commode dans la pratique pour caractériser l'incertitude d'un résultat final, mais, afin de ne pas perdre d'information, il faut indiquer également le nombre de degrés de liberté (ou le nombre de mesures

  Par ailleurs il serait souhaitable de normaliser le niveau de confiance correspondant (95 % de préférence).

  Enfin l'utilisation des limites de confiance pour les calculs intermédiaires est à proscrire.

QUESTION 3 a et 3 b

- Y a-t-il une différence essentielle entre erreurs aléatoires et erreurs systématiques ? Faut-il recommander une règle pratique permettant de savoir à quel type d'erreur on a affaire ?

.../...

REPONSE :

- Il existe une différence entre erreurs aléatoires et erreurs systématiques. L'erreur systématique pouvant être définie comme l'erreur qui, lors de plusieurs mesurages effectués dans les mêmes conditions, de la même valeur d'une certaine grandeur, reste constante en valeur absolue et en signe ou qui varie selon une loi définie quand les conditions changent.

  L'erreur aléatoire (ou fortuite) peut être définie comme l'erreur qui varie d'une façon imprévisible en valeur absolue et en signe lorsqu'on effectue un grand nombre de mesurages de la même valeur d'une grandeur dans des conditions pratiquement identiques. (c.f. Norme NFx07-007). Ceci peut être interprêté statistiquement de la façon suivante. Pour chaque mesure il existe un écart appelé "erreur", entre la valeur mesurée et la "valeur vraie". Cet écart peut être considéré comme une valeur prise par une variable aléatoire dont l'espérance mathématique, ou "biais", représente la partie systématique et dont l'écart type caractérise la partie aléatoire.

  Si le biais est négligeable par rapport à l'écart-type, on dit que l'erreur est à caractère aléatoire dominant. Dans le cas contraire l'erreur est dite à caractère systématique dominant. Entre ces extrêmes, toutes les situations peuvent se présenter.

  Cette distinction entre erreur systématique et erreur aléatoire n'est possible que si le biais a pu être déterminé et dans ce cas on en tient compte en effectuant des corrections systématiques qui ont pour but de le réduire. Cependant en général le biais est inconnu et on a simplement une estimation subjective de sa dispersion autour d'une valeur que l'on suppose habituellement nulle.

Par ailleurs il convient de faire les remarques suivantes :

- Erreurs aléatoires et erreurs systématiques correspondent à des notions physiques différentes il est donc utile de conserver les deux appellations.

- Augmenter le nombre de mesurages permet de réduire la valeur de l'écart type mesure de la dispersion des erreurs aléatoires, mais ne peut en aucun cas diminuer l'influence des erreurs systématiques (ceci peut être obtenu par des règles expérimentales et non par répétition des mesures).

QUESTION 4 :

- Faut-il recommander une règle pratique pour l'expression des erreurs systématiques ?

REPONSE :

- Il serait souhaitable d'arriver à une expression uniforme qui permettrait de caractériser la dispersion du biais. Il semble que la pratique actuelle qui consiste à utiliser les limites supérieures d'erreur conduise à des résultats pessimistes. Cependant tout

.../...

autre règle pratique et plus réaliste est alors conventionnelle et dépendante de la grandeur mesurée (voir de la méthode utilisée).

QUESTION 5

- Faut-il recommander une règle pratique pour combiner entre elles les erreurs systématiques ?

REPONSE :

Lorsqu'on est en présence de plusieurs erreurs systématiques, on dispose d'estimations subjectives pour caractériser la dispersion des différents biais. On peut envisager dans ces conditions, pour estimer la dispersion totale, et si l'on peut raisonnablement admettre que ces biais sont indépendants, de combiner ces dispersions comme s'il s'agissait d'erreurs aléatoire

- On doit faire des réserves sur les points suivants :

  - le nombre d'erreurs systématiques doit-être suffisammen important (et en tout cas supérieur à 4).
  - Seul l'écart type pourra alors être calculé, on ne pourra calculer des limites de confiances qui nécessiteraient la connaissance de la loi de distribution des biais.

Cependant il convient d'insister sur le fait que le non respect des hypothèses sous-jacentes à cette règle de combinaison des erreurs systématiques risque de détériorer la qualité de la mesure en sous estimant l'importance des erreurs systématiques.

