Hans Andersson

The Use of Uncertainty Estimates in Testing

SP
Swedish National Testing and Research Institute
SP REPORT 1993:47

Abstract

The statistical background for assessment of measurement uncertainty is reviewed, and related to the recommendations which have recently now been presented by ISO and WECC. It is found that the recommendations, although they are now being introduced as mandatory requirements in laboratory accreditation, are not well-adjusted, neither technically nor with regard to customer needs, to the area of testing.

The impact of uncertainty estimates on risk assessments and on comparisons with conditions for approval is discussed. It is pointed out that as well the distributions of risk as those of the product properties and test methods should be regarded in a decision on rational procedures for assessment of uncertainty in testing. Further it is recommended that information and knowledge is disseminated to the standardization society.

SP
SP-Rapport 1993:47
ISBN 91-7848-433-2
ISSN 0284-5172
Borås 1993

Swedish National Testing and Research Institute
SP-Report 1993:47

Postal address:
Box 857, S-501 15 BORÅS
Sweden
Telephone + 46 33 16 50 00
Telex 36252 Testing S
Telefax + 46 33 13 55 02

Contents

Summary

The present ISO- and WELAC-recommendations on how to estimate and declare uncertainty in test results from accredited laboratories are critically assessed from a laboratory standpoint.

It is found that the assumptions and theoretical basis for the recommended treatment of random and systematic uncertainty components are such that the approach may work well in metrology where many variables are included, and statistical evaluations can be used extensivelty. The assumptions can always be debated on a principal level but this is not meaningful for laboratories to pursue however tempting it may be.

Instead the practical consequences in a typical test assignment are considered. Then it can be stated that the assumptions behind the recommendations are not compliant with the technical conditions, and that the way to express uncertainty may be misleading, or even impossible to follow. To achieve quality in the test assignment the declaration of uncertainty should be adjusted to customers´ needs, to standardization, and to the intended use of the results for comparisons with design rules and regulations.

Standardization organizations should take on the task to include and integrate uncertainty declarations in testing standards, taking technology and inteded use into consideration. Laboratories should produce background material and disseminate it to accreditation and standardization organizations for an improved view on things in the future.

1 Purpose

Presently there is a strong trend to introduce quality systems and accreditation in laboratories. In Europe this trend is largely motivated by the establishing of the common market, where openness and transparency requires ways for all laboratories which so desire to get access to the market on equal terms. As regards the treatment of uncertainty, the requirements of the accreditation bodies have caused considerable unease among laboratories. In many cases the requirements seem unnatural and not well fitted for their purpose.In this note some views of a test engineer are expressed on how uncertainty of test results should be estimated and declared.

The accreditation organisations are in general closely liased with metrology. This may explain why their requirements regarding uncertainty to a large extent stems from, and are adjusted to, technology and aims in this field. They are connected to development work and the strive to understand and successively eliminate causes for discrepancies in comparisons of national standards, and in the chains of traceability to industrially used measuring instruments. Generally, the personnel in a calibration laboratory consists of highly educated specialist, devoting their time to the exploration of a few physical entities. They need working tools to discriminate possible differences, and they can afford repeated measurements to get the necessary background for statistical analysis.

In the testing laboratory, on the contrary, the work is performed by test engineers, specializing in the properties of products and materials. They use measurements to evaluate these properties, and to report the results to customers who demand a guarantee for the correctness of the result, in order to feel sure with regard to product liability, design requirements etc. Normally they are end users, not interested in carrying the traceability further, but to use a tolerance limit in discussions with authorities or commercial partners.

Many industrial sectors have long-standing traditions as regards the expression of technical data and how to use them for control and design, e.g. in codes and regulations. Few of the users are, or will be, familiar with subtle distinctions, such as error limit vs standard uncertainty. Therefore the requirements of the accreditation bodies to use complicated guides, as the ISO/TAG4/WG3 [1], as a universal tool for assessing and expressing uncertainties in test results may end up with serious trouble. This was expressed e.g. at the EUROLAB Workshop in December 1992 [2], where the group studies clearly indicated a need for moderation. An indiscriminate requirement to use the methods of [1] in testing laboratories may cause not only considerable costs but also severe misunderstandings, which is indeed not the intention of quality assurance work.

2 Uncertainty of experimental results

2.1 Basic considerations

The activities in a testing laboratory generally consist of standardized experiments to evaluate, as accurately as is meaningful with regard to economics and safety requirements, the properties of materials or components. Then measurements are made of various quantities. Each measurement is afflicted with errors, influencing the end result. These errors are not known, but there is an uncertainty about their exact magnitude in each measurement. In standardized testing one can generally state realistically small limits within which the errors must almost certainly be lying.

In most cases it can be presumed that there is a well-established theoretical functional relationship between the measured quantities, x_i, i = 1, N, and the desired value of a property, y,

$$y = f(x_1, x_2, ..., x_N)$$

If the deviations, errors[1], in the measurements of x_i, are denoted δ_i, the deviation in y can be written, to the first order,

$$dy = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i} \delta_i$$

Further, if the limits for δ_i can be stated as $\pm\delta_{i0}$ the deviation *dy* will be at most

$$dy_{max} = \pm \sum_{i=1}^{N} \left| \frac{\partial f}{\partial x_i} \right| \delta_{io} \qquad (1)$$

Since this formula normally will exaggerate the real deviation of y, partly since the limits δ_{i0} are usually estimated too generously and partly since it is highly unlikely that all errors are maximal and of the same sign, other methods to express *dy* have been suggested, which are based on the theory of statistics.

