XX IMEKO World Congress
Metrology for Green Growth
September 9-14, 2012, Busan, Republic of Korea

HOW TO PROVE AND ASSESS CONFORMITY OF GUM-SUPPORTING SOFTWARE PRODUCTS

N. Greif, H. Schrepf

Physikalisch-Technische Bundesanstalt, Institute Berlin, Abbestr. 2-12, 10587 Berlin, Germany,
Norbert.Greif@ptb.de, Heike.Schrepf@ptb.de

Abstract: This paper describes how to prove and assess GUM conformity of software products which claim to offer a GUM-conform calculation of measurement uncertainties. To bridge the gap between the GUM guideline and the required test specification, an analysis of the GUM from the perspective of software testing is presented. Roughly, the benefit and the limits of the developed validation procedure are outlined.

Keywords: Measurement uncertainty, GUM conformity, measurement software quality, software validation.

1. INTRODUCTION

An increasing number of software products that claim to offer a GUM-conform calculation of measurement uncertainties are available on the market. In order to ensure that these products perform the calculations in accordance with the GUM [1, 2], a specific validation of the software products with respect to the GUM is necessary.

Additionally, to guarantee comparability of the measurement uncertainties calculated by different software products, a defined comparability of the software products themselves is required. Consequently, a reusable, automated test environment has been developed which supports both a GUM-oriented validation and GUM-related comparisons of different software products by tracing back the product features to the rules and requirements of the GUM.

Roughly, the paper presents the benefit of the test environment, but also the limitations of validation and product comparisons.

To bridge the gap between the GUM guideline and the required test specification, an analysis of the GUM from the perspective of software testing is presented. This detailed analysis of the GUM has uncovered some issues and inconsistencies within the GUM. Included are, for example, non-testable GUM statements, alternative options of implementations of GUM statements, and missing definitions to ensure unambiguous implementations of the GUM.

2. MOTIVATION AND AIM

For several years, the authors have been involved in the evaluation of software products which implement the GUM [3, 4, 5]. Recently, three further software products were comparatively evaluated concerning GUM conformity. During this work, special experience was gathered and an implementation-oriented view on the GUM has emerged. One aim of the paper is to outline this special experience.

For example, the following fundamental questions have come up regarding an implementation of the GUM:

- **Completeness**
  What does it mean when a software product claims to implement the GUM? Is the GUM completely implementable? Is it possible to reformulate each of the GUM statements so that it can be represented as a computational step? Is a pocket calculator already compliant to GUM when it correctly implements only one formula like average (as described in GUM 4.2.1)?
- **Correctness**
  What does it mean when a software product claims to be correct? Is the product able to calculate the correct values? Or is it able to calculate the correct values, round them with a correct rounding procedure, and display them with a correct number of digits?
- **Compliance**
  What does it mean when a software product claims to be compliant to the GUM? Is there a conformity statement?

Already these few questions lead to one of the core problems of both testing GUM software and comparing GUM test results: The need to trace back each computational step and each test result to a certain well-defined, well-understood and uniformly interpreted GUM statement. Consequently, traceability should be the precondition for the validation of a specific software product as well as the comparison of different products.

To get repeatable, comparable, and traceable validation results, the questions mentioned above and some further queries have to be answered.

The corresponding answers have an important impact on the set-up of the test environment. For the software specification step in between, specific guidance is given in [6].

3. ANALYSIS OF THE GUM FROM THE PERSPECTIVE OF SOFTWARE TESTING

In this main section of the paper, some problems of testability and ambiguity of GUM statements are analysed in detail. These issues have a straight influence on the traceability and comparability of validation results belonging to different software products.

In the following, the GUM issues under discussion are classified according to these four main categories:

- **Testability of GUM statements**
  Non-testable statements are analysed with respect to additional context information.
- **Strictness of GUM statements**
  Diffuse statements with regard to the possibility to use alternative options are analysed.
- **Ambiguity of GUM statements**
  Ambiguous statements regarding informal wording are discussed.
- **Residual category**
  Missing GUM statements, GUM inconsistencies and the handling of calculation results not covered by the GUM are analysed.

For each problem, the necessary decisions to guaranty testability, the unambiguous definition of the validation procedure, and the direct consequences for the development of the test environment are derived (cf. the examples with accepted solutions). Some solutions can not be realised within an automated test environment.

3.1 Testability of GUM statements

The GUM includes a series of statements which are not testable, even in the core sections 4 through 8. These statements require additional decisions or context information to guaranty the unambiguous definition of the test process. Software packages should be able to ask for the necessary context information. This has not been the case for all software packages validated so far.

