

# On the transferability of the GUM-S1 type A uncertainty

Gerd Wübbeler and Clemens Elster

Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin, Germany

E-mail: [gerd.wuebbeler@ptb.de](mailto:gerd.wuebbeler@ptb.de)

Received 8 August 2019, revised 21 October 2019

Accepted for publication 24 October 2019

Published 23 January 2020



## Abstract

A key requirement for the evaluation of measurement uncertainty in metrology is its transferability. This is particularly relevant for ensuring traceability through a chain of measurements. The GUM [JCGM 100] requires that ‘*it should be possible to use directly the uncertainty evaluated for one result as a component in evaluating the uncertainty of another measurement in which the first result is used*’. The GUM implements the transfer of uncertainty by applying variance propagation to linear models or models that can be well approximated through a linearization. GUM-S1 [JCGM 101] has been released to broaden the scope of uncertainty evaluation to cover also non-linear models. We demonstrate in terms of examples that the GUM-S1 type A evaluation does not satisfy the requirement of transferability, i.e. re-using the probability distribution produced by GUM-S1 in a subsequent uncertainty exercise can lead to inadequate results for non-linear models. Furthermore, already for linear models the type A evaluation of GUM-S1 is shown to produce unsatisfactory solutions. These findings are discussed and the underlying reason is identified, namely that the use of non-informative priors as in GUM-S1 does not lead to results that can always be reliably transferred. We discuss possible alternatives and finally draw some conclusions.

Keywords: GUM-S1, type A uncertainty evaluation, Bayesian inference, non-informative prior, long-run success rate

(Some figures may appear in colour only in the online journal)

## 1. Introduction

The dissemination of the Guide to the Expression of Uncertainty in Measurement (GUM) [1] has facilitated the worldwide harmonization of uncertainty evaluation in metrology. The GUM uncertainty framework uses the propagation of variances based on a linear, or linearized, model for the measurand. Supplement 1 to the GUM (GUM-S1 [2]) expands the scope of the GUM to non-linear models by replacing the propagation of uncertainties with the so-called propagation of distributions. In the presence of an input quantity for which type A information is available, the distribution produced for the measurand has been shown to be equivalent to a Bayesian inference that uses specific non-informative priors for the measurand and for the variance of the distribution from which the data of the type A input quantity are assumed to have been drawn [3–7]. Since the GUM type A evaluation can be viewed

as a frequentist one [8], GUM-S1 also evoked a change of paradigm towards the Bayesian point of view regarding type A evaluation of measurement uncertainty, along with an inconsistency of corresponding results.

Several papers [9–12] have raised criticisms of GUM-S1. While [10, 11] provided examples for which GUM-S1 showed an unsatisfying long-run success rate, [9] criticized GUM-S1 for approaching the Bayesian point of view and, particularly, for using non-informative priors. In [12], on the other hand, it is argued that the uncertainty produced by GUM-S1 can be a poor estimate for the standard deviation of the estimate of the measurand. In this paper, we add a further critical aspect by showing that the GUM-S1 type A evaluation does not satisfy the requirement of transferability, a key requirement for any uncertainty evaluation method applied in metrology.

One essential task of metrology is to ensure the traceability of measurements performed on the shop floor level to

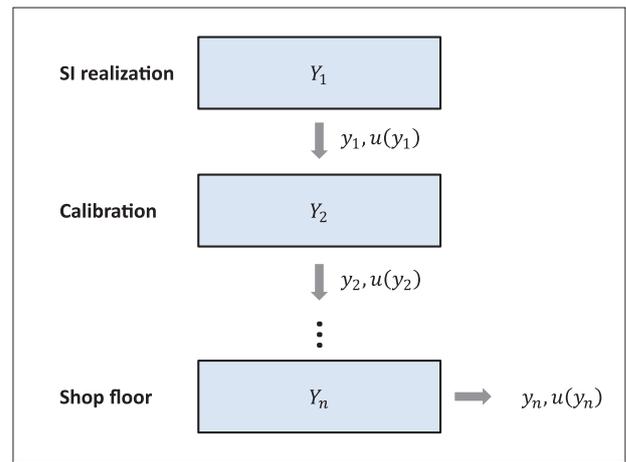
the international system of units (SI), see figure 1. The result of each step in such a measurement chain, and in particular the uncertainty associated with the reported estimate, is taken as the input of the subsequent measurement. It is mandatory that the applied method of uncertainty evaluation is transferable in the sense that ‘it should be possible to use directly the uncertainty evaluated for one result as a component in evaluating the uncertainty of another measurement in which the first result is used’ [1]. And this statement intends to not just require the technical possibility of re-using uncertainties subsequently, but that such a re-use leads to reliable results. The GUM calculus for the propagation of variances meets this requirement for linear models which may be one reason for its successful application throughout so many applications in metrology.