Assuming that the deviations in x_i are due to the fact that they are random variables, the general formula for combining variances can be used,

$$\sigma_y^2 = \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_{x_i}^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \sigma_{x_i} \sigma_{x_j} \rho_{ij} \qquad (2)$$

[1] In testing it is meaningful to talk about errors, and true values, contrary to what is the case in fundamental metrology, since in testing the quantities measured have errors which are several orders of magnitude bigger than the uncertainties of the corresponding basic physical entities, as voltage, length and time. Then these entities can be considered as exact references.

Here σ denotes standard deviation and ρ_{ij} correlation coefficients. In the following it is presumed, for simplicity in the principal reasoning, that the x_i are uncorrelated, which means that $\rho_{ij} = 0$.

If the standard deviations of x_i were known the standard deviation of y would be known. In a practical situation the standard deviations have to be replaced by estimates, s_i, from experiments or experience, and the result will be an estimate of σ_y, s_y. Special treatment, not very probable in testing, is required in non-linear cases, e.g. when $\partial f / \partial x_i = 0$ locally. In [1] and [10] it is suggested that (2) is used to determine s_y as a standard uncertainty also in cases where not all the x_i are random variables but have constant, unknown errors, and to use the quantity $2s_y$ as a measure, on a 95 % confidence level, of the uncertainty of y. Theoretically this is based on the assumptions that y is normal and that $s_y = \sigma_y$.

The formula (2) combines variances of the distributions of x_i to the variance of the distribution of y. However, the distribution of y is generally not a simple function of the distributions of x_i. Only in the case where all x_i are normal, also the distribution of y is normal. In real life one can be sure that the distributions of x_i are not normal, but limited and of some unknown shape. Then, the distribution of y will not be normal either, but some limited distribution, and the standard deviation σ_y can not, principally, be used to make predictions about significance levels with the normal distribution as a basis, particularly when N is small. In practice, it appears that, except for some extreme situations, which have been discussed in [3] and [4], and are illustrated below, the distribution of y is nearly normal in its central parts, and that the confidence level of 95 % for $2\sigma_y$ is not much affected. This has been explored in detail by e.g. Dietrich [5].

Consider, as an illustration, a case where all the x_i have identical rectangular distributions with unit width, and that all $\delta f/\delta x_i = 1$. Then, if first N = 2 is chosen, the convolution of x_1 and x_2 will give a triangular distribution for y, as depicted in Figure 1. Here, $\sigma_1 = \sigma_2 = 1/2\sqrt{3}$, which means that a 1σ - limit covers 57 % of the distribution area while the 2σ-limit exceeds the limits of the distribution by 15 %.

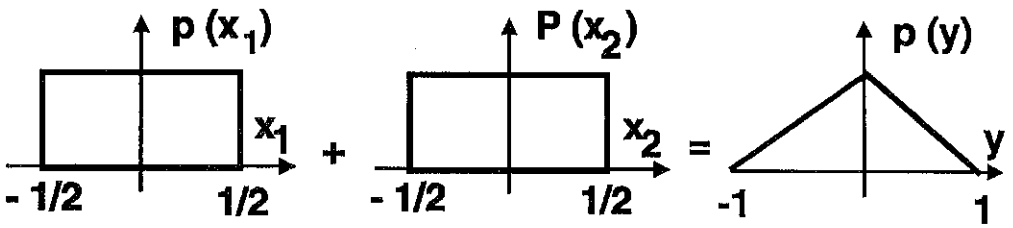

Figure 1. Combination of two rectangular distributions with unit width.

For the resulting distribution $\sigma_y = \sqrt{2}\,\sigma_1 = 1/\sqrt{6}$, according to (2). Then, the 1σ-level covers 65 % and the 2σ limit 96 % of the distribution, which is close to the corresponding figures for the normal distribution. When four rectangular distributions are combined, the difference between the resulting distribution for y and a normal distribution with the same σ is nearly indistinguishable [5]. The differences will be even smaller when more natural distributions than the rectangular one are chosen.As mentioned, problems may arise when the uncertainty of one of the x_i is dominating. Then, the 2σ limit may exceed the sum of δ_{io}, which is, of course not desirable.

If N equal rectangular distributions are assumed the formulas (1) and (2) yield

$$\frac{dy_{max}}{2\sigma_y} = \frac{\sqrt{3}\sqrt{N}}{2}$$

Hence, the relationship between dy_{max} and $2\sigma_y$ increases with the number of terms. For N=1-3 it is around unity, making the need to use (2) instead of (1) less important. Still, the percentage of the distribution outside $2\sigma_y$ stabilizes very quickly to approximately 5 %, which should remembered in the sequel of this note.