Example 1 (GUM 4.2.1, GUM 4.2.3):

The GUM describes the computation of an arithmetic mean for an observation series and allows the use of the computed mean as an estimator for the quantity's value as long as the observation values are measured or collected under repeatability conditions ("…under the same conditions of measurement …").

Being allocated a list of observation values, no software package is able to decide whether the required repeatability conditions have been met.

Accepted solution 1: The package asks the user to check the conditions.

Accepted solution 2: The package always implicates certain repeatability conditions. The user manual points out the responsibility of the user.

Example 2 (GUM 3.2.3, GUM 3.2.4, GUM 8.1):

The GUM states that systematic deviations have to be incorporated into the model equation in the form of correction terms.

Being allocated a model equation, no software package is able to decide whether the model equation is complete in this sense.

Accepted solution: The package always implicates completeness of the model equations. The user manual points out the responsibility of the user.

Example 3 (GUM G.2.1):

The GUM states that the output quantity is approximately normally distributed if its variance is "much larger than …".

A software package is not able to compare two values in this informal way.

Accepted solution: The package offers all information necessary to decide on distribution of the output quantity to the user (distributions of all input quantities, their uncertainty contributions, the linearity of the model equation). The user should be able to decide whether the distribution of the output quantity may be understood as normal or is "unknown". In the case "unknown", the package should not calculate and report the expanded uncertainty of that output quantity.

Example 4 (GUM 5.1.2, GUM 5.1.5):

The GUM states that "higher terms (of the Taylor series of the model function) must be negligible".

Should a software package neglect a term when its value is 1/20, 1/100, or 1/1000 of the sum of the low-order terms?

Accepted solution 1: The software package calculates results for both, the standard GUM case (first order of Taylor series), and the sophisticated case (including higher order). If the results for $u_c(y)$ are equal after rounding and shortening, then the higher order terms are negligible.

Accepted solution 2: Alternatively to solution 1, the package can check the linearity of the model equation (e.g. via difference quotients).

Example 5 (GUM F.1.2.1 a) and c)):

The GUM explains that the covariance of two input quantities may be treated as insignificant if certain conditions are met.

The software package is not able to decide whether the required conditions have been met.

Accepted solution: The software package calculates both, the result with and without correlation. If the results for $u_c(y)$ are equal after rounding and shortening, then the correlation is negligible.

3.2 Strictness of GUM statements

The possibility to use alternative options requires additional decisions or assumptions to ensure testability. Such options have to be exercised, for example, in case of the formulation of model equations, or in case of the evaluation of sensitivity coefficients.

Example 6 (GUM 3.1.7, GUM 4.1.1, GUM 4.1.2):

GUM 4.1.1 presents the model relationship as an equation solved for the scalar output quantity.

GUM 3.1.7 mentions that the concept is applicable to vector results, too. However, there is no further treatment.

GUM 4.1.2 states that the relation between input quantities and output quantities are not necessarily an explicit functional relationship. Instead, an algorithm or a computer program which is able to produce result values y for certain input values x_i may be used.

Accepted solution: The model equation may be represented by an explicit functional relationship or not. Each variant is considered to be compliant.

Example 7 (GUM 5.1.3, GUM 5.1.4):

GUM 5.1.3 describes the sensitivity coefficients as partial derivatives of the output quantity with respect to the input quantities at the point of the estimates of the input quantity values. Note 2 of the same section states that the partial derivatives may be calculated using common numerical methods.

GUM 5.1.4 allows the experimental determination of these sensitivity coefficients.

Accepted solution: Sensitivity coefficients may be determined analytically, numerically, or experimentally. Each variant is considered to be GUM-compliant.

Example 8 (GUM 4.1.4):

GUM 4.1.4 states, that the estimated value of the output quantity is calculated using the estimated values of the input quantities and the model equation.

The following note says that the estimate of the output quantity may also be calculated as the average of several output values, each of them calculated from a set of input values and the model equation.

Accepted solution: The value y may be calculated as a function value or as an average of function values.

Each variant is considered to be GUM-conform. In case of linear models, the results do not differ.

Example 9 (GUM G.4.1, Note 1):

GUM G.4.1 describes the treatment of a degrees-of-freedom value calculated by the Welch-Satterthwaite formula. To derive the coverage factor, two different methods are allowed, interpolation or truncation.

Accepted solution: Each variant is considered to be GUM-compliant.

3.3 Ambiguity of GUM statements

Ambiguity is caused by informal GUM wording which shall improve readability and comprehensibility.

Example 10 (GUM 7.2.6, GUM H):

GUM 7.2.6 explains that the uncertainty should be given with "at most" two significant digits. More digits are allowed to avoid rounding errors in subsequent calculations.

