Chapter 9
The *International Vocabulary of Metrology* and the *Guide to the Expression of Uncertainty in Measurement:* Analysis, Criticism, and Recommendations

9.1 Introduction

As an independent scientific discipline, metrology needs its own terminological dictionary. Beginning from 1984, ISO has published three editions of the International Vocabulary of Metrology (VIM). The first such dictionary – "International Vocabulary of Basic and General Terms in Metrology" appeared in 1984. In 1993 came out the second edition of this document, and in 2007 the third (and current) edition. All three editions were prepared under the auspices of BIPM. The third edition has a new name – "International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM)" [1] – in order to stress the fundamental differences from the prior editions. The new VIM indeed differs significantly from the previous ones. Instead of the traditional philosophical foundations of metrology the new VIM adopts the philosophy of the "Guide to the Expression of Uncertainty in Measurement (GUM)" – the document prepared under the auspices of BIPM and published by ISO in 1995 [2]. This philosophy was named in VIM the "uncertainty approach".

The history of the uncertainty approach started from the article [19], which appeared in BIPM journal "Metrologia". The main idea of that paper was that the term "*measurement error*" appears to be used in two different senses. In one sense it expresses the difference between measurement result and the true value of the measurand. In this case, in the opinion of the authors of [19], one would use an expression such as "the error is +1%". In the other sense it reflects the uncertainty of the measurement result, where one would say "the error is ±1%". In order to distinguish the meaning of the word *error* in these two cases, the paper proposed to use the word *uncertainty* in the second case instead of the word *error*.

In fact, the terminological ambiguity the paper addressed was caused simply by an erroneous terminological shortcut. To be precise, the expression "the error

Electronic Supplementary Material The online version of this chapter (doi:10.1007/978-1-4614-6717-5_9) contains supplementary material, which is available to authorized users.

S.G. Rabinovich, *Evaluating Measurement Accuracy: A Practical Approach*,
DOI 10.1007/978-1-4614-6717-5_9, © Springer Science+Business Media New York 2013

±1%" means that the measurement error of *one measurement* is simultaneously both +1% and −1%. But this cannot be, since there can only be one result of the measurement, a fixed numerical value, which would obviously have one value of error. Thus, this expression is incorrect. In this case, in accordance with the definition of the term *error,* one should say *"the error falls within the range* ±1%" *or "the limits of error are* ±1%" If this correct expression were used, then the ambiguity pointed out in [19] would not occur.

Still, while strictly speaking unnecessary, the proposal in [19] to use the term "uncertainty" in the second case was useful because it allowed one to provide a single term for a multi-word expression, avoid the confusion with the erroneous shortcut, and divide the terms "error" and "uncertainty", which previously were used interchangeably.[1] Unfortunately, this seemingly small issue has led to consequences that were hard to foresee at the time.

The most significant consequence that appears to stem from the above terminological change was the mentioned above *uncertainty approach* introduced as a term in VIM but which represented the philosophy and a group of methods formulated in GUM. We will show in Sect. 9.4 that some of these methods are unfounded, some are arbitrary, and some are simply wrong.

At the heart of the problems with the uncertainty approach is a foundational mistake, which in essence is that GUM takes standard deviation instead of confidence interval as the indicator of measurement accuracy. GUM does not explicitly note this crucial substitution; rather it simply replaces the term "standard deviation" by "standard uncertainty" while what the term means remains exactly the same. We analyze and compare in Sect. 9.2 standard deviation and confidence interval as indicators of measurement accuracy and show that standard deviation is not a suitable indicator in most cases. Based on this analysis, we discuss shortcomings of VIM and GUM in Sects. 9.3 and 9.4, respectively. We should note that the comprehensive analysis of VIM goes outside the scope of the present book and thus we concentrate only on the terms and concepts concerning the estimation of measurement accuracy.

VIM and GUM attempt to define the vocabulary and foundations of metrology; as such both documents are vitally needed. However, given the fundamental nature of their shortcomings, fixing the current documents through corrections and additions seems impossible – they need to be re-written. All criticisms in this chapter are accompanied by proposals for addressing the identified problems.

9.2 Comparison of Standard Deviation and Confidence Interval as Measurement Accuracy Indicators

Before analyzing GUM and VIM, we need to establish precisely the meaning of the terms *standard deviation* and *confidence interval* as indicators of measurement accuracy. Both these terms are mathematical concepts used in statistics and in

[1] Indeed, the reader would note that this terminology, which separates the terms error and uncertainty, is followed in the present book.

experimental data processing in metrology. In statistics, the accuracy of an estimate of mathematical expectation is characterized by confidence interval while its efficiency by variance, i.e., the square of the standard deviation. In GUM, however, the standard deviation is used to express the accuracy of measurement. Before we analyze the appropriateness of such usage, we should make a disclaimer that we will limit ourselves to a multiple measurement free of systematic errors, whose widely accepted mathematical model is a random quantity. These concepts were discussed in Chap. 3, but let us recall a few basic notions.