The importance of the case where one or two terms are dominating, which is usual in testing, is accentuated by the, also usual , situation where the functional relationship can be written

$$y = \prod_{i=1}^{N} x_i^{k_i} .$$

Then, some simple calculations give, from (2),

$$\left(\frac{\sigma_y}{y}\right)^2 = \sum_{i=1}^{N} k_i^2 \left(\frac{\sigma_{x_i}}{x_i}\right)^2 \qquad (2')$$

If for some i both $k_i > 1$ and σ_{x_i} / x_i are considerably larger than all other components, this particular component, and its shape of distribution will be decisive for the whole situation.

Formula (1) is valid for all cases where maximum error estimates can be made for all i-values. The use of formula (2) requires that for all i-values the errors of the x_i-s are random variables, and consequently gives the variance of the the error distribution of y.

As far as the aims and the needs have been decided upon, the choice between (1) and (2) should be uncontroversial. The recommendation in [1] and [10] to use (2) also when some of the error components are constant, but unknown within certain limits, is based on a wish to achieve harmonization, and the need, in metrology, to add uncertainty components in chains of traceability. The consequences of such a use of (2) is discussed next.

2.2 The treatment of error components which are not random variables

When the error of a measured value x_i is a random variable, changing with each experiment within a laboratory, its properties can be investgated and assessed. This type of error is called a type A uncertainty in [1] and [10] and it is the one which can be treated rigorously by (2).

In many cases parts of the errors are constant within the laboratory, due to method, equipment or operator, i e they do not change for consequtive experiments in one laboratory. They may or may not be variables changing *between* laboratories. Such errors are called type B uncertainties in [1] and [10]. The inclusion and treatment of type B uncertainties as random variables in (2) has been, and is, causing much debate, and it has to be commented upon also in this note.

Each measurement of a variable contains several errors of type A and/or type B, i e

$$x_i = x_{ic} + \sum_j m_{ij} + \sum_k \varepsilon_{ik}$$

where x_{ic} are correct values, m_{ij} are constant within the laboratory and ε_{ik} vary randomly. If the m_{ij}-s which are constant in all laboratories are excluded, which is an à priori assumption in many measurements related to fundamental metrology, comparative experiments to find repeatability within laboratories, related to ε_{ik}, and reproducibility between laboratories, related to m_{ij} and ε_{ik}, may be performed according to e g ISO 5725 [6] to obtain assessments of the effects on y from the error components.

In the following it is presumed, for simplicity in a principal reasoning, that only one ε- and one m-component, ε_i and m_i, have to be considered for each i. For a certain test the total error can then be written as

$$dy = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}(\varepsilon_i + m_i) = \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right)\varepsilon_i + \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right) m_i$$

Here, m_i are constant, while ε_i vary as random variables. A series of experiments in two laboratories will then yield results according to Figure 2. If a number of experiments is performed in one laboratory, the results will vary within the range of the distribution for $\sum(\partial f/\partial x_i)\varepsilon_i$. The variance of this distribution can be assessed by the aid of (2), and the coverage can be given by e g

$$\pm\sum\left(\frac{\partial f}{\partial x_i}\right) m_i + 2\left(\sum\left(\frac{\partial f}{\partial x_i}\right)^2 s_{x_i}^2\right)^{1/2}$$

If the m_i-s are the dominating terms and if the estimate of their sum is underestimated there is a risk that the results of that laboratory are given with too low levels of uncertainty. Particularly when N is small, say 2 or 3, the risk is particularly big. In the Appendix this effect is further demonstrated by an example.

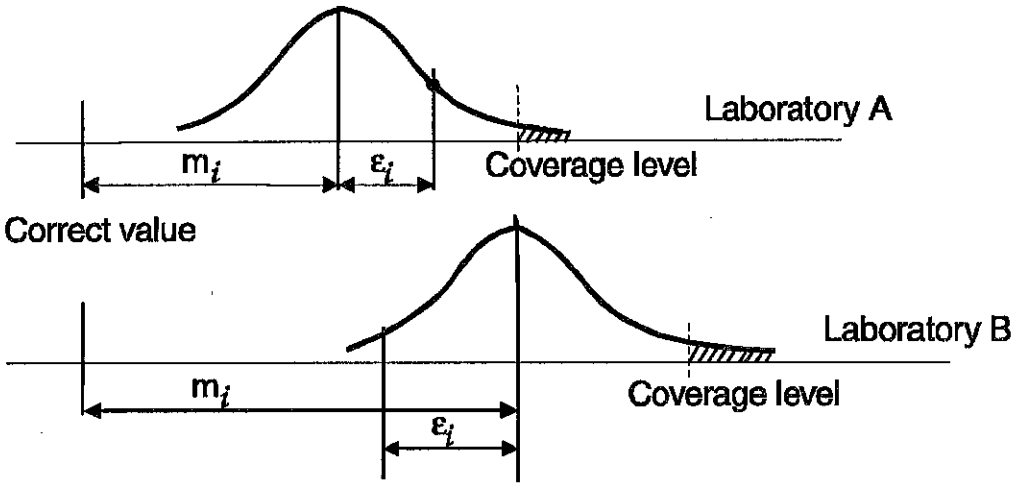

Figure 2. Results of experiments in two laboratories.