GUM annex H mostly uses two, in some cases only one digit for uncertainty values (H.3, H.5, H.6).

What should the programmer of a GUM package do regarding the question of digits? How should the tester of a GUM package formulate the nominal output for a test case?

Accepted solution: The software package should use two digits by default. It should allow an adjustment if necessary.

Example 11 (GUM 7.2.6):

GUM 7.2.6 states that "it may sometimes be appropriate" to round uncertainties up rather than to the nearest digit. Two examples are given: A value like 10.47 should be better rounded up to 11 instead of rounding it to the nearest digit, i.e. 10. In another case, a value like 28.05 should be rounded to the nearest digit, i.e. 28, instead of rounding up to 29.

The GUM obviously uses a rounding principle that is describable as "rounding up or down with a fraction limit somewhere between 0.1 and 0.4, instead of 0.5 as is usual". Since this is not formulated explicitly, each programmer is free to use rounding up or rounding to the nearest digit (and half up).

Accepted solution: Concerning testing, the decision was made to expect rounding to the nearest digit (and half up).

Example 12 (GUM G.6.6):

GUM G.6.6 explains that in certain cases one may use the coverage factor values of 2 (to get a level of confidence of nearly 95 %) or 3 (to get nearly 99 %). Afterwards, the GUM discusses that in these cases significant over- and underestimations of the confidence interval may occur and that a better estimation may be necessary. The user is recommended to choose a better estimation if the approximation is not sufficient for his purposes.

The question arises whether a GUM package should use the (GUM-compliant) approximation or the (GUM-compliant) better estimation.

Accepted solution: The solution is based on a case-by-case analysis of the important influencing factors. One possible solution for analysing the factors and operating the software package is the following:

1. The user gets all information from the package (distributions of all input quantities, their uncertainty contributions, the linearity of the model equation, effective degrees of freedom).
2. The user decides on the distribution of the output quantity (normal distribution or not).
   2.1. Normal distribution of output quantity:
   2.1.1. The user delivers the level of confidence p and the package calculates the coverage factor t(ν) based on a t-distribution, or
   2.1.2. the user delivers the level of confidence p and the package calculates the coverage factor k based on a normal distribution, and it delivers the deviation between k and t(ν).
   2.2. Distribution of the output quantity is not normal:
   2.2.1. The distribution is unknown; the package does not calculate the coverage factor.
   2.2.2. The distribution is known; the user delivers the coverage factor.

Example 13 (GUM F.2.3.3):

GUM F.2.3.3 discusses the case in which only a minimum and maximum value (and therefore the half width a) for an input quantity is available. The suggestions for the uncertainty of this input quantity vary from $a/\sqrt{3}$ (for a uniform distribution assumption), to $a/\sqrt{6}$ (for a triangular distribution assumption), and to $a/\sqrt{9}$ (for a normal distribution assumption).

Accepted solution: The package has to ask the user. He has to decide which assumption holds.

3.4 Residual category

Finally, some problems regarding missing GUM statements, GUM inconsistencies, and the handling of calculation results not covered by the GUM are considered.

Example 14 (GUM G.4):

The problem of effective degrees of freedom of the output quantity is discussed in relation to the problem of the output quantity's distribution and other aspects (central limit theorem). The given formula (G.2b) and the reference to section GUM 5.1.3 suggest that the formula is valid for uncorrelated input quantities only, but this is not expressed explicitly or discussed in detail.

In particular, there is no explicit prescription not to use formula (G.2b) in case of correlated input quantities.

Accepted solution: The calculation of a degree-of-freedom value for the output quantity in case of correlated input quantities is not considered to be compliant. A value may be given, but its calculation has to be documented, and it has to be marked as outside the GUM scope.

Example 15 (GUM 4.3.8, GUM G, GUM F):

A topic which is discussed very roughly is the usage of input quantities with asymmetric distributions. In this case, GUM statements consist of a single section in the main text, a short discussion of negligibility in annex G, and the discussion of a particular case in annex F.

The question arises: How should the user deal with asymmetrically distributed input quantities? They cannot be omitted, since GUM does not prohibit its use.

Accepted solution: The distributions of the input quantities do not influence the computation of the value of the output quantity y and the standard measurement uncertainty $u_c(y)$. To prove GUM conformity, the asymmetric distribution of input quantities can be ignored, knowing that these results may not be correct.

Example 16 (GUM 6, GUM G):

The problem of how to evaluate the expanded uncertainty of an output quantity (which is in practice of greater interest than the standard uncertainty) is only briefly discussed. GUM 6 suggests to use a coverage factor between 2 and 3, and mentions that the selection of a proper value depends on experience or, alternatively, on knowledge about the output quantity's distribution. The details of this discussion take place in annex G.