QUESTION 6

- Faut-il recommander une règle pratique pour combiner les erreurs aléatoires et systématiques ?

REPONSE :

- Dans l'éventualité de l'existence d'une règle pratique de combinaison des erreurs systématiques entre elles, telle que celle développée a point 5, on peut envisager une règle qui combinerait les erreurs systèmatiques et aléatoires de la façon suivante : pour obtenir la variance globale des résultats, on ajouterait la variance estimée de la moyenne des mesures à la variance obtenue par combinaison des variances estimées correspondant aux différents biais systématiques.

- Si au contraire une telle règle ne pouvant être retenue l'approche qui consiste à exprimer les erreurs systématiques en "valeur maximale" et à ajouter aux limites de confiances (des erreurs aléatoires) les bornes calculées (des erreurs systèmatiques) nous semble être, bien que pessimiste, celle permettant de ne pas masquer l'importance des erreurs systématiques.

.../...

QUESTION 7 :

- Faut-il recommander une règle pratique pour l'expression de l'incertitude finale ?

REPONSE :

Même dans le cas où les différentes erreurs peuvent être traitées par une règle de combinaison des erreurs systématiques entre elles (voir point 5) et par une règle de combinaison des erreurs systématiques et des erreurs aléatoires (voir point 6), la représentation de l'incertitude globale sous forme d'un intervalle de confiance correspondant à un niveau de confiance donné n'est en général pas possible. On est donc amer à exprimer cette incertitude en se basant sur la variance globale de résultats. Toutefois pour certaines applications (certificats d'étalonnage par exemple), il serait souhaitable que l'on puisse uniformiser cette expression sous forme d'un multiple de l'écart-type estimé.

Cette restriction quant à l'impossibilité de définir un intervalle de confiance présente en effet des difficultés pour la présentation des résultats.

En conclusion au présent questionnaire nous tenons à souligner les dangers d'une approche entièrement statistique des erreurs systématiques, approche que nous avons essayée de développer au cours des réponses. Il est indispensable que les hypothèses soient vérifiées, notamment celles relatives au nombre d'erreurs systématiques prises en compte et, à leur indépendance. Si tel n'était pas le cas sous estimation de l'importance des erreurs systématiques contribuerait à la dégradation de la qualité des résultats de mesure.

Des références bibliographiques, ainsi que quelques exemples typiques d'applications (certificats d'étalonnage, documents internes) issus des laboratoires ayant contribué à ce travail sont joints en annexe.

APPENDIX IV

Reply by the Physikalisch-Technische Bundesanstalt

The PTB has recently formed a working group charged with both the terminological and the practical aspects of the problem statement of uncertainty of measurement.

After a series of discussions, it became clear that it is of paramount importance to define the relevant terms as a basis for meaningful international negotiation.
Being as yet only the outcome of our initial discussions in the working group with its small circle of participants, including some DIN experts, our recommended glossary of terms as well as our answers must be regarded as a draft.

We expect that there will be further questions and a stipulation of priorities, and in this event will then elaborate our ideas in more detail.
Finally, the purpose and aim of a statement of uncertainty of measurement should be described and explained. This is the reason for the following short clarification of the aim and characteristics of three metrological levels:

1. High-level metrology

- The aims of the measurements are e.g. determination of fundamental constants with very low uncertainty, and international comparisons between measuring setups for primary standards.

- In test protocolls or publications all the individual information should be quoted which is relevant for further treatment (processing) of the measured results.

2. Medium-level metrology

- The aims of the measurements are e.g. determination of physical quantities in the pure and applied sciences with medium measuring uncertainty, and the determination of the calibration factor of secondary standard measuring devices.

- Depending on the field of work, component uncertainties of a measuring result must be quoted as well as the total uncertainty in the form of a single figure.
  The latter must be available so that it may be used later on as a component uncertainty in other experiments where necessary.

3. Low-level metrology

- The aims of the measurements are e.g. the determination of physical quantities in routine measurements in commerce, trade and in medicine, as well as the determination of the calibration factor of routine-work instruments.
- Only the total uncertainty (not the component uncertainties) of a measured result (or of the calibration factor of a routine-work instrument) and the confidence level are required to give an indication of the reliability and to allow faulty devices to be rejected during instrument comparisons. It is not envisaged that the quoted uncertainty of measurement be processed further.