We will base our discussion on the normal distribution – the most frequently used distribution in statistics and metrology. The normal distribution is specified by two parameters, mathematical expectation A and variance σ^2 although our discussion will be mostly concerned with not variance but standard deviation $\sigma = +\sqrt{\sigma^2}$. A multiple measurement represents a series of repeated measurements of the same quantity under the same conditions. These individual measurements are called observations. Under the conditions of the experiment, all observations have equal probability and considered as a sample from the general population of observations with a certain distribution – the normal distribution in our case. The task of processing the observations obtained in the course of measurement is to find the most accurate estimate of the mathematical expectation of the above distribution function and to determine the accuracy of this estimate. In metrology, the mathematical expectation of the distribution function is called the true value of the measurand. This point corresponds to the abscissa of the knee point of the normal distribution function, or the maximum point of the probability density function. As defined in Sect. 1.1, the accuracy of measurement expresses how close the result of measurement is to the true value of the measurand. In practice, instead of the positive notion "accuracy", its negative dual concept – error – is commonly used.

We now turn to comparing the accuracy indicators. Confidence interval and the methods of its construction were described in Sect. 3.6. In the practice of measurements, it is overwhelmingly constructed using Student's distribution. As shown in Sect. 3.6, Student's distribution defines interval

$$|\bar{x} - A| \leq t_q S(\bar{x}),$$

where t_q is the q th percentile point of Student's distribution, $\bar{x}$ is the mean of the observations, which also represents an estimate of the true value of the measurand, and $S(\bar{x})$ is an estimate of standard deviation of the mean of the observations.

The above interval specifies directly the limits of the measurement error. It can also be represented in a different form:

$$\left(\bar{x} - t_q S(\bar{x})\right) \leq A \leq \left(\bar{x} + t_q S(\bar{x})\right).$$

This form shows the limits of an interval that covers the true value of the measurand.

An important aspect of Student's distribution is that it depends only on integral estimates $\bar{x}$ and $S(\bar{x})$ of the parameters of the distribution. Therefore, the confidence interval built based on Student's distribution has low sensitivity to the distribution of input data and can be used with any distributions as long as they are convex and symmetrical. The percentile t_q depends on the number of observations, which determines degree of freedom v, and on the confidence probability α.

The confidence interval, as we know (see Sect. 3.6), with probability α covers the true value of the measurand. This means that if we repeat the same measurement under the same conditions many times, in α fraction of cases the confidence intervals will overlap, and their common part will cover the true value. Thus, the uncertainty expressed as a confidence interval correctly reflects the accuracy of the estimate of the true value. Since confidence probability is known a-priori as it is set by the experimenter to suit specific objectives of the measurement, confidence interval is an unambiguous and precise (in the probabilistic sense) indicator of measurement accuracy.

Now let us consider standard deviation. As a parameter of the distribution function, standard deviation characterizes how widely the random variable is spread out relative to the mathematical expectation (i.e., the true value in our context). But the *estimate* of standard deviation $S(\bar{x})$ characterizes how widely the observations are spread out relative to their mean, which – while taken as the measurement result – differs from the mathematical expectation. Thus the estimate of standard deviation does not take into account the distinction between the mean of the observations and the mathematical expectation (or, equivalently, true value). But as we saw earlier, this difference can be as large as half the confidence interval, or $t_q S(\bar{x})$. In fact, this represents the measurement uncertainty itself – therefore, one cannot neglect it or not account for it in estimating measurement accuracy. In summary, the estimate of standard deviation characterizes how spread-out the observations are and hence the repeatability of the measurement but not its accuracy.

One might say that the only issue here is how to define measurement accuracy. If we just defined the accuracy as the standard deviation of the mean, the above problem would disappear. However, the estimate of this standard deviation in no way reflects how close the mean is to the mathematical expectation, i.e., to the true value of the measurand. Thus, standard deviation cannot reflect the accuracy of measurement and hence cannot be used as the definition of accuracy.

Another drawback of the estimate of standard deviation as measurement accuracy indicator is that it does not offer any information on how likely a repeated measurement of the same measurand with the same number of observations would produce the same accuracy. Indeed, having a group of observations, we can construct an interval that would exclude from the group some extreme observations as outliers. We can compute the percentage of the remaining observations in that interval and assume that these are trustworthy. However, we have no basis to specify any probability of having the same percentage of trusted observations in subsequent measurements. Thus standard deviation as a measurement accuracy indicator carries no information about the probability to obtain the same accuracy of a future measurement of the same quantity.