From the tester's viewpoint, the GUM doesn't state that the annexes are as mandatory as the main text, so the question arises as to how this important part of uncertainty evaluation should be dealt with.

Accepted solution: The annexes are considered not so mandatory. The user should be able to decide about the distribution of the output quantities (cf. examples 3 and 12).

Example 17 (overall GUM):

It is common sense that a correlation matrix should be checked with respect to its being non-negative definite. Most of the GUM packages allow the user to do this check, but it is not discussed in the GUM.

Accepted solution 1: The software package checks the non-negative definiteness of the correlation matrix.

Accepted solution 2: The Software package does not check the non-negative definiteness of the correlation matrix. Instead of that, before the output of the standard measurement uncertainty $u_c(y)$ of the output quantity, the package checks that the expression for $u_c^2(y)$ is non-negative.

Example 18 (overall GUM):

The experience from the GUM packages that have been validated is that most of these packages compute
- confidence intervals for output quantities with rectangular distribution,
- effective degrees of freedom in case of correlated inputs, and
- confidence intervals for correlated output quantities, irrespective of the fact that the GUM does not prescribe anything in these cases.

Accepted solution: Because these calculation results are not covered by the GUM, they do not belong to a validation of a package with respect to GUM conformity. On the other hand, however, these results are important in practice. With regard to the test process, testing of these calculations is performed, but the corresponding test cases are marked as "outside GUM conformity testing".

4. ROUGH SURVEY OF THE TEST ENVIRONMENT

In this section of the paper, the test environment as it has been developed for the validation and GUM-related comparison of software products is described very roughly. A schematic overview of the test environment is illustrated in figure 1. A detailed presentation is given in [5].

The following description is restricted to the overall understanding of the test concept and to some aspects which are of importance for the analysis of benefits and the problems mentioned above. Implementation details are omitted.

The objective to validate software products that implement the GUM is best achieved by establishing a well-defined, GUM-oriented test process supported by a reliable technical test environment. The environment itself has to obey certain quality requirements, for example, correctness and completeness. Especially, the test cases must be designed in a way that they generally fit for any GUM-supporting software product under test. Consequently, comparability of certain validation results and after all the comparability of the whole validation process has to be ensured.

To meet these requirements, the test environment consists of the following components:

- **Data model** defining the structure of information necessary for uncertainty calculations and corresponding tests.
- **Set of universal test cases** which do not contain any product-specific or technical information.
- **Test case converter** which translates universal test cases into product-oriented specific ones; the converter needs information about the package to be tested, the underlying operating system, the test tool which will be used, etc.
- **Several sets of product-oriented specific test cases**, each of which belongs to a specific software package to be tested.
- **Capture-replay test tools** to operate the test cases and to repeat automatically the overall test process.

The universal and package-specific test cases are arranged concerning a well-defined classification scheme. The respective position of a test case in the scheme corresponds to the purpose of the test. In this way, the classification scheme allows a certain control of completeness and traceability of the validation process. The main levels of the classification hierarchy are:

- Assignment of the test cases to the set of software quality characteristics according to the international software standard ISO/IEC 25010 [7], for example, *functionality*, *usability*, and *reliability*.
- Subdivision of test cases into positive cases (prove that the GUM is correctly implemented) and negative cases (prove that in case of the non-applicability of the GUM no calculation is carried out).
- Specific subdivisions depending on the value for the first level.