In the following answers to the questionnaire we attempt to take the needs of the different levels of metrology into account.
An internationally uniform representation and determination of the uncertainties according to simple rules (as far as possible is both very desirable and expedient.
For the glossary of terms the following documents have been taken into account:

British Standards Institution Glossary of Terms used in Metrolo (1974)

IEC Publication 359, 1971

ISO R 645, 1967

VDI/VDE 2600, 1973

OIML Vocabulaire de Métrologie Légale (1969)

CCIR Draft Report 7/179, 28 March 1978

PTB Draft for a Glossary of Terms connected with Measuring Uncertainties

1. *Quantity to be measured, measured quantity.*
   A quantity subjected to a process of measurement.

2. *Value (of a quantity).* The quantity expressed as the product of a number and the unit of measurement.
   e.g. 5.3 m, 20°C

3. *Influence quantity* . A quantity which is not the subject of the measurement but which influences the value of the quantity to be measured, or the indication of the measuring instrument, or the value of the material measure reproducing the quantity.
   NOTE. The influence quantity can arise from the ambient conditions or from the instrument itself.
   e.g. Temperature, mains frequency, self-heating of an instrument, response time.

4. *Correct value (or conventional true value).*
   a) Value determined with an uncertainty low enough for each particular case. b) With respect to one influence quantity: value to be expected when the influence quantity is at its reference value (i.e. the value to which the calibration refers).

5. *Error (of indication* of a measuring instrument or of a material measure). The difference: " measured value minus correct value" or "designated value minus correct value".
   The following two terms are given with some reservation as being useful only when simplifying assumptions can be made:

5.1 *Random error (error of repetition).* The difference: "measured value minus value of the mean of a set of measurements". The random error is due to influence quantities which fluctuate during the measurement.
It cannot be uniquely defined when the time period of the fluctuations is comparable with the measuring time.

5.2 *Systematic error.* A concept for the possible difference: "measured value minus correct value", due to steady influence quantities or biases introduced by foregoing measurements which cannot be repeated. (known systematic errors have to be corrected for; an uncertainty remains).

6. *Uncertainty (of a measured value).* It expresses the magnitude of a possible deviation of the measured value from the correct value. To quote the uncertainty quantitatively it can be given as an interval around the measured value or around the best estimate of a quantity which embraces the correct value with a certain probability. Frequently it is possible to distinguish the following two components:

6.1 *Random uncertainty (uncertainty of repetition).* A certain multiple of the experimental standard deviation of the mean of a set of measurements. It can be diminished by increasing the number of measurements.

6.2 *Systematic uncertainty.* This uncertainty can only be estimated on the basis of knowledge and/or experience obtained outside the current measurement.

7. *Component uncertainty (or partial uncertainty).* Uncertaint due to the effect of single influence quantities (see No.3 calculation errors, biases or to the fluctuation of indicated values.

8. *Total or overall uncertainty.* A certain combination of all types of component uncertainties.

PTB Draft answer to BIPM Uncertainty Questionnaire

We wish to point out that our own national standard dealing with these and other questions of metrology (DIN 1319 parts 1-3) is being completely revised. At the present time therefore our answers can be regarded as guideline only in the discussion which shall be continued in the PTB as in DIN.

Questions 1 and 2: This should be left to individual decision, but the choice should be explicitly indicated. In both cases the number of degrees of freedom should be quoted.

Question 3 a: The answer depends on the type of experiment considered. It is "yes" only in cases where the fluctuating influence quantities can be clearly divided into those of which the time constants are either short or long compared with the measuring time. In many experiments the two times are comparable and this distinction is not applicable; the experimental standard deviation varies according to the measuring time for the sample.

Question 3 b: see answer to 3 a.

Question 4: The term "systematic error" should not be used in this context, only "systematic uncertainty". With this alteration the answer is "yes".

Question 5: Yes, one should recommend rules for the combinatic of systematic uncertainties (see 7 and 8 in the glossary).

Question 6: Yes, one should recommend appropriate rules for combining systematic and random uncertainties.

Question 7: Same answer as question 6

Question 8: One should discuss what information should be given in test certificates in different fields of work.