We demonstrate in terms of simple examples that the GUM-S1 type A evaluation of uncertainty fails to generally ensure the requirement of transferability. It is shown that even for linear models unsatisfactory results can be obtained. An asymptotic analysis is carried out that explains the behavior for the considered sequence of linear models. We discuss the underlying reason for our findings, which turns out to be the automatic choice of the non-informative prior involved in the GUM-S1 type A evaluation. Our reasoning is not based on an objection against Bayesian methods in general, or the use of non-informative priors in particular; rather, we argue that results obtained in a Bayesian inference based on a non-informative prior are generally not suitable for re-use in a subsequent analysis. In fact, we show for one of the non-linear examples that a more suitable choice of the non-informative prior applied to all data and the whole sequence of models at the same time leads to satisfactory results. However, such a procedure is neither practicable nor desirable for applications in metrology. Possible alternatives to the GUM-S1 type A uncertainty evaluation are discussed, namely a subjective Bayesian inference and the use of vaguely informative proper priors.

The paper is organized as follows: in section 2, the GUM-S1 results for three different examples are presented. Section 3 then provides an asymptotic analysis for the sequence of linear models that explains the observed behavior. The unsatisfactory transferability of results for the non-linear models is then discussed in section 4 including results obtained by the reference posterior for one of the non-linear models. Section 5 contains a critical assessment about the use of non-informative priors in metrology and discusses possible alternatives. Finally, we draw some conclusions from our findings.

## 2. Performance of GUM-S1 for generic examples

In this section, the performance of the GUM-S1 type A uncertainty evaluation is explored for three sequences of models: one that consists of linear models and two that include non-linear models. The sequences of models are designed in such a way that they will mimic measurement chains as illustrated in figure 1. For the sake of simplicity, we consider the case in which all input quantities are ‘type A input quantities’,



**Figure 1.** Measurements performed on the shop floor are traced to the SI realization through an unbroken chain of calibrations. Estimates and uncertainties are transferred at each step.

i.e. information about the input quantities shall be based on observed data. This reflects a situation in which the type A input quantities provide the dominant source of uncertainty and allows to focus on the suitability of the GUM-S1 type A evaluation method. It is likely that prior information about the type A input quantities is available; however, such information is not considered here in accordance with GUM-S1. All examples are simulated examples so that the ground truth is known, and we will assess the performance of an uncertainty evaluation in terms of how well the expected size of the deviations of the estimates from the known ground truth is characterized by the uncertainties obtained.

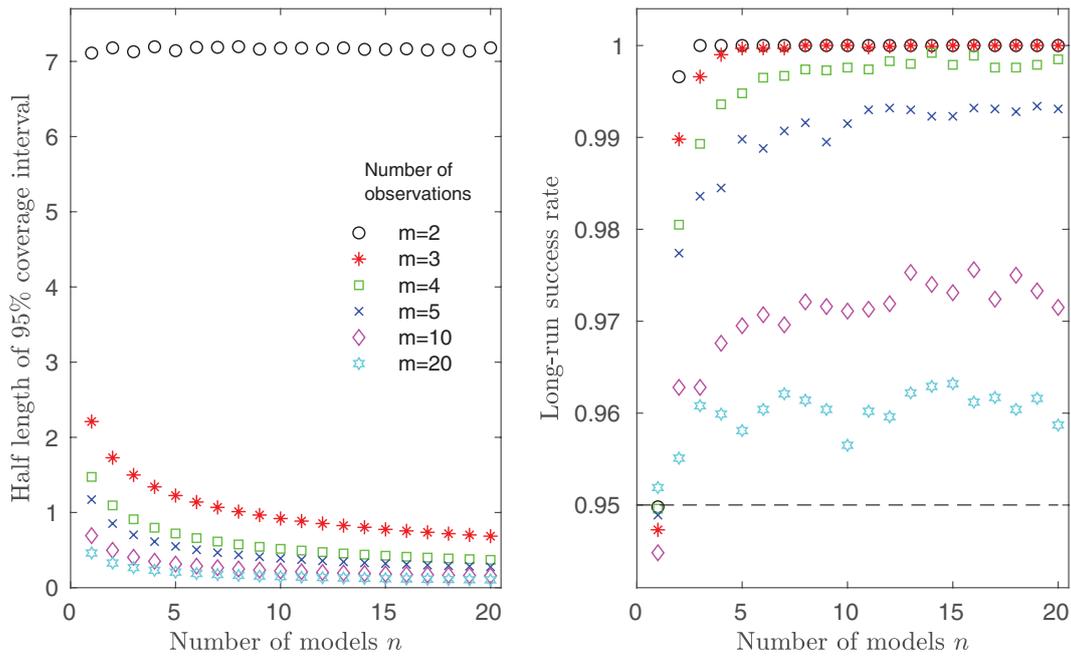
We note that such long-run success rates, and in particular frequentist properties for repeatedly sampling the data, are generally not viewed as a relevant criterion for a subjective Bayesian inference (e.g. [13, 14]); we will come back to this issue in the discussion in section 5. Here, we just mention that, since GUM-S1 does not consider the use of subjective priors, and since a GUM-S1 type A uncertainty evaluation is based on the observed data only, its result ought to be reliably linked with the ground truth used for producing the data when repeatedly applied.