It should be noted here, as was also made in [1], that estimates of limits for ε- and m-errors (type A and type B uncertainties with the nomenclature of [1] and [10]) are about equally easy, or difficult, to assess in practise. For type A components standard deviation estimates have usually to be found from sample sizes of around ten, giving estimates which in themselves have uncertainties with relative standard deviations of about 10-20 %. For m-components, there is often experience available enabling the estimation of limits of m with a comparable confidence. The use of such estimates to make seemingly acurate standard uncertainty estimates for constant error components, of type B, is discussed in the sequel.

Obviously a safe, but strongly exaggerating way to estimate the influence of the m_i-components would be to add the maximum limit estimates by the aid of (1). This would function well compared to using (2) and standard uncertainty estimates of type B uncertainties, when N is small, as was mentioned above. Then no result in any laboratory would deviate more from the correct value than the stated limits.

The risk to come even close to these limits is, however, very small when N grows, and particularly for metrology purposes other ways have been explored. A first step of motivation may be to assume a situation where there is only a number of type B errors, and that they are all equal and of the same magnitude, Δ, but that there is an even chance that they have plus and minus signs. Then for N=1,2 and 3 the probabilities for the combined error will be as depicted in Figure 3. Even from this extreme example it is clear that when more components are added the probability will be very small for the maximum possible error to occur.

It has been proposed [1] to extend this kind of reasoning to cases where each constant m_i may be considered as being situated anywhere in a continuous interval, and not only at the end points. Within this interval it is probable that the m_i-values are different in different laboratories, except when they represent universal errors. In such cases there is a probability that they have values other than the estimated limits, and that different universal m_i-s are situated at different positions in their respective intervals. Even if these distributions do of course not exist, the belief in the probabilities is thought to be possible to quantify as "standard uncertainties" expressing, more precisely, the belief in finding an actual m_i-value, if it were possible to find, within an interval of plus and minus one "standard uncertainty" in two cases out of three. This approach does not work in cases of extreme distributions like the ones in Figure 3. In cases where only a limiting interval for an m_i-value can be assessed, it is recommended in [1] to assume a rectangular distribution, which is motivated by the assumptions related below. This gives the "standard uncertainty" as $1/\sqrt{3}$ times the limiting interval. according to the example above.

Using Bayesian statistics, which is a rigorous, but disputable, formulation of the above reasoning, and assuming the principle of maximum entropy (equal probability for all possible values) Weise and Wöger [7] have shown that "standard uncertainties" of type B error components should be estimated and combined just like standard deviation estimates of type A errors, with an assumption of rectangular distributions.This approach, and these assumptions, which are discussed further in the Appendix, are the ones which form the basis of the ISO/TAG4/WG3 recommendations [1], and the WECC document 19 [10].

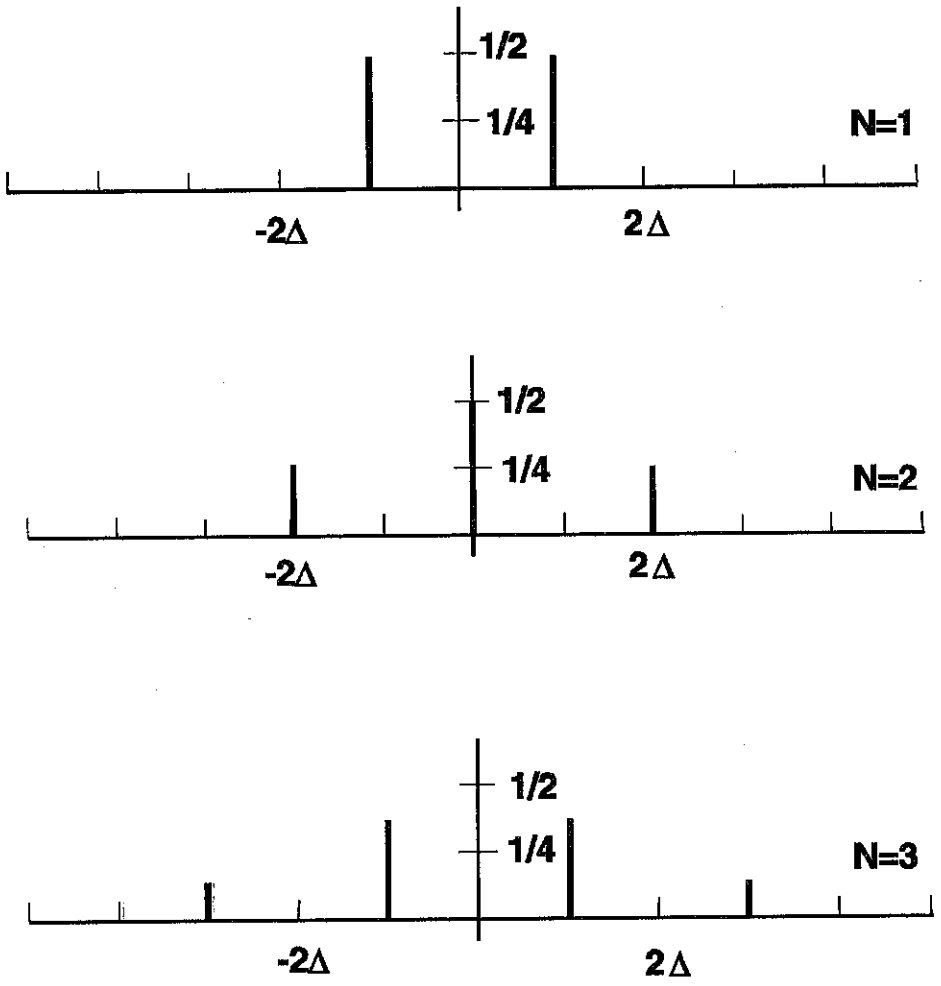

Figure 3. Resulting error distributions for sums of variables with equal, extreme error distributions, as a function of the number of variables.