A further drawback of standard deviation is that the reliability[2] of its estimate strongly depends on the number of observations obtained in the course of the measurement, and this dependency is not reflected in the overall accuracy indicator (when the standard deviation is used in this role). Indeed, Sect. 3.7 showed that with the normal distribution of observations, this dependency is expressed as

$$\varphi = s[S(\bar{x})]/\sigma(\bar{x}) = 1/\sqrt{2(n-1)},$$

where $S(\bar{x})$ is the estimate of standard deviation of the sample mean of size $n, s[S(\bar{x})]$ is the estimate of the standard deviation of $S(\bar{x})$, and $\sigma(\bar{x})$ is the true standard deviation of the sample mean. One can get a sense for the accuracy reliability of typical estimates of standard deviation from the example in Sect. 3.5, which considered a sample of $n = 10$ observations from normal distribution that had $S(\bar{x}) = \tilde{\sigma}_{\bar{x}} = 1.2 \times 10^{-5}$. In this case, we have $\varphi \cong 24\%$. Obviously, the obtained estimate is inaccurate. To see possible limits of its inaccuracy, we obtained in Sect. 3.5 the confidence interval covering $\sigma(\bar{x})$. In our example, for confidence probability 90%, this interval was $[0.88,\ 2.0] \times 10^{-5}$. Thus, in repeated samples of the same size from the same distribution, the estimates of the standard deviation may deviate, with rather high probability, almost by the factor of two from the estimate that would be reported as the accuracy of the measurement.

This example had 10 observations. In practice, measurements often have even fewer observations, resulting to even higher inaccuracy of the estimate of the standard deviation of the mean. Only in measurements where the number of observations reaches a 100, can the accuracy of estimate of the standard deviation be considered sufficient in practice. But such measurements are relatively uncommon.

The above discussion leads to a conclusion that standard deviation is not a suitable indicator of accuracy of a measurement. In fact, the arguments used in this discussion are rather obvious, and one may wonder why in practice standard deviation is still used in this role. The answer lies in the classical theory of indirect measurements.

As we know (see Chap. 5), indirect measurements can be with independent or dependent arguments. In the former case, the classical theory allows one to compute both the standard deviation and confidence interval of the measurement result. However, in the latter case, the classical theory can compute the standard deviation – if the computation accounts for correlation coefficients between the arguments – but not the confidence interval because we do not know how to find the degree of freedom that accounts for correlation coefficients.

Multiple indirect measurements with dependent arguments are often used in scientific experiments. One of typical problems encountered in these experiments is to understand conditions under which the parameter of interest to the experimenter is sufficiently stable, so that it can be measured in other laboratories. In this case, the scientist is primarily interested in repeatability rather than accuracy of the measurement. Furthermore, measurements in these experiments usually include

[2] Recall from Sect. 3.6 that by reliability of an estimate we mean an indication of how much different estimates, obtained from different samples of observations, can differ from each other.

many observations and hence the estimate of standard deviation becomes sufficiently accurate. Thus, the use of standard deviation as measurement accuracy indicator by scientists in this case arose from the combination of several factors, namely, the primary need for estimating repeatability rather than accuracy of measurement, the sufficient accuracy of the standard deviation estimate in these types of measurements, and the inability to compute confidence interval.

However, in the 70s of the last century, a new method was found – the method of reduction, which allows one to construct confidence interval in dependent indirect measurements. This method is described in detail in Sect. 5.6, where we also highlight its other benefits, namely that it removes the need in correlation coefficients and that it is in principle more accurate than the traditional method in estimating standard deviation. Thus, one can now construct confidence intervals for all types of indirect measurements, which eliminates the last niche to use standard deviation as a measurement accuracy indicator.

A general conclusion from this discussion is that the question on whether confidence interval or standard deviation is a better indicator for measurement accuracy is in essence improperly stated. Indeed, standard deviation only characterizes how spread-out – and not how accurate – the repeated observations are. The only indicator of measurement accuracy is confidence interval.

9.3 Critique of the *International Vocabulary of Metrology*

The “International Vocabulary of Metrology – Basic and General Concepts and Associated Terms” (VIM) was prepared by Working Group 2 of JCGM and published by ISO/IEC in 2007. The Foreword of VIM states that this document replaces all previously published editions of the International Vocabulary of Basic and General Terms in Metrology. The new VIM differs significantly from the previous ones and reconsiders the definitions of many terms.

In Introduction it brings several arguments in justifying this change. These arguments include the detailed explanation of the reasons to move from the traditional approach to experimental data processing to a group of methods that VIM named the *Uncertainty Approach,* and which were established in GUM [2]. However, these arguments are not compelling, and often invalid.

For example, the Clause 0.1 General of VIM states that the traditional approach[3] cannot solve the problem of combining systematic and random errors: “The

[3] Note that while we in this book have also transitioned from traditional theory of measurements to what we call physical theory, our change retains the concepts of true value of a measurand or measurement error, and does not substitute standard deviation for measurement accuracy indicator. Our physical theory avoids the use of Taylor series and thus increased accuracy of estimating measurement uncertainty, as well as solves two fundamental problems in theory of measurement: constructing confidence intervals for indirect measurements and universal method for combining systematic and random components of measurement error. The uncertainty approach introduced by VIM and GUM does not solve new problems, and – as we show in this chapter – produced some incorrect methods.

objective of measurement in the Error Approach is to determine an estimate of the true value that is as close as possible to that single true value. The deviation from the true value is composed of random and systematic errors. The two kinds of errors, assumed to be always distinguishable, have to be treated differently. No rule can be derived on how they combine to form the total error of any given measurement result, usually taken as the estimate" (page vii in [1]).