Literature: S. Wagner: PTB-Bericht FMRB 31/69 Nov. 1969
H. Reich: PTB-Mitteilungen 86 Pg. 421 1976

APPENDIX V

REPLY TO BIPM QUESTIONNAIRE ON UNCERTAINTIES

BY THE NATIONAL PHYSICAL LABORATORY

Preface

In replying to the BIPM questionnaire on uncertainties the NPL is of the opinion that the most important task a BIPM topic group should undertake is to define clearly the basic concepts involved, in order to provide a secure foundation on which to base practical guidelines. Bearing in mind the difficulties of this venture, and while wishing it well, some doubts must be expressed about the likelihood of its success.

Introduction

There is no doubt that much discussion on uncertainties is unproductive. The reason for this is usually because the participants in the discussion are using the same words to mean different things. Before getting down to discuss practical rules for the handling of uncertainties it is absolutely essential that the study group should first obtain agreement on the basic concepts and terminology, otherwise they will spend their time arguing at cross purposes.

Before answering the specific questions on the BIPM questionnaire, this memorandum discusses these basic concepts under three headings:- (a) distinction between errors and uncertainties, (b) categorisation and definition of uncertainties, and (c) nomenclature of uncertainties. Included under this topic would be the question of whether such debased words as error, accuracy and precision can have an internationally agreed meaning.

Distinction between errors and uncertainties

When science takes over existing words from a language it is highly desirable that any restricted meanings given to them do not differ much from the common everyday meanings of the words. Thus in normal English usage the word "error" is more or less synonymous with "mistake, wrong, difference, disagreement, discrepancy", whereas the word "uncertainty" is more or less synonymous with "doubt, vagueness, indeterminacy, ignorance, imperfect knowledge". In simple experimental theory therefore the error of a measurement is the difference between the measured value and the true value of the quantity that is being measured. The uncertainty is the doubt that is expressed about the accuracy of the measurement, i.e. the range within which the true value is thought to lie.

In many cases it is possible to argue that the quantity that is being measured must have a true value. Stochastic quantities however will have expectation values. In other cases the quantity that is being measured will possess an inherent uncertainty. In all cases however the true value is unknown and unknowable, and hence the error of the measurement must also be unknown and unknowable. Discussion of errors is therefore usually unprofitable, and in most cases when errors are being discussed, it is really the uncertainties that are meant, and a clear distinction between the two terms must be made at all times. To use phrases like "propagation of errors" when what is meant is "combination of uncertainties" is indefensible, as it leads to erroneous mental concepts and arguments at cross purposes.

Categorisation and definition of uncertainties

Discussion about uncertainties has been strongly influenced by the old distinction between random errors, which cause fluctuations in the measurements, and systematic errors (or bias) which cause constant errors in the measurement. It is not very useful however to divide uncertainties in the same way. Instead it is proposed that uncertainties be divided into two categories depending on their method of derivation and not on how the measurements are affected.

The two categories can be thought of as dividing the facts from the guesses. On the one hand we have standard deviations, confidence limits, etc., which are calculated from the results of repeated measurements; given a copy of the experimental readings, anyone should be able to calculate exactly the same values for these uncertainties. On the other hand we have estimates of uncertainties derived, for example, from deliberate variation of external influences, additional measurements to determine correction factors, examination of longer-term variabilities in the results, and sometimes just intelligent estimation by the experimenter; a different person examining the same experimental results might arrive at a different answer for this uncertainty.

It is not easy to give precise definitions for these two types of uncertainties, but the study group should make an attempt. The following are suggested:-

(a) uncertainties on the measured value estimated by statistical methods from the results of repeated measurements.

(b) uncertainties estimated by non-statistical methods to allow for the effect on the measurement of both fixed and variable influences.

Nomenclature of uncertainties

Obviously the definitions just given of these two uncertainties are too long to be used as names. Many different names have been put forward in the past, notably by workers at NBS, whose papers on the subject are collected in the excellent NBS Special Publication 300. They usually suffered from being too unwieldy, and were not widely copied. In 1973 the above categorisation and the names random uncertainties and systematic uncertainties were put forward by Burns, Campion and Williams in a Letter to Metrologia. They suffer from the obvious disadvantage of being very similar to the terms random errors and systematic errors which are divided in a different way. Nevertheless, the terms are brief and have since been widely and increasingly used in the scientific literature.