3 Applicability in testing

The suggested procedure in [1] appears to work well in cases where it has been possible to test [3], with a considerable number of error components of similar sizes. This is compliant with the assumptions related above. So, although the suggested procedure is limited, due to the model assumptions, it has some value through its capacity of harmonization, and it can produce reasonable results if used with care.

Presently, there are, however, differing interpretations in national codes, e g [3], [8] and [9], with modifications in relationship to [1]. In e.g. [3] the important case with one or two dominant systematic uncertainties, which is very common in testing, has been taken care of explicity.

Hence, to testing laboratories the difficulty is not principal but related to the way the procedure is implemented and imposed in testing. For instance, the document [1] is too long to be a practical working document, around 100 pages, which should not be necessary, and so filled with subjective justifications that strong suspicions are aroused among straight-forward practicians.

Further, the testing situation is governed by a few natural and important conditions which are not compliant with the recommended procedure. Testing consists of measuring a few values, to be combined by a formula into an end result. This result is normally used for comparison with a criterion for approval. The measurements are made by instruments which are periodically checked to be within certain error limits. Hence, in the testing situation

- there is a few, two to four, systematic uncertainties, consisting of stated limits from periodic checks ("the pressure gauge has a relative error of less than 1 % between 20 % and 100 % of its measuring range"),

- there are generally no statistical analyses made, and hence type A uncertainties occur sparsely or are very small in relationship to the others,

- one or two of the uncertainties are dominant, and the other ones are one or more orders of magnitude smaller,

- one knows from experience that the "pseudo-probability" of the uncertainty may well be similar to that shown in Figure 4. For e.g. load cells, the manufacturer picks out the best ones for high precision classification. This means that the relative error of a "1 % cell" is most probably larger than 0,5 %, since those with a smaller error have been sold as "0,5 % cells",

- the customer wants a safe error limit, rather than a small one, and he does not want to risk to rely systematically on one laboratory, which may be frequently out of its stated uncertainty due to the chance of having underestimated the systematic uncertainty.

All these factors lead to the conclusion that the statement of uncertainty of a test result, although the result on the level of one significant digit may not differ too much from the one obtained from [1] or [10], should be evaluated in a much more robust way, e.g. according to [3], or even with formula (1).

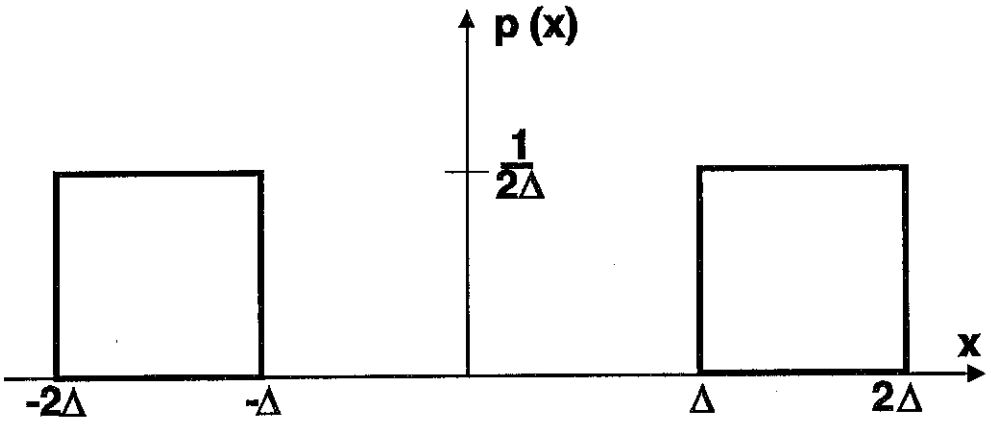

Figure 4. Possible distribution for the error of a commercial measuring instrument.

Further, the result should be expressed in a way that is not a standard uncertainty, open to later manipulation, but with limits which are natural for the customer to believe in as "practically safe". The various practises in giving uncertainties as 1s values in parentheses is totally unacceptable in testing services, regarding the structure of customers.

One additional ambiguity is the unrealistic accuracy of the recommended standard uncertainty estimates. In reality the reliance on rectangular distributions, "two times out of three bettings" etc must be considered to give "standard uncertainties to the standard uncertainties", of the order of 20-40 %, and a correspondingly small number of equivalent or effective degrees of freedom. If a small number of dominant uncertainties with considerable built-in uncertainties are combined, the result will also have a significant uncertainty. Hence, the reliance on a 95 %-level for a k-factor of 2 is not to be taken literally. If, for example, the combined standard uncertainty, s, is 20 % too large the coverage factor for 2s is not 95 % but 99 %, and if it is 20 % too small the coverage factor of 2s is 90 %. If a value, x, is stated, according to the recommendation as e.g.