However, the above statement is mistaken. Long before the uncertainty approach was introduced in VIM, or the methods underlying this approach were established in GUM, the fact that the two types of errors, while estimated differently, must in the end be combined was commonly accepted and several concrete methods of solving this task were proposed. The most widely used methods are described, for example, in Sect. 4.11 of the present book but they originated in my own work and the work of my group long ago – a universal (i.e., suitable for an arbitrary confidence probability) solution to this problem was developed back in 1970 [48], analyzed in 1978 [46], and later included in the standard "Metrology. Basic Terms and Definitions" [12] (in Russian, clause 9.30, formulas 9.9). In English, well-known methods suitable for confidence probability 0.95 were described in book [14], first published in 1973, and in standard [6], which included formulas for probabilities 0.95 and 0.99. The universal method was described, analyzed and compared with other methods in my book *Measurement Errors: Theory and Practice,* published in 1993, as well as in the subsequent editions of that book [44].

Furthermore, as we show in Sect. 9.4.4 below, GUM and the uncertainty approach do not actually solve the problem of combining systematic and random errors. Indeed, we will see that the method for combining these error components formulated in GUM (which called them uncertainty A and uncertainty B), is incorrect.

Because of the prominence given to the term uncertainty in GUM and then VIM, this term has acquired special significance. Thus, it is interesting to trace how its interpretation changed over time. In the first edition of VIM (1984), the term *uncertainty* (clause 3.09) was defined as follows: *An estimate characterizing the range of values within which the true value of a measurand lies.* In other words, this term meant an interval that characterized the accuracy of measurement. The second edition (1993) has changed this definition to read: *Parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.* This definition, unlike the previous one, is rather vague. It is unclear what "values that could reasonably be attributed to the measurand" may mean. Moreover, the note to the definition says that uncertainty can be standard deviation or half of the listed interval with a stated level of confidence (clause 3.9), thus associated two distinct meanings to the same term – obviously an undesirable outcome. The new VIM (2007) has retained the definition of uncertainty (clause 2.26) from the second edition, only in the note the expression "stated level of confidence" is replaced by "stated coverage probability".

Let us now consider specific terms in the latest VIM that have a particular bearing on the present book – *measurement result*, *true value, error,* and *uncertainty.*

The clause 2.9 of the VIM defines *measurement result* as a "*set of quantity values being attributed to a measurand together with other available relevant information*."

Note 1 clarifies that " this may be expressed in the form of a probability density function (PDF)." According to this definition, a set of observations obtained from the multiple measurements represents the result of the measurement. However, as known, the goal of the measurement is always a single estimate for the measured quantity obtained from the analysis of this set, often augmented with an indication of its accuracy, but not the set itself. Having a single estimate allows one to use measurement results in mathematical formulas expressing natural laws. One cannot replace values with distribution functions in these calculations. Therefore, this definition of *measurement result* is not productive, and the traditional definition should be retained, which is that the *measurement result is a value attributed to a measurand, obtained by measurement.*

The definition of *true value* (VIM, clause 2.11) says that it is the "*quantity value consistent with the definition of* a *quantity*." However, the value assigned to the measurand as the result of the measurement is also consistent with the definition of the quantity – otherwise, it would be useless. In other words, this definition of the true value suggests that the measured value of the quantity and its true value are the same. In contrast, the established meaning of the term true value is that it is an abstract, unreachable, property of the measurand. Without this established meaning, it is impossible to define the accuracy of a measurement. Therefore, the following definition, given in this book, is advisable: *true value – the value of a quantity that, were it known, would ideally reflect the property of an object with respect to the purpose of the measurement.* Note: as any ideal, the true value is impossible to find.

The definition of *true value* in VIM has three notes, two of which require a discussion. Note 1 states: "In the Error Approach ... a true quantity value is considered unique and, in practice, unknowable. The Uncertainty Approach is to recognize that, owing to the inherently incomplete amount of detail in the definition of a quantity, there is not a single true quantity value but rather a set of true quantity values consistent with the definition. However, this set of values, in principle and in practice, unknowable. Other approaches dispense altogether with the concept of true quantity value and rely on the concept of metrological compatibility of measurement results for assessing their validity." Two aspects of this note are objectionable.