### 2.1. Example 1—sequence of linear models

Consider the following sequence of linear models

$$\begin{aligned}
 Y_1 &= X_1, \\
 Y_2 &= Y_1 + X_2, \\
 &\vdots \\
 Y_n &= \frac{1}{n} (Y_{n-1} + X_n), \tag{1}
 \end{aligned}$$

where the result of each model is subsequently fed into the next model. More precisely, for the  $i$ th model  $Y_i = Y_{i-1} + X_i$ , it is assumed that information about  $Y_{i-1}$  and  $X_i$  is available only, but not about  $X_1, X_2, \dots, X_{i-1}$ . The particular sensitivity coefficients in (1) have been chosen for the sake of convenience, and the obtained results can be considered representative for



**Figure 2.** Mean value of half length of 95% coverage intervals (left graph) and long-run success rates (right graph) obtained by applying GUM-S1 for the sequence of linear models (1) in dependence on the number of models  $n$ , and for different numbers of observations  $m$  per input quantity. The dashed line in the right graph indicates a success rate of 95%.

chains of linear models provided that the sensitivity coefficients are balanced and the final model for  $Y_n$  is not dominated by one or only a few of all input quantities involved.

For each input quantity  $X_i$ ,  $m_i > 1$ , repeated observations  $x_{i1}, \dots, x_{im_i}$  shall be available which are assumed to have been drawn independently from a normal distribution with mean  $\mu_i$  and variance  $\sigma_i^2$ . The  $\mu_i$  are the true values of the input quantities  $X_i$  and they are unknown to the analyzer, as are the variances  $\sigma_i^2$ . For the simulation of example data, we chose the specific values  $\mu_1 = \mu_2 = \dots = \mu_n = 0$  and  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = 1$ . Furthermore, the same number  $m$  of repeated observations was used for all input quantities, i.e.  $m_1 = m_2 = \dots = m_n = m$ . Here, the focus is on the uncertainty evaluation associated with  $Y_n$ , the measurand obtained at the end of the measurement chain. We note that, in [9], a model similar to (1) has been analyzed for the case  $n = 2$  with respect to the resulting coverage factors obtained by applying GUM and GUM-S1.

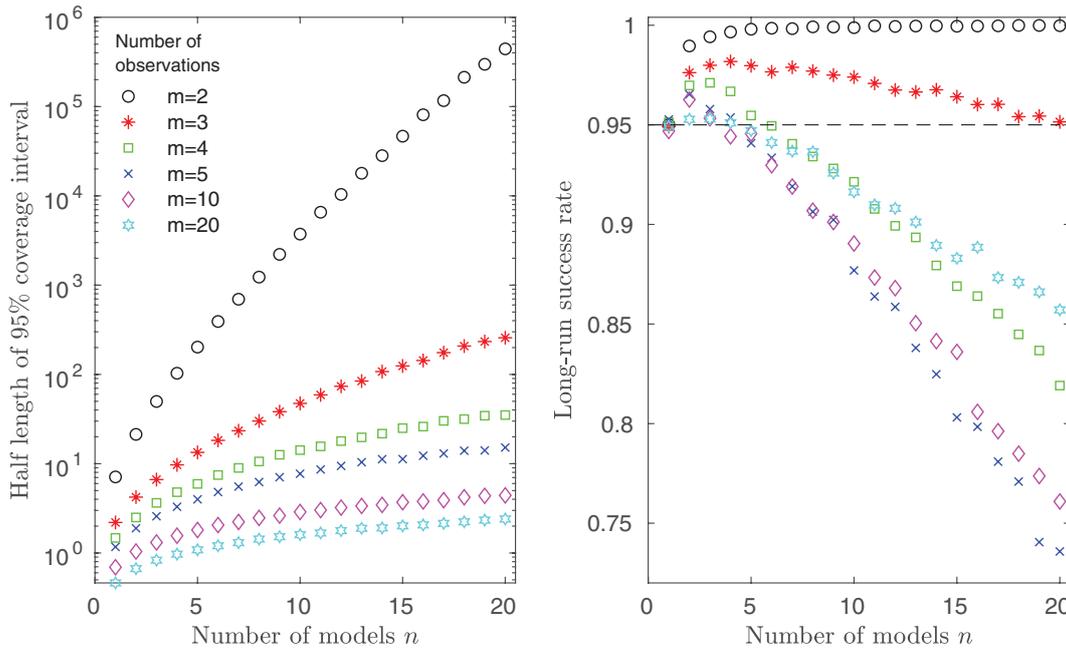
A simulation study has been conducted in which data for the input quantities have been simulated, to which GUM-S1 has been applied as follows. A scaled and shifted  $t$ -distribution is assigned to the first input quantity  $X_1$  with scale and shift calculated from the observations on  $X_1$  as described in table 1 of GUM-S1. This distribution is taken as the distribution for  $Y_1$ . The distribution for  $Y_1$  is then taken as an input for the evaluation of the second model, along with a second scaled and shifted  $t$ -distribution for  $X_2$ . Applying the propagation of distribution procedure from GUM-S1 yields the distribution for  $Y_2$ , which is then taken as an input for the third model, and so on<sup>1</sup>. In this way, the distribution for  $Y_n$  is obtained. From