$x = 23.45 \pm 0.76$

with 0.76 expressing a 2s uncertainty on the 95 % level, based on type B estimates, this is thought to be in most cases misleading. At most, the result should be possible to state as

$x = 23.5 \pm 0.8$

"with only a few percent risk of having the true value outside this interval". In [3], where a 99 % level has been chosen, this has been considered, and it is stated that the confidence level is such that not more than one or two estimates of one hundred should be in excess of the given limits.

The way of expressing the uncertainty is one example of the conditions of quality in testing services, namely the need for a user-adjusted quality concept, which may include even subjective elements. It is not only the obvious condition of a correct result which has to be fulfilled. It is imperative that the testing assignments

- are performed as efficiently as possible, i.e. quickly and at a low cost

- are performed with a high service level. Among other things this means that the customer shall have a report which can be easily understood and used in each specific industrial area, considering its tradition, level of standardisation and other needs. This includes a relevant statement of uncertainty.

With this in mind it must be stated that it is bad quality, and contradicting the general aim of quality assurance by accreditation, to impose, however harmonized it may be, a way of expressing results and uncertainties which risks being misunderstood or considered difficult and out of the real need.

In testing, the accuracy of a result can be estimated and expressed in principally different ways which naturally differ from what has been discussed above.

There are test methods, where the result is a "go/no go" result and not quantitative, but depending on measured variables. One example is fire testing, where the input parameter is the furnace temperature as a function of time, and the test result is the capability of a building element to withstand the fire for a certain amount of time. The temperature profile is allowed to lie between two limiting curves, and due to the varying geometries of furnaces of different laboratories etc the reproducibility is accepted to be low. Another example is the testing of safety belts for cars, where input variables are the impact velocity of the sledge, the properties of a standardized dummy (mannequin), and the retardation profile, which is, again, to lie between two limiting curves. The output is, of course, that the belt shall withstand the test without "visible" damage.

For this type of test, such transducers, amplifiers and readout units as those for temperature, velocity and acceleration shall, of course, be calibrated, or controlled, to standardized classes of accuracy, but it is totally irrelevant to make analyses of uncertainty, since the dispersion of properties of test specimens and of the test conditions are much larger than the measurement uncertainty.

There are also test methods specifying in detail the measurements to be made and hence also specifying the suitable way to express a significant result. One example of such a method is the ASTM E 399 [11] for fracture mechanics testing. Here, a test specimen is to be manufactured, with accurately prescribed dimensions. The specimen is provided with a fatigue crack, and then loaded to fracture. The accuracy of the load measurement, and the way to measure the crack length of the broken specimen are prescribed, as is the accuracy of a function f(a/w) necessary for evaluation of a parameter

$$K_{IC} = \frac{P_{crit} \cdot \sqrt{a}}{t \cdot w} f(a/w)$$

where K_{IC} is the fracture toughness, P_{crit} the fracture load, a the crack length and t and w geometric measures of the specimen. With such a detailed procedure, which is not uncommon in standards, the way to express the resulting uncertaintity is more or less specifying itself, and also explicitly stated in the standard. In the case of fracture toughness, and many others, the resulting quantity, here K_{IC} is a material property which is a random variable in itself, with a standard deviation which is considerably larger than the uncertainty of the testing procedure.

A requirement on an accredited laboratory to have as well documented methods as special routines, e.g. according to [1], for calculating and expressing uncertainties in reports are hence in most cases redundant. Recommendations like [1] should be imposed on, or rather disseminated to, those who produce standardized test methods.

4 Comparison of test results with requirements

The uncertainty has to be considered when a test result is to be compared with a required level for approval. The situation may be illustrated by Figure 5.