First, we would like to disagree with the notion that the incomplete amount of detail in the definition of a quantity entails a set of true values rather than a single true value for the quantity. It is well known that the goal of any measurement is to obtain a numeric value that reflects the measured quantity. Measurement results realize this goal. It is this aspect of measurements that allows us to apply mathematics to natural sciences, and it is only possible if every measured quantity has a single true value. Indeed, if we assumed that the measured quantity had multiple true values, it would be impossible to associate it with a single number and use it in subsequent mathematical formulas. Although a measurement result often includes an indication of its accuracy, and this indication is often expressed as an interval, any measurement result still assigns a value (usually taken as the most likely value within the interval) to the measurand.

The concept of the true value of a measured quantity is considered in detail in Sect. 1.4 of the present book. That section also considers the example of the measurement of the thickness of a sheet of a material, which is presented in GUM (Sects. D.3.2 and D.3.4) to motivate the idea of a measured quantity having a set of true values. We explained that when the thickness of the sheet is different in different places and one must reflect these different thickness values by measuring the thickness in different places, we have in fact several distinct measurements, one in each place of the sheet. Each given point of the sheet has its own true value of thickness and will have its own measurement result. There is no single measurement result here, and the set of true values does not have to do with individual measurements of the sheet thickness in different points. Thus, this example does not show the need or the usefulness of having a set of true values for one measured quantity.

Regarding the inherently incomplete amount of detail reflected in the definition of the quantity, the definition of the quantity must only reflect the property that is of interest to the experimenter. The lack of detail in the definition of the quantity is not a reason for introducing a set of true values for the quantity.

Second, we question the usefulness of distinguishing two approaches to estimation of the accuracy of measurements. Defining new approaches is beneficial only if they enable solutions to new problems. However, the VIM does not present any new problem solved by the Uncertainty Approach with its set of true values for a quantity. Thus, its introduction appears unwarranted. Further, the sentence following the note in question mentions additional approaches but leaves it unclear what these approaches are. From the above considerations, we conclude that the notion of a "set of true values" must be removed from VIM.

Note 3 also raises objections. It represents an attempt to justify an erroneous concept of the "Guide to the Expression of Uncertainty in Measurement" [2] of the equivalency between the true value and the value of the measured quantity. However, the true value is an unreachable ideal concept, while the value of a measured quantity is a measurement result. Thus, the two cannot be equivalent no matter the accuracy of the measurement in question. We return to this issue in more detail in Sect. 9.4.

These considerations lead to a conclusion that Notes 1 and 3 should be removed from VIM.

Clause 2.16 defines *measurement error* as "*measured quantity value minus a reference quantity value*." Unfortunately, the above sentence cannot be considered a definition because it does not explain the meaning of the term. Instead it attempts to provide an algorithm for its calculation but this algorithm is unrealistic: it follows from clause 5.18 that the reference quantity value in measurements refers to the true value, which is always unknown. Furthermore, this definition narrows the meaning of the term since it only covers the absolute error, leaving a commonly used relative error aside.

I consider *measurement error* to be properly defined as *a deviation of the result of measurement from the true value of the measurand.* This definition is not algorithmic and makes it clear that just like the true value, measurement error is impossible to obtain. In fact, the above consideration warrants the following note to this definition: Because the true value is always unknown, the error of measurement is estimated indirectly, by analyzing the accuracy of measuring instruments, measurement conditions, and the obtained measurement data. In single measurements under reference condition of the instruments involved, the measurement error is determined by the limits of the permissible error of the instruments and is expressed in the form of limits of measurement error. In multiple measurements, the measurement inaccuracy is usually estimated using statistical methods, in which case the measurement inaccuracy is characterized using the concept of measurement uncertainty rather than the limits of error. The proposed definition of the term "error" is close to that given in [10].

The definition of *uncertainty* in VIM (clause 2.26) is provided with a note saying that uncertainty "may be, for example, a standard deviation called standard measurement uncertainty (or a specified multiple of it), or the half-width of an interval, having a stated coverage probability." This note creates ambiguity that is unacceptable in scientific terminology. Indeed, what is the uncertainty, a standard deviation or an interval? Giving two different meanings to one term must be avoided in a terminological dictionary.

9.4 Critique of the *Guide to the Expression of Uncertainty in Measurement*

Another important document published by ISO is the "Guide to the Expression of Uncertainty in Measurement" (GUM) [2]. The goal of GUM was to unify the methods of measurement uncertainty estimation and its presentation. The uniformity of estimation and expression of inaccuracy of measurements is a necessary condition for the economic development of every country and for international economic cooperation. Thus, GUM was enthusiastically received by the metrological community.

However, a number of shortcomings among GUM recommendations have transpired subsequently. In [16], it was noted that "the evaluation methods in the GUM are applicable only to linear or linearized models and can yield unsatisfactory results in some cases." The same article reported that to address these issues, Addition 1 to GUM had been prepared and that furthermore, Working Group 1 JCGM decided in 2006 to prepare a new edition of GUM. Other critical comments regarding GUM can be found in [32]. Our own criticism appeared in [44] and, in more detail, in [42].