this distribution, a 95% coverage interval is then calculated as described in GUM-S1.

Data simulation for all input quantities and calculation of a 95% coverage interval for  $Y_n$ , has been repeated a 10000 times, and the number of cases in which a 95% coverage interval contained the true value of  $Y_n$  was counted. The whole procedure was carried out for different lengths  $n$  of the measurement chain and different numbers  $m$  of repeated observations for the input quantities. The left graph in figure 2 shows the resulting mean values of the half lengths of the 95% coverage intervals (i.e. expanded uncertainties) in dependence on the number of input quantities  $n$  and the number of observations  $m$  per input quantity. For  $m > 2$ , a decrease in the average coverage interval length with an increasing number of input quantities can be observed, which could be expected since model (1) equals the average of the input quantities. Interestingly, this decrease is not observed for  $m = 2$ ; instead, the coverage interval lengths turns out to be almost independent with respect to the number of input quantities. The success rates for 95% coverage intervals to contain the true value for  $Y_n$  shown in the right graph of figure 2 reveal that, for a single input quantity  $n = 1$ , GUM-S1 yields perfect long-run success for all values of  $m$ . When the number of input quantities increases, the success rates quickly approach values above 0.95. In terms of long-run success, this indicates that the coverage interval length has been overestimated by GUM-S1. The small fluctuations in the long-run success rates shown in figure 2 are due to the finite number of 10000 repetitions used.

These results demonstrate that, for the sequence of linear models (1), the GUM-S1 type A uncertainty evaluation leads to a marked overestimation of the expanded uncertainty with respect to the long-run success criterion, especially when

<sup>1</sup>  $Y_n$  can equivalently be obtained through the convolution of  $n$   $t$ -distributed random variables.



**Figure 3.** Mean value of half lengths of 95% coverage intervals (left graph) and long-run success rates (right graph) obtained by applying GUM-S1 for the sequence of products models (2) in dependence on the number of models  $n$ , and for different numbers of observations  $m$  per input quantity. Note the logarithmic scale of ordinate used for the half length of 95% coverage intervals (left graph). The dashed line in the right graph indicates a success rate of 95%.

the number of observations of each input quantity is small. Overestimation of uncertainty may be seen as critical, particularly when it causes expensive (and unnecessary) actions, e.g. in conformity assessment. Moreover, according to the GUM [1], uncertainties should be ‘realistic’; the use of ‘safe’ or ‘conservative’ uncertainties is deprecated. An asymptotic analysis of the GUM-S1 results for the sequence of linear models (1) is provided in section 3.

2.2. Example 2—sequence of products of quantities

Consider the following sequence of products of quantities

$$\begin{aligned}
 Y_1 &= X_1, \\
 Y_2 &= Y_1 X_2, \\
 &\vdots \\
 Y_n &= Y_{n-1} X_n,
 \end{aligned}
 \tag{2}$$

where, similar to the sequence of linear models (1), the result of each model is subsequently fed into the next model. In the context of the  $i$ th model, the only information used shall be the measurand  $Y_{i-1}$  of the previous model and the current input quantity  $X_i$ . Non-linear models of the kind in (2) arise, for example, when multiplicative correction factors are applied.