At first glance the foregoing discussion may seem highly philosophical but it is not. It is intended to provide a firm practical foundation on which a superstructure of rules and recommendations can be built. The study group should first spend its time agreeing on these basic concepts.

After this they may then go on to discuss practical rules for handling uncertainties. However it may never be possible to arrive at a completely uniform system because different requirements and different traditions may demand different kinds of statements and different methods of derivation. But if in a statement of uncertainty the actual meaning of the statement is uncertain, then the statement is valueless.

ANSWERS TO QUESTIONS ON BIPM QUESTIONNAIRE

It should be borne in mind that the questionnaire is not asking at this stage for individual or collective views on how uncertainties should be handled. It is asking instead for advice about what topics it would seem most profitable for the proposed BIPM study group to spend its (expensive) time discussing.

Question 1

The study group could usefully clarify the difference between a statement of the repeatability of an instrument under test (usually characterised by the standard deviation of individual readings) and a statement of the random uncertainty of a calibration factor or measurement since the random uncertainty can contain many components in addition to that arising from the dispersion of the readings taken during a calibration or set of measurements.

Question 2

There are two common ways of stating the random uncertainty of a measurement:-

(a) standard error of mean, together with effective number of degrees of freedom.

(b) confidence limits to a stated probability level, together with effective number of degrees of freedom.

If one is given, the other can be calculated assuming the distribution. It would seem unprofitable to spend much time discussing this as different circumstances may determine which method is used. For the choice of probability level, see answer to question 7.

Question 3

This is an important topic, and is discussed in the introductory sections to this memorandum.

Question 4

There are many experimental techniques available that can be used to help in obtaining an estimate of how a measurement might be affected by external and internal influences. The study group might consider whether it is possible to propose some broad guide lines on these techniques, although a detailed discussion would be impracticable because they differ so much from one field of science to another. However, having carried out these supplementary tests and applied suitable corrections to the result, one is left with residual uncertainties which are the systematic uncertainties which have to be estimated and written down.

There is one objective way in which these uncertainties can be stated, and that is to quote uncertainty limits so wide that the influence being considered could not possibly affect the result by a greater amount under any possible circumstances. This however is not the way it is done, for it would make the uncertainty limits useless from any practical point of view of comparing results. What is expected of the experimenter is that he should quote the narrowest limits that he can convince himself are reasonable, taking all considerations into account. The systematic uncertainty limits are therefore estimated using subjective judgement. And being a subjective judgement, the only way to describe the criteria adopted in estimating the uncertainty is by means of a subjective phrase.

Although it sounds objective, the commonly-used term "maximum limits of error" is highly subjective, but implies a safe estimate with rather wide limits, perhaps comparable with 99.9% confidence limits. On the other hand one term that has been suggested "as likely as not" implies narrow limits, implying some sort of comparability with the statistical "probable error", i.e. 50% confidence limits, while another term that has been suggested "twice as likely as not" is intended to be comparable with 67% confidence limits or the standard error of the mean.

The term used in some parts of NPL is "certain beyond reasonable doubt" which is of course the criterion used in obtaining verdicts from juries in the English judicial system, and may be regarded as suitable for combination with 99% confidence limits. If the study group agrees with this assessment of how systematic uncertainties are estimated , then they could usefully discuss:-

(a) the choice of subjective probability level to be adopted (see also answer to Question 7).

(b) the choice of phrases that might be used to describe this subjective probability level.

Question 5

There are at least five different methods in common use for combining systematic uncertainties:-

(a) add the uncertainties linearly (most workers would agree from experience that this is an overestimate).

(b) add the uncertainties quadratically (many workers feel, again from experience, that this is an under-estimate).

(c) If one or two uncertainties are much larger than the others, add them linearly to the quadratic sum of the smaller uncertainties.

(d) pick a plausible figure somewhere between the linear sum and the quadratic sum.

(e) use a mathematical formula, assuming some systematic error probability distribution, to calculate a value aimed at being comparable with a statistical function such as the standard error of the mean.

What is required is a convincing theory based on the concept of uncertainty (not errors) from which a method for combining systematic uncertainties can be derived. If this is not found to be possible, then a second best would be a clear classification and description of the various methods, with some guidance as to their relative advantages and areas of application.