Still, the recently published VIM (which we discussed in the previous section) clearly reflects GUM's influence. For example, VIM repeatedly uses the notion of a set of true values of a measured quantity, which as we showed in Sect. 9.3 is misguided. In Note 3 to clause 2.11 it makes an attempt to justify a mistaken concept from GUM about the equivalency of the true value and a value of a quantity. Apparently, past criticisms of GUM were not sufficiently convincing, and we revisit its drawbacks here.

9.4.1 Scope of GUM

GUM begins with a statement that "The Guide establishes general rules for evaluating and expressing uncertainty in measurement that can be followed at various levels of accuracy and in many fields – from shop floor to fundamental research." Unfortunately, the rest of GUM's content does not support this intended scope since it is devoted exclusively to multiple measurements. Single measurements, although being the basic type of measurements in industry, trade, quality assurance, clinical medicine, and other fields, are not even mentioned. This limited scope is a significant limitation of GUM.

9.4.2 Philosophy of GUM

The foundational premise of GUM is that the concept of true value of a measurand is not needed because it is equal to the value of this measurand. This premise is formulated explicitly in "Guide Comment to Sect. B.2.3" (page 32 of GUM) and also in Annex D (Sect. D.3.5). However, this premise is in contradiction with VIM, as well as with fundamental conventions of physics and statistics. According to VIM, clause 1.19, the value of a quantity is a number and reference together expressing the magnitude of a quantity. In other words, it is the product of a number and the unit of measurement. This value is obtained as the result of a measurement. In contrast, the true value is a purely theoretical concept and cannot be found (see clause 2.11 of the VIM). Thus, the terms "true value" and "value of a quantity" cannot be considered the same and the latter cannot replace the former.

In statistics, the terms "parameter" (true value) and "estimate of the parameter" (the obtained value of the parameter) are strictly distinguished. In physics, the equations between physical quantities would be impossible without the concept of a true value; indeed, physical equations would always be only approximately correct for obtained values of the quantities. Finally, as we will see bellow, the GUM itself needed a distinction between the true value and the value of the measurand and was forced to introduce rather awkward new terminology in its place. These considerations bring a conclusion that during the new edition of GUM it should revert to traditional philosophy.

9.4.3 Terminology of the GUM

The elimination of the term “true value” was motivated by the desire to eliminate the term “error.” Consequently, the GUM uses the term “uncertainty” in place of “error” throughout the document. The goal was to eliminate synonymia in using both terms throughout the document. This goal can be accomplished, however, without excluding the term “true value” and the corresponding concept; in fact, by defining the terms “error” and “uncertainty” precisely, we could distinguish the two clearly and at the same time not impoverish the metrological language by eliminating the term “error” but, to the contrary, enrich it by giving the two terms different meaning.

Metrology offers every prerequisite to achieve this. Indeed, the uncertainty of a measurement result is calculated usually from its components and with the help of statistical methods. In contrast, in the case of a single measurement using measurement instruments under reference conditions, the measurement inaccuracy is fully determined by the limits of error of the instrument, and statistical methods are not applicable.

Consequently, the term “uncertainty” may be used for probabilistic estimates of inaccuracy and the term “limits of error” when the inaccuracy estimates have no probabilistic interpretation. Moreover, according to VIM clause 2.26, the term “uncertainty” is associated with the result of measurement. Thus, it cannot replace the term “error” in other cases; for example, it cannot be used for components of uncertainty or to express the inaccuracy of a measuring instrument. We conclude that the total replacement of “error” with “uncertainty” is unjustified.

The GUM introduces two new terms “type A and type B evaluation of uncertainty,” defining them as methods of evaluation of uncertainty (clause 2.3.2 and 2.3.3) but using them as components of uncertainty. Indeed, clause 5.1.2 describes how to combine uncertainties type A and type B; clearly, methods cannot be combined and they are treated there as components of uncertainty in this context. Such inconsistency should be avoided in a document aiming to introduce rigorous language for others to follow. In addition, these terms are not expressive. It would be much better to use the common term “random error” instead of “type A uncertainty” and the term “rated error” (if the term “systematic error” is undesirable).

Another inconsistency in the GUM is with the terms “standard uncertainty,” “combined uncertainty,” and “expanded uncertainty.” The first two are defined as simply standard deviation and the combined standard deviation, respectively. But “expanded uncertainty” is presented as an interval. It is confusing to use the same term “uncertainty” as the basis for derived terms having drastically different meaning – a standard deviation in one case and an interval in the other.

In general, to calculate measurement uncertainty, the terms “standard deviation,” “combined standard deviation,” and “uncertainty” itself would be sufficient. The GUM introduced duplicate terms “standard uncertainty” and “combined standard uncertainty” as the terms that “are used sometimes for convenience” (clause 4.2.3). But it uses them exclusively throughout the rest of the document, creating an impression that this is the proper terminology to be used. These duplicate terms cause inconvenience in practice. For example, to follow this terminology, one has to always point out that standard uncertainty is equal to standard deviation, which is then computed using known statistical methods. As a typical example, Kacker and Jones [31] repeatedly use in their article passages the following: "According to the ISO Guide (Sect. 4.2), the type A standard uncertainty associated with z_A from classical statistics is $u(z_A) = s(z_A) = s(z)/\sqrt{m}$."