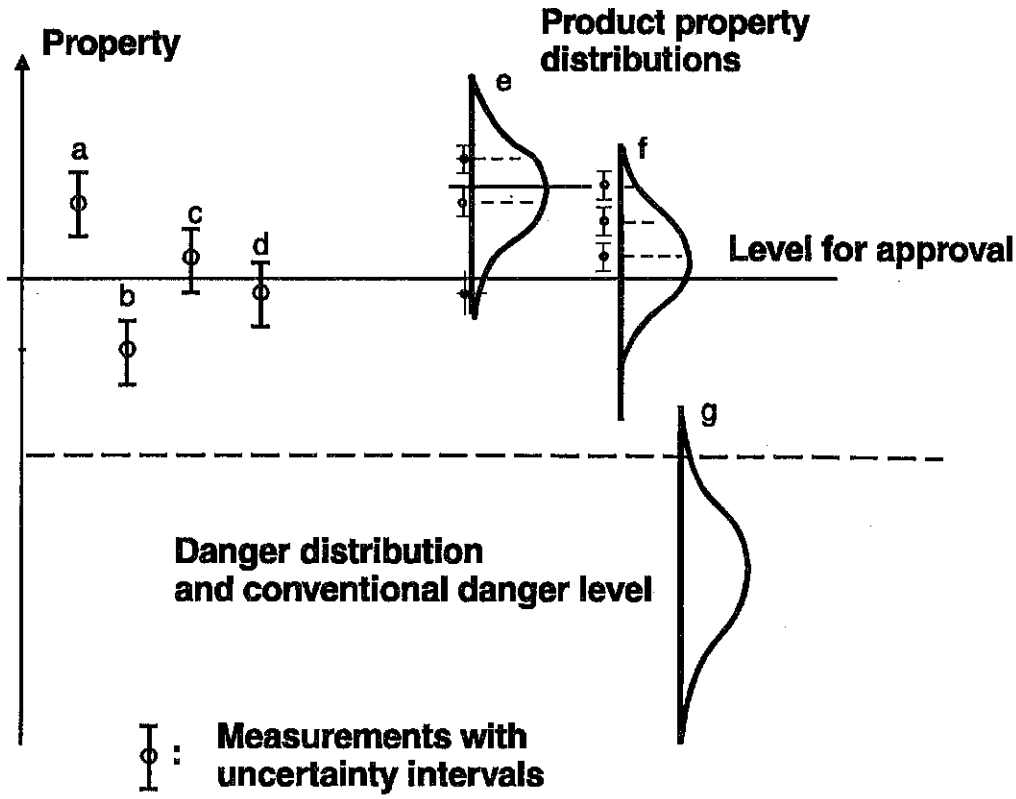

Figure 5. Relationships to consider when test results are used for safety assessments.

Two levels are shown. The solid line represents the level of the standardized requirement. This level is often set to give an "absolute safety". The dashed line denotes a level where a failure is realistic to expect, or has been experienced to occur. In terms of the failure distribution g) this level is often of the order of 3-4 standard deviations above the mean value, i e at a risk level of 10^{-4} - 10^{-3}.

One of the problems frequently discussed in connection with estimations of uncertainty is that of the acceptance of a single result illustrated in a) - d) of Figure 5. In a) and b) the result is obviously to be approved and disapproved, respectively. In c) and d) the situation is disputable, but must be resolved in practise. It has been argued that the customer should be given the "benefit of the doubt". Against this may be said that declaration of unduly large uncertainties would constitute an instrument of competition for the laboratories, which would contradict the general efforts to increase quality through smaller uncertainties. On the other hand it has also been said that laboratories with less knowledge tend to neglect uncertainties, and hence to declare too small uncertainties. One solution would be to agree, through reproducibility investigations on a reasonable "credit to the customers" for each type of testing. This is normally used implicitly today in many technical areas.

In Figure 5 e) and f) a perhaps even more important case is illustrated. Three, or another number, samples of a product are tested with a test having a given uncertainty. What decision should be taken as a function of the values of test results (and uncertainties) as related to the level of approval? In many cases of product testing the uncertainty is much smaller than the width of the distribution of the property of the product. Then it is customary to approve the product if all mean values are above the norm, to require re-testing if one value is below the norm, and to reject the product if two or more results are below the norm, irrespective of the absolute positions of the values. In some cases there are no measured results; the test is a compliance/non-compliance test.

The uncertainty intervals depicted in a) - f) of Figure 5 are those estimated by the procedures in the sections above. They could have approximate coverage factors of 95-99 %, and they should, much more than is usual, be used either to estimate the distribution f), for a probabilistic risk assessment together with the danger distribution g), or for a more direct comparison with this distribution.

In practise, these problems can be simplified by rather evident assessments of the order of magnitude of the parameters involved. In many cases this is implicit in the tradition of various areas of technology.

The relationships to consider are the ones between

- the gap between the two levels, and the coverage of the "danger level", the dashed level in Figure 5, which implies knowledge of the width of the distribution g);

- the width of the uncertainty, and its approximate coverage factor;

- the width and other properties of the distribution f) of the material or product.

Of course, the situation which is most easy to handle is the one when the uncertainty interval can, with simple means, be shown to be much smaller than both the width of f) and the gap between the levels. This is very usual in product testing. The discussion about whether the cases c) or d) should be approved or not is not

resolved by reduction of the uncertainty interval or exact determination of the coverage factor. That this should be the case is sometimes used as an argument to use recomendations like [1]. The important thing is to make an estimate which can be related to the other parameters. Only when they are of the same magnitude it is of importance to make a detailed assessment of the uncertainty.

So, one conclusion of this section is that the question of how to assess and express uncertainty can not be discussed, or regulated, in absolute terms. It should be considered in relationship to the properties tested and to the total risk assessment procedure. This sounds self-evident but obviously is not.

A second conclusion is that determination of uncertainties and risk assessment procedures should be integrated in the standardization work, in order to cover the whole problem in a consistent way. Consequently, the standardization committees, in addition to laboratories, should be a main target group for information and influence.

5 Conclusions and recommendations

The inclusion of the requirement to produce "error budgets", or analyses of uncertainty, in accredited laboratories according to the ISO/TAG 4/WG 3 recommendation, or the WELAC document 19 has aroused considerable opposition and confusion in the testing society. The objections have not been very well structured or expressed. At the same time it is of utmost importance to clarify the situation, since accreditation is spreading quickly as a very expensive, however necessary instrument for testing laboratories to gain access to the market

In this note an effort has been made to sort things out from the perspective of a test engineer, with the present situation as a background and with the aim to contribute to a constructive and conclusive discussion in the accreditation **and** testing societies. Up to this point the development has been governed by metrology experts associated to the accreditation bodies and with strong reference to chemical analysis.