This question cannot be divorced from the combination of systematic and random uncertainties, so see also the answer to question 7.

Question 6

Most people would accept that random and systematic uncertainties should not be combined unless they are both estimated at approximately the same probability level (a subjective probability in the case of systematic uncertainties). This cannot usefully be discussed separately from the next question.

Question 7

As suggested in the questionnaire, it is desirable to quote the random and systematic uncertainties separately in measurements of the highest level of accuracy. In this case the individual estimates of each of the component uncertainties should be listed, and given all this information it is possible for a reader of a report to combine them in any method he thinks fit.

The main need for a statement of the overall uncertainty is felt at the lower levels in the calibration hierarchy, and one may reasonably ask what practical use is made of such a statement. The answer, more often than not, is that it is needed for legal, or pseudo-legalistic, purposes. For example, take the case of a manufacturer who purchases an expensive measuring instrument, specially calibrated by a secondary standard laboratory, and, using this instrument as his working standard, he then manufactures a large number of components whose dimensions are critical to their use. If the calibration of his working standard turns out to have a 1% error (i.e. a 1% discrepancy when compared later with another secondary standard) then the manufacturer is entitled to feel aggrieved if the calibration certificate for that instrument stated an overall uncertainty of $\pm$ 0.1% . It would be of little interest to either party in the dispute to know that of that 0.1% the random uncertainty was $\pm$ 0.07% and the systematic uncertainty was $\pm$ 0.07%, and that they had been combined quadratically.

In most cases a manufacturer will choose a measuring instrument whose uncertainty is considerably smaller than his required manufacturing tolerances. In these cases it matters little what methods are used to combine and state random and systematic uncertainties. However, this is not always possible, and it is when the uncertainty of calibration approaches the required manufacturing tolerances that the statement of uncertainty becomes important. It has been claimed a number of times that confidence limits at the 99% probability level are only appropriate at places like primary standards laboratories, and that elsewhere 95% or even 90% confidence limits are regarded as acceptable and that the systematic uncertainties should be estimated at the same subjective probability level. However, what this means in principle is that an average of one instrument in twenty (or 1 in 10 for 90% probability levels) will have a true response that lies outside (by an unknown amount) the uncertainty limits around the stated calibration factor. This probability of error may in some cases be unacceptable to manufacturers.

What this might indicate is that systematic uncertainties should be estimated using criteria like "maximum limits of error" or "certain beyond reasonable doubt", and that they should be combined with confidence limits estimated at the 99% level. This provides guidance on the answers to Question 2 and 4. Whether it implies that systematic uncertainties need to be combined linearly is open to doubt, but the study group will have to consider this. The route by which the uncertainties can be combined should also be discussed, whether it can be direct or via the intermediate calculation of standard errors of means and a comparable quantity for systematic uncertainties.

There is one tricky point concerned with systematic uncertainties that the study group may like to consider. Take the case of an experiment in which the individual systematic uncertainties have been estimated at a subjective probability level of 99%. Suppose now someone subsequently wants to use the result of this experiment as supplementary to another experiment in which the systematic uncertainties have been estimated at a subjective probability level of 67%. Can a method be proposed to convert the first set of uncertainties to be compatible with the second set?

A similar point is the combination of the random uncertainties of supplementary or previous experiments, the results of which are to be used in another experiment. Should the random uncertainties of these other experiments now be regarded as random or systematic uncertainties in the later experiment?

It might also be worth examining modern information theory to see if it can throw any light on the amount of information lost when detailed lists of uncertainty limits are compressed into single overall uncertainty limits.

A more minor but nevertheless very practical point is the question of the treatment of those readings which lie a considerable distance from the mean.

Supplementary information

The enclosed document describes the procedures used in the British Calibration Service. It has been pointed out elsewhere that it is rather unlikely that a single procedure will be found which will be suitable for all levels in a calibration hierarchy. The attached document therefore is offered as an example of the necessarily shortened form of an uncertainty statement which has to be used at a level someway removed from the apex of a hierarchy and is not to be interpreted as a statement of NPL policy in this area.

Enclosure: British Calibration Service guidance publication number 3003 "The expression of uncertainty in electrical measurements"

5 May 1978