For the sequence of products (2), we again assume  $m$  repeated observations  $x_{i1}, \dots, x_{im}$  to be available for each input quantity  $X_i$ , where these observations have been drawn independently from a normal distribution with unknown mean  $\mu_i$  and unknown variance  $\sigma_i^2$ . For the simulation of example data, we chose the specific values  $\mu_1 = \mu_2 = \dots = \mu_n = 1$  and  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = 1$ . Similarly to the example with

the sequence of linear models, 10 000 data sets were simulated and subsequently analyzed by applying GUM-S1. Following table 1 in GUM-S1, scaled and shifted  $t$ -distributions were assigned to the input quantities  $X_i$ . The GUM-S1 performance was then assessed in terms of the resulting half length of the coverage intervals, and in terms of the long-run success rate given by the portion of 95% coverage intervals for  $Y_n$  containing the true value of  $Y_n$ .

Figure 3 (left) shows the resulting mean values of the half lengths of 95% coverage intervals in dependence on the number of models  $n$ , and for different numbers  $m$  of repeated observations. As can be seen, the half lengths of the coverage intervals increase with an increasing number of models  $n$ , while for  $m = 2$  observations per input quantity large coverage intervals are obtained (note the logarithmic ordinate in the left graph of figure 3). The corresponding long-run success rates are shown in figure 3 (right), where perfect long-run success rates for all values of  $m$  were recorded for  $n = 1$ . Note that, for  $n = 1$ , model (2) equals (1). Similarly to the sequence of linear models (1), long-run success rates of almost 1 arise for  $m = 2$  and  $n > 2$ , indicating that applying GUM-S1 to the sequence of products (2) can lead to overestimated expanded uncertainties. However, the situation for model (2) appears to be more complex, since for this model also long-run success rates well below 0.95 can be obtained. For example, for  $n = 20$  and  $m = 10$ , the long-run success rate is only about 0.75, which indicates an underrating of the expanded uncertainty. These findings illustrate that application of the GUM-S1 type A uncertainty evaluation to a sequence of non-linear models does not in general lead to results that can be transferred reliably.

2.3. Example 3—sequence of sums of squared quantities

Consider the following sequence of sums of squared quantities

$$\begin{aligned} Y_1 &= X_1^2, \\ Y_2 &= Y_1 + X_2^2, \\ &\vdots \\ Y_n &= Y_{n-1} + X_n^2. \end{aligned} \tag{3}$$

Similarly to the previous examples, the result of each model is subsequently fed into the next model and it is assumed that, for the  $i$ th model, the only information available is on  $Y_{i-1}$  and  $X_i$ . Here, we shall assume, that for each input quantity  $X_i$ , a single observation  $x_i$  drawn from a normal distribution with unknown mean  $\mu_i$  and known variance  $\sigma_i^2$  is given. In the sense of GUM-S1, a normal distribution with mean  $x_i$  and variance  $\sigma_i^2$  is then assigned to each input quantity  $X_i$ . For the simulation of example data, we chose the same values  $\mu_1 = \mu_2 = \dots = \mu_n = \mu$  and  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = 1$  for all input quantities. The focus is again on the uncertainty evaluation associated with  $Y_n$ , the measurand obtained at the end of the measurement chain. Sum-of-squares models similar to (3) arise, for example, when evaluating the energy of a discrete time-dependent signal where the uncertainty for each observation is known [15].

Sequential application of GUM-S1 to the sequence of models (3) is equivalent to one application of GUM-S1 to the single model

$$Y_n = X_1^2 + X_2^2 + \dots + X_n^2, \tag{4}$$

for which in [11] already the insufficiency of the long-run success performance of GUM-S1 was demonstrated, and a similar conclusion for a related model was drawn in [10]. Our point in iterating some of those results for this model is that of viewing it as a sequence of models, for which at each stage  $i$  only the information on the previous measurand  $Y_{i-1}$  and the current input quantity  $X_i$  is assumed to be available. This prevents one from taking a more favorable non-informative prior in view of the final model, which will be discussed in section 4.

Figure 4 shows the distribution of  $Y_n$  obtained by means of GUM-S1 for a typical set of simulated data when choosing  $n = 20$  models and  $\mu = 1$ , together with the underlying true value. The GUM-S1 distribution is clearly separated from the location of the true value around which its probability density is practically zero; thus, applying GUM-S1 would result in a completely incorrect conclusion about the value of the measurand here.