Regarding the background of the theory of the present recommendation the following has been found. The method to treat all uncertainties in the same way, and to add them like variances, is not meaningful to criticise as such from the testing society. The approach is attractive in that uncertainties can be propagated in traceability chains, and in that it offers a unified approach. Theoretically it is built on assumptions leading to a statistical model. This is necessary and the assumptions can always be disputed. The adoption of the assumptions and the model should be decided by their ability to yield realistic results. In this respect, the recommended method, with various modifications, has proved to be successful, under conditions consistent with the assumptions, i e when there is a large number of uncertainty components, and when these are of a comparable magnitude. When there are only a few significant components, and when one or two is dominant, which is usual in testing, the approach does not work well and has to be modified as in [3].

The implied accuracy of the uncertainty determination, and the recommended way to express it are thought to be misleading. The uncertainties of uncertainty estimates are so large in practical cases, that the 95 %, or 99 %, statements are most uncertain. The way to express the basic result, on the 60 % level as a standard uncertainty, can be very confusing in quite normal contacts between customers and laboratories. To express uncertainties with more than one significant digit, and meaning "almost sure uncertainty limits", can not be recommended in testing.

In testing laboratories, and especially so in connection with accredited testing, the testing work, and reporting is performed according to documented methods. As a rule the methods contain specifications which imply the resulting accuracy of the test result, often they also specify how to express the result. These specifications are agreed between interested parties in industry and authorities, often in committees of international standardization. It is therefore in many cases redundant to imply additional requirements on the laboratory level. The needs and desires for a harmonized approach should result in recommendations to standardization bodies in the first hand.

Uncertainty estimates should not be prescribed in absolute terms or even in a uniform manner, but be related to the safety levels and the properties of the tested product. This integration should be handled in the standardization process or during the development of the test method.

In the present situation it is recommended that

- the testing laboratory society in its discussions with accreditation organizations, through e g EUROLAB, tries to explain the importance of customer adjustment to QA in testing laboratories, and recommends that the standardization bodies, in the first hand take on a unified approach to express uncertainty in test results according to the needs.

- the laboratories are very careful when they document methods and procedures to be accredited, so that the needs of customers are not abused, and that they make emerging anomalies known, so that experience can be used and included in continued re-assessments of the state of the art.

Appendix

Consider the case where the testing procedure is such that the result includes one type A uncertainty which is rectangular in [-1,1] and one type B uncertainty which is also, with Bayesian statistics, rectangular in [-1,1]. This means that in an ensemble of laboratories the probability to find a certain magnitude of the type B component in one laboratory is rectangular in [-1,1].

By the use of the recommendation [1], one of the laboratories in the ensemble would get the uncertainty as

$$s = \sqrt{s_A^2 + s_B^2} = \sqrt{\left(1/\sqrt{3}\right)^2 + \left(1/\sqrt{3}\right)^2} = \sqrt{2/3}$$

and the 2s limit for 95% coverage as

$$2s = 2\sqrt{2/3} = 1.63$$

In this example a coverage is then obtained of $1-((2-1.63)/2)^2 = 0.966$ from the standpoint of the laboratory and the recommendation (see also Figure 2).

However, for the customer this is true only if for each test he chooses randomly in the ensemble of laboartories. In the practical situation, where ha has to rely on one laboratory, verified on the same level as all the others, the situation changes. For some unlucky customers, using the laboratories where the type B error happens to be maximum, the coverage is reduced to (2-1.63)/2=0.82, i e in 18 % of the tests the value given is outside the stated 95% coverage area.

Although the effect is reduced when several error components are present, the example demonstrates that the principle of adding type A and B uncertainties according to [1] presumes that the customers´ choice of a laboratory is included in the sampling procedure of each test.

References

[1] ISO/TAG4/WG3-document: "Guide to the expression of uncertainty in measurements", June 1992.

[2] EUROLAB Workshop and Seminar on Measurement Uncertainty in Testing, documentation. LGAI, Barcelona, 1992.

[3] ANSI/ASME PTC 19.1-1985 (Rev. 1990) "Instrument and Apparatus - Measurement Uncertainty, part 1". ASME, New York, 1990.

[4] McNish, A.G. and Cameron, J.M. "Propagation of Error in a chain of standards", IRE Transactions on Instrumentation, 1960.

[5] Dietrich, C.F. "Uncertainty, Calibration and Probability", Adam Hilger, Bristol (1991).

[6] ISO 5725, "Precision of Test Methods - Determination of Repeatability and Reproducibility", 1981.

[7] Weise, K. and Wöger, W. "A Bayesian theory of measurement uncertainty", Meas. Sci. Technol., 3, pp 1-11, (1992).

[8] NIST Technical Note 1297, "Guidelines for evaluation and Expressing the Uncertainty of NIST measurement Results", January 1993.

[9] NATA-document "Assessment of Uncertainties of Measurement for Electrical Testing", NATA, Australia, (1992).

[10] WECC doc 19-1990, "Guidelines for the Expression of the Uncertainty of Measurements in Calibrations", WECC, (1990).

[11] ASTM E 399 "Standard Test Method for Plane Strain Fracture Toughness of Metallic Materials".