In other words, when saying "standard uncertainty," a methrologist must remember that in fact the term refers to "standard deviation." The same holds for the term "combined standard deviation."

Another terminological difficulty has to do with the concept of confidence interval. As it is known, it is the interval that, with given probability, contains the true value. Thus, it needs the concept of true value, which the GUM was trying to eliminate. In an attempt to resolve this logical gap, the GUM replaces the term "true value" with the expression "letter Y that represents the value attributed to the measurand" (clause 6.2.1 and Annex G) or "measurand Y." This proliferation of nondescriptive terms makes the terminology nonintuitive, and it is unnecessary since descriptive terms exist.

9.4.4 Evaluation of the Uncertainty in the GUM

GUM uses standard deviation as measurement uncertainty, calling it standard uncertainty. The adoption of standard deviation as accuracy indicator forms the foundation of the entire document. As discussed in Sect. 9.2, standard deviation fundamentally cannot serve as measurement accuracy indicator because its estimation is calculated relative to the mean of observations and does not reflect the offset of the mean from the true value of the measurand. Thus, standard deviation can only be an indicator of repeatability of a measurement but not of its accuracy. Thus fundamental mistake entails all other drawbacks of GUM.

The GUM contains the terms standard uncertainty, combined uncertainty, and expanded uncertainty. The first two are just different names for the standard deviation and combined standard deviation. They are computed using known formulas. But expanded uncertainty is represented as an interval. In Chap. 6 of the GUM this interval is called *coverage interval,* which is defined as "an interval about the measurement result that encompasses a large fraction p of the probability distribution of values that could reasonably be attributed to the measurand" (clause 6.1.2). GUM further describes the calculation procedure for the coverage interval using two additional new terms, *coverage probability* and *coverage factor*. However, how to find the above probability distribution, the coverage factor, and therefore the coverage interval, remains unspecified and is unknown. Changing the terminology obviously does not solve the problem of obtaining the expanded uncertainty (or confidence interval in the traditional terminology).

The root of the problem with computing the expanded uncertainty is that the GUM does not provide a method for combining systematic and random errors of a measurement result. Consequently, clause 6.3.3 recommends calculating the expanded uncertainty simply as the product of combined uncertainty and factor 2 or 3; the result is assigned, without any justification, probability 0.95 in the first case and 0.99 in the second.

Besides assigning unjustified confidence probability, the above method selects the factors 2 and 3 so that they are almost the same as percentiles t_q from Student's distribution with $v = \infty$. However, measurements often do not have large enough observations to justify the assumption that $v=\infty$. Furthermore, Student's distribution is not applicable in this case. Indeed, recall that estimate of combined variance is a sum of estimates of variance of random errors (uncertainty A according to GUM) and conditional constant errors (uncertainty B). Thus, the combined standard deviation represents the standard deviation of the sum of random and conditionally constant systematic errors. Student's distribution establishes the connection between the mean of a group of observations and the standard deviation of this mean. In the case in question, the mean is calculated using data having only random errors, while the standard deviation – the square root of the sum of the estimates of the variances of random and conditionally constant errors – reflects both random and systematic errors. Therefore, using Student's distribution in this case is incorrect.

Clauses G.3.1 and G.3.2 of Annex G offer a different method for calculating the expanded uncertainty. This method is based on the Student's distribution, which in this case is not applicable as we just argued.

Another mistake has to do with calculating the effective degree of freedom. Its essence is that the concept of "degree of freedom" is not applicable to a random quantity with fully known distribution function. For the model of systematic errors the GUM takes the uniform distribution with known limits, and this distribution cannot be assigned degree of freedom $v = 1$, or any other number.

We should note that there is a known method for computing the uncertainty of a measurement result with given confidence probability, which accounts for both systematic and random errors of the result. This method is described in [44, 46] and discussed in detail in the present book.

The forgoing discussion shows that the upcoming new edition of the GUM must extend beyond revising its philosophy and terminology and revise its recommendations for data processing as well. Such revision is possible on the basis of existing methods and traditional philosophical foundation.

The revision of the GUM should utilize the method of reduction for dependent indirect measurements. In fact, the GUM already mentions the method of reduction as a second approach (see the note on page 10 in Sect. 4.1.4), but does not discuss its advantages over the primary method recommended in the main body of the document. These advantages were pointed out in this book, and the main ones being that this method allows one to construct the confidence interval for dependent indirect measurements and that it eliminates the need for the correlation coefficient. These benefits of the method of reduction are hard to overstate.

Further, we would like to point out again that the revision of the GUM must also include methods of estimating the inaccuracy of single measurements. These methods also exist already and are discussed in this book.