As in to the previous examples above, many data sets have been simulated according to the models (3) and analyzed by applying GUM-S1. Owing to a possible skewness of the GUM-S1 distribution arising for this example, the long-run success rates were evaluated here using 95% highest posterior density intervals [16]. Figure 5 (left) shows corresponding results in dependence on the number of models  $n$ , and on the mean values  $\mu_1 = \dots = \mu_n = \mu$ . Reasonable success rates above, say, 90% are achieved only for small values of  $n$  and/or when  $\mu$  is large. For increasing values of  $n$ , a clear drop of the

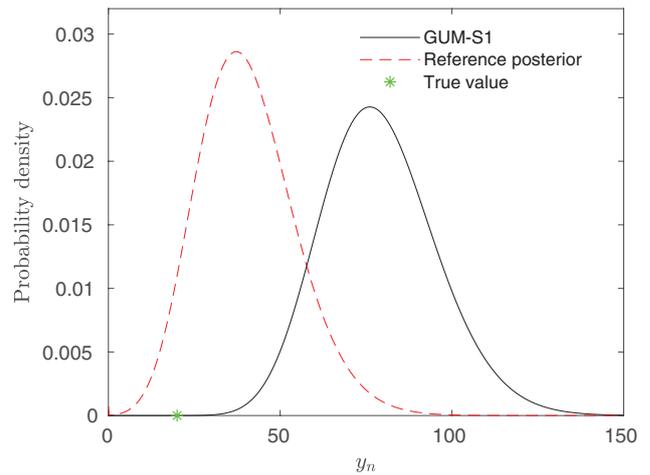


Figure 4. Posterior distributions obtained for the sequence of non-linear models (3) via GUM-S1 (black) and reference posterior (red) for a typical example when  $\mu_1 = \dots = \mu_n = 1$  and a measurement chain of  $n = 20$  models. The green star indicates the location of the true value  $y_{\text{true}} = 20$ .

success rate is observed; for a setting of, for example,  $n = 20$  and  $\mu = 1$ , the coverage rate is practically zero (see figure 4). That is, for this setting, the true value is almost never found within a 95% GUM-S1 coverage interval.

3. Asymptotic behavior of the sequence of linear models

In the following, the results obtained for the sequence of linear models (1) are studied analytically as the number of subsequent measurements becomes larger. The analysis of the sequential model (1) is equivalent to that of

$$Y_n = \frac{1}{n} \sum_{i=1}^n X_i, \tag{5}$$

with data  $x_{ij} \sim N(\mu, 1), j = 1, \dots, m$ , where  $\mu$  denotes the common value of all input quantities  $X_i, i = 1, \dots, n$ , and where  $m \geq 4$  is assumed. Let  $x_i$  denote the mean and  $s_i^2$  the squared standard deviations of the  $m$  repeated observations for  $X_i$ . Since GUM-S1 assigns scaled and shifted  $t$ -distributions to each  $X_i$  with means  $x_i$ , the estimate produced by GUM-S1 is given via

$$y_n = \frac{1}{n} \sum_{i=1}^n x_i. \tag{6}$$

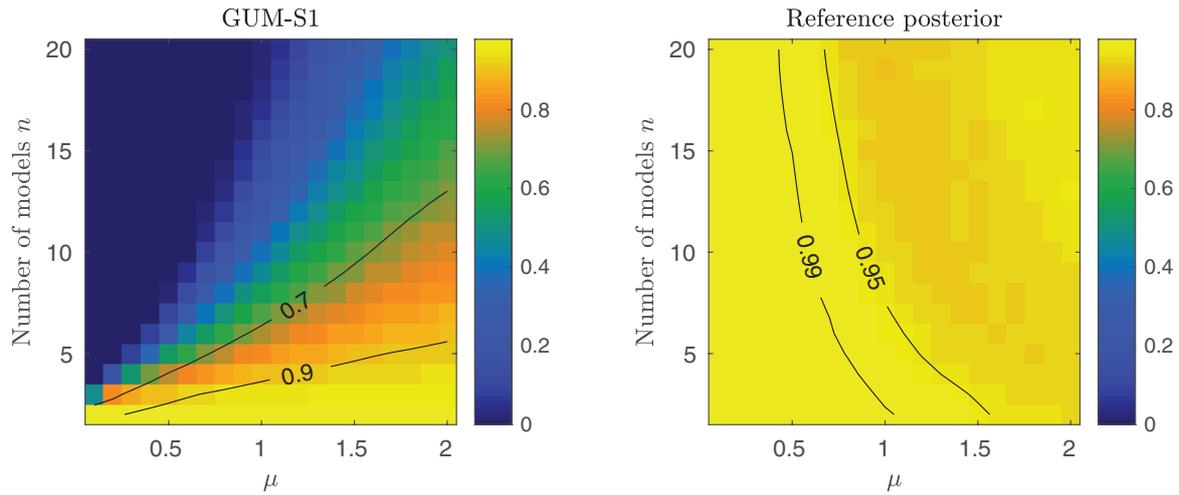
Since  $x_i|\mu \sim N(\mu, 1/m)$ , one obtains

$$y_n|\mu \sim N(\mu, \varphi/n), \tag{7}$$

where

$$\varphi = \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{m} \right) = \frac{1}{m}. \tag{8}$$