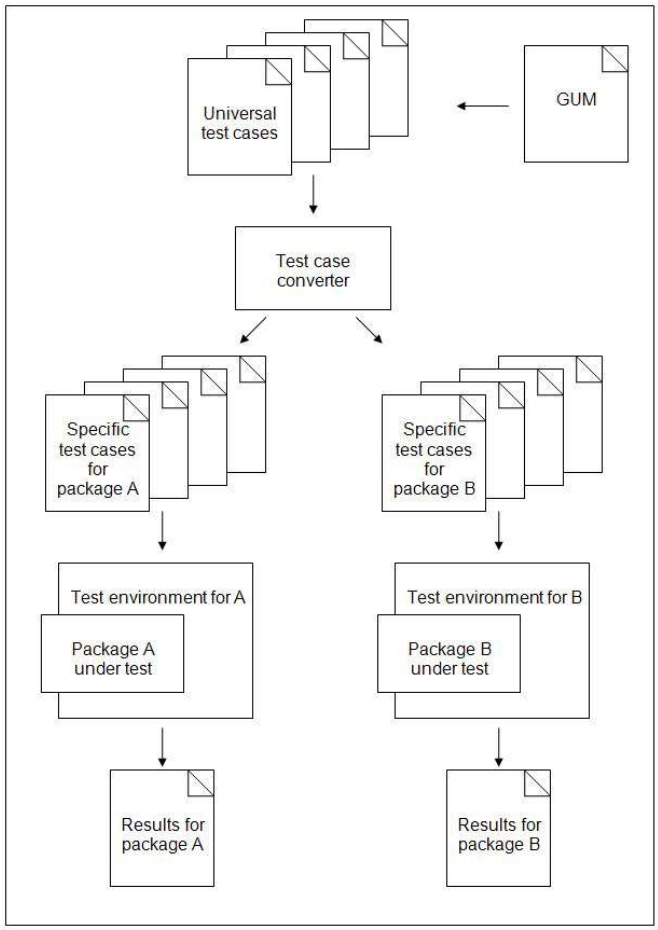

Figure 1: Schematic overview of the test environment

An example for the third classification level is closely connected with the software quality characteristic *functionality*. In this case, the classification hierarchy represents the detailed calculation steps needed to prove the conformity of the software packages to the core sections and formulas of the GUM. The calculations are split into the following steps (branches of the hierarchy):

- Calculations for processing a Type A input quantity (without correlation of inputs);
- Calculations for processing a Type B input quantity (without correlation of inputs);
- Interpretation of model equations and calculation of sensitivity coefficients;
- Calculation of values, standard measurement uncertainties, and coverage intervals of output quantities without and with the correlation of input quantities;
- Calculation of the correlations between output quantities (vector results);
- Calculations of the examples from GUM Annex H.

For each of these calculation steps, further classification levels depending on the degree of complexity of the test cases can be defined. Normally, we use between five and nine classification levels.

In addition to the quality characteristic *functionality*, the characteristics *usability* and *reliability* were used to design and implement test cases. In future, the characteristic *efficiency* might become relevant to include response time evaluations of Monte Carlo simulation engines.

5. CONCLUSIONS

A number of software packages which claim to implement the GUM are on the market. However, they differ in functionality and the have deficiencies which are not obvious. Thus, a validation of these packages with respect to the GUM is necessary.

The PTB test environment has been used successfully to validate and compare three different GUM-supporting software packages.

To bridge the gap between the GUM guideline and the explicit test specification, a detailed analysis of the GUM from a tester's perspective and certain decisions regarding the test process (cf. the accepted solutions) were necessary. Based on the results of this analysis, unambiguous and detailed test cases could be developed. The benefits of the test environment and the validation procedure are:

- General procedure usable for any GUM-supporting software product;
- Automated and reusable process;
- Comparability of the validation procedure and, especially, of the validation results;
- Automated documentation process.

However, there are also limitations in the validation procedure, and in the process of product comparison. The current procedure does not consider Monte Carlo simulations and does not regard the handling of complex numbers. The general limitation is, that several obstructive characteristics of GUM statements (with regard to software testing), such as ambiguities, missing or inexact specifications/definitions, do restrict the applicability and the objectiveness of the test environment. Thus, some of the accepted solutions can not be realised within an automated test environment.

Concerning the software quality characteristics, up to now, the validation procedure does not include *efficiency* testing (e.g. duration of Monte Carlo simulations).

In principle, the test environment is prepared to realise the extensions mentioned above. Some extensions concerning Monte Carlo simulations and vector results are already under construction.

6. REFERENCES

[1] ISO/IEC Guide 98-3:2008, Uncertainty of measurement - Part 3: Guide to the expression of uncertainty in measurement, 2008.

[2] Evaluation of measurement data - Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method, JCGM 101, 2008.

[3] N. Greif, H. Schrepf, D. Richter, Software validation in metrology: A case study for a GUM-supporting software, Measurement 39 (2006) 849-855, Elsevier, 2006.

[4] N. Greif, H. Schrepf, Validierung von Software zur Bestimmung von Messunsicherheiten, VDI-Berichte 1947, Messunsicherheit praxisgerecht bestimmen, 409-418, 2006.

[5] N. Greif, H. Schrepf, V. Hartmann, G. Kilz, A test environment for GUM conformity tests, Physikalisch-Technische Bundesanstalt (PTB), Braunschweig und Berlin, PTB Report, to appear, 2012.

[6] M. G. Cox, P. M. Harris, I. M. Smith, Software specification for uncertainty evaluation, NPL Report MS 7, March, 2010.

[7] ISO/IEC 25010:2011, Systems and software engineering - System and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models.