The above problems with GUM's recommendations regarding the estimation of the uncertainty of a measurement result have been recognized by JCGM, and Supplement 1 to the GUM is devoted to rectifying these issues [13]. Supplement 1 addresses them through the use of the Monte Carlo method. However, as we discussed earlier, there exist more accurate and much simpler approaches. Note that being able to solve these problems without the Monte Carlo method would not obviate the need for Supplement 1 in the form of a separate recommendation devoted expressly to the Monte Carlo method, which can have its own significance in metrology (see Sect. 5.12).

9.5 Roots of the Drawbacks of GUM and VIM

Given the importance of GUM and VIM for metrology and the severity of the flaws in the current versions of these documents, it is important to think of ways to correct these problems. A necessary preliminary step would be to analyze the reasons that might have caused the problems in the current documents. This section contains the author's thoughts and speculations in this regard.

The flaws in GUM could be explained by two main reasons. The first reason is that the development of GUM was not properly organized. Indeed, BIPM, for over 130 years of its existence, has dealt with measurements of the highest levels of accuracy necessary to create measurement standards (etalons). These were always multiple measurements, both back when the etalons were prototypes and now when most etalons are based on stable quantum effects and speed of light and their accuracy has increased dramatically. Thus, the task of developing the foundational documents that concerned the whole metrology including everyday measurements did not match the experience and culture of BIPM.

In fact, it is probably for this reason that CIPM transferred this task to ISO, motivating the decision by its belief that ISO would be able to better reflect the interests of industry and trade (see the Foreword to GUM). However, the composition of the working group formed by ISO remained the same as during the time when the work was conducted under the direction of BIPM. The working group still included representatives from BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML. All these organizations are as authoritative in their respective areas as BIMP in metrology, but mostly similarly far removed from everyday measurements. Only IEC and OIML had necessary experience for this task, but we speculate they – being in a minority – could not set the tone for this work. Thus, the development of GUM was assigned to organizations that were not suitable for the task.

The second reason might be the way the discussion of the document draft was carried out. As noted in GUM, its draft was distributed for discussion to national metrological organizations. Given the great authority and reputation of BIPM, one could easily see some of these organizations to defer to BIMP without seriously considering the document. Others, even if they wanted to consider it, could have suffered from the common issue at the time – a strong dichotomy in metrologists' expertise, which was either centered on practical measurements but lacked rigorous mathematical background or focused on applied statistics but was far removed from measurement practice. Those focused on practical measurements could not completely assess the document full of mathematical formulas and references. At the same time, those who could fully understand the mathematics in the document did not have deep understanding of practice of measurements to understand the document's implications in this aspect. And even if there were some comments straddling the two sides of the coin, they were probably ignored by virtue of being in a minority. Thus, the GUM draft may not have been properly discussed.

In summary, GUM's flaws could be that on one hand, its development was assigned to organizations that did not have experience and culture of dealing with practical measurements, and on the other hand, it was adopted without effective discussion. With this understanding, we can now discuss avenues for correcting GUM and VIM.

9.6 Perspectives on Fixing GUM и VIM

According to the foreword to VIM and paper [16], the work on enhancing and correcting GUM and VIM, which used to be under direction of BIPM, was reorganized in 1997, and a Joint Committee for Guides in Metrology (JCGM) was created to direct this work. However, the chairperson of JCGM is the director of BIPM and the committee itself still consists of representatives of the same organizations that originally developed GUM and VIM. Only at a later stage JCGM added ILAC. Thus the reorganization that was carried out has not let to any significant change.

The first document prepared under the direction of JCGM is Addition 1 to GUM [13]. Addition 1 was presented as the correction of a mistake in GUM in computing the expanded uncertainty using the Monte Carlo method. However, the Monte Carlo method inherently includes the inaccuracy of moving from experimental data to their approximated distribution functions, which is not accounted in the final result. Thus, the Monte Carlo method is not suitable for this problem. We refer the reader to Sects. 5.7 and 5.8 for the more appropriate methods.

The above considerations lead to twofold suggestions for reorganizing the work on GUM and VIM. Our first suggestion concerns the organization that would direct the work. The work on GUM under the direction of BIPM took 17 years and in the end produced a flawed recommendation as showed earlier. Reorganization of this work through the creation of JCGM has not resulted in a meaningful change. To make the reorganization effective, the work needs to be assigned to an organization that possesses the necessary experience in the development of documents of this kind, such as, for example, OILM.

And second suggestion concerns the problem of organizing the discussion refereeing of document drafts. The goal here should be to engage specialists and obtain their input directly, and not through the bureaucratic administration layers of metrological organizations. This goal can be achieved if the task of considering the drafts would be assigned to a specially appointed commission of experts, which would be selected by an authoritative neutral organization. Such an organization could be, for example, ISO, assuming string rules for expert selection are adopted to avoid any conflict of interest that might affect the refereeing.