Note that  $y_n|\mu$  in (7) denotes the distribution of  $y_n$  under repeated sampling given the values of  $\mu$  and  $\varphi/n$ . It follows that  $y_n$  is unbiased (i.e.  $E(y_n|\mu) = \mu$ ); since  $\text{Var}(y_n|\mu) \rightarrow 0$



**Figure 5.** Long-run success rates for the sequence of non-linear models (3) for various settings of the length  $n$  of the measurement chain and mean values  $\mu_1 = \dots = \mu_n = \mu$  of the input quantities. The left graph shows the success rates obtained for GUM-S1, while the right graph shows those of the reference posterior (see section 4). The color scale indicates the range of long-run success rates and the black lines in both graphs show isocontour lines at the levels indicated.

as  $n \rightarrow \infty$ ,  $y_n$  also is a consistent estimator of the measurand. Because

$$\left( \frac{1}{nm} \sum_{i=1}^n s_i^2 \right) \xrightarrow{P} \varphi, \tag{9}$$

as  $n \rightarrow \infty$  (where  $\xrightarrow{P}$  indicates convergence in probability), and since, for  $y_n = \sum_i x_i/n$ , application of the central limit theorem yields  $\sqrt{n}((y_n|\mu) - \mu) \xrightarrow{D} N(0, 1/m) = N(0, \varphi)$  (where  $\xrightarrow{D}$  stands for convergence in distribution), it follows that

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left( y_n | \mu \in \left[ \mu \pm \sqrt{\frac{1}{nm} \sum_{i=1}^n s_i^2} \sqrt{\frac{1}{n}} z_{[1-(1-\alpha)/2]} \right] \right) \\ = 2\Phi(z_{[1-(1-\alpha)/2]}) - 1 = \alpha, \end{aligned} \tag{10}$$

where  $\Phi$  denotes the distribution function of the standard Gaussian distribution and  $z_q$  its  $q$ th quantile (see, e.g. example 2.10 in [17]). Hence,

$$\left[ y_n \pm \sqrt{\frac{1}{nm} \sum_{i=1}^n s_i^2} \sqrt{\frac{1}{n}} z_{[1-(1-\alpha)/2]} \right] \tag{11}$$

is asymptotically an exact  $100\alpha\%$  confidence interval around the estimate produced by GUM-S1. However, (11) does not equal the  $100\alpha\%$  coverage interval produced by GUM-S1.

The distribution produced by GUM-S1 is that of the average of  $n$  scaled and shifted  $t$ -distributions with  $m - 1$  degrees of freedom, each having mean  $x_i$  and variance  $s_i^2/m \times (m - 1)/(m - 3)$  (recall that  $m \geq 4$  is assumed). The mean of this distribution equals the average of all  $x_i$ , and its variance equals  $\sum_i s_i^2/m \times (m - 1)/(m - 3) \times 1/n^2$ . Let  $\tilde{Y}_n$  denote the random variable characterized via the GUM-S1 distribution. Application of the central limit theorem yields that  $\tilde{Y}_n$  asymptotically follows a Gaussian distribution with mean  $y_n = \sum_i x_i/n$ , and variance  $\sum_i s_i^2/(nm) \times (m - 1)/(m - 3)/n$ .

Hence, the GUM-S1  $100\alpha\%$  coverage interval is asymptotically given by

$$\left[ y_n \pm \sqrt{\frac{1}{nm} \sum_{i=1}^n s_i^2} \sqrt{\frac{(m - 1)}{(m - 3)}} \sqrt{\frac{1}{n}} z_{[1-(1-\alpha)/2]} \right]. \tag{12}$$

Comparison with (11) shows that the  $100\alpha\%$  coverage interval of GUM-S1 is larger than the exact  $100\alpha\%$  confidence interval and that GUM-S1 thus overestimates the expanded uncertainty by the factor  $\sqrt{(m - 1)/(m - 3)}$ , irrespective of the chosen coverage level. From the asymptotic distribution of  $y_n|\mu$ , it follows that the long-run success rate of the GUM-S1  $100\alpha\%$  coverage interval (12) equals  $2\Phi(\sqrt{(m - 1)/(m - 3)}z_{[1-(1-\alpha)/2]}) - 1 > 2\Phi(z_{[1-(1-\alpha)/2]}) - 1 = \alpha$ . The results shown in figure 2 indicate that approximate asymptotic behavior can be reached already for  $n \approx 20$ , meaning that the long-run success rates determined for  $m = 10$  and  $n > 15$  are in accord with the asymptotic value of 0.974 obtained via the above analysis.

#### 4. Reference prior for sequence of non-linear models

The results produced by the application of GUM-S1 to data from the two sequences of non-linear models in section 2 are unsatisfactory. Figure 4 demonstrates that the estimation of the measurand can even fail completely. It is evident that the choice of non-informative prior made implicitly by applying GUM-S1 is not adequate for the particular example of figure 4. In fact, when applying the reference prior principle [18–20] to the non-linear model (4), the resulting reference posterior (see [21]) also shown in figure 4 yields excellent long-run success rates (see figure 5). The non-informative prior underlying GUM-S1 appears to simply be inadequate.

However, in order to apply the reference prior to the sequence of non-linear models, the whole sequence of models needs to be known in advance and all data must be treated

simultaneously. This is not in accordance with the rationale of the GUM to completely capture the result of one experiment in an estimate and its associated uncertainty, or in a distribution. With these constraints in mind, the reference prior principle would not have been applicable for this sequence of non-linear models. Furthermore, when considering an input quantity itself, the non-informative prior underlying GUM-S1 equals the corresponding reference prior. From figure 2, we can observe that the behavior of the inference on one of the input quantities obtained for  $n = 1$  in the sequence of linear models in (1) is excellent.

For many models, the reference prior principle yields reasonable inferences with good frequentist properties. The essential point is that, via this principle, the prior is chosen in dependence on the properties of the statistical model involved. However, since prior to the re-use of a result one does not know about future models, a Bayesian inference in combination with the reference prior principle does not guarantee transferability in metrology.

The conclusion drawn for the reference prior is valid with the same reasoning for any other principle used to select a non-informative prior that depends on the statistical model employed. On the other hand, a non-informative prior chosen on *a priori* grounds (such as the constant prior) will not perform well for all possible models, as demonstrated by the results of the constant prior for the sequence of models (4).

## 5. Discussion

Metrological tasks such as calibration or ensuring traceability require an uncertainty evaluation method whose results can safely be transferred and re-used as inputs in subsequent experiments. Bayesian uncertainty evaluation appears to be a natural choice in this regard. The prior belief about all unknowns is updated in the presence of new data by using Bayes' theorem, resulting in the posterior distribution, which then completely characterizes one's degree of belief about the measurand considered. The posterior distribution can be re-used as a prior distribution in a subsequent measurement in which the current measurand is entered as an input quantity. In this way, Bayesian uncertainty evaluation suggests itself and the paradigm shift for type A evaluation made by GUM-S1 towards the Bayesian point of view appears to be natural. However, the shortcomings of GUM-S1 observed in the context of the simple examples studied in section 2 are discouraging. In the following, we reveal the reason for the poor behavior observed for GUM-S1 and discuss possible alternatives.

### 5.1. GUM-S1 and automatic Bayes

The GUM-S1 type A uncertainty evaluation is a Bayesian inference with an automatically chosen non-informative prior. The use of non-informative priors is controversial within the Bayesian community, see arguments against and in favor of it given in [22, 23].

One motivation for using non-informative priors is to avoid the specification of an informative prior distribution. For example, it can be difficult or expensive for the analyst to specify a proper probability distribution that precisely reflects their state of knowledge. Another motivation could be that the results should depend only on the data. At the same time, the advantages of Bayesian inference, e.g. to make probability statements about the quantities of interest after the data have been observed or to re-use a posterior as a prior in a subsequent analysis, shall be shared. However, there exists no unique non-informative prior, and many different principles have been proposed for selecting a non-informative prior, see [24]. As a consequence, applying a Bayesian inference with a non-informative prior cannot be seen as expressing the analyst's degree of belief unless the number of observations is large enough that the data dominate the posterior<sup>2</sup>. Furthermore, since no unique non-informative prior exists, the posterior of an automatic Bayesian inference depends not only on the data, but generally also on the particular non-informative prior chosen. Hence, Bayesian inference using non-informative priors can be viewed as being in between a frequentist and a subjective Bayesian inference. Since the posterior cannot be strictly seen as expressing the analyst's degree of belief, and as the prior is chosen automatically in the same way whenever a model of the considered type is faced, the behavior of the resulting inference under repeated applications becomes relevant, and long-run success rates are a valid criterion for judging a non-informative prior [23, 25, 26] (see also [27] for a discussion on the success rates of Bayesian credible intervals and frequentist confidence intervals). It is for these reasons why we have looked at the performance of the GUM-S1 type A evaluation in the way described in the previous sections. This differs completely from a subjective Bayesian inference, see below.

### 5.2. Vaguely informative priors

Vaguely informative priors have become popular in metrology recently (see, e.g. [28, 29]). Vaguely informative priors are proper priors that capture available prior knowledge while preserving some sort of vagueness at the same time. For example, when identifying the quantities  $\mu_1, \dots, \mu_n$  in the sequence of non-linear models (4) with some physical quantity, the analyst could easily provide reasonable lower and upper bounds for them in a safe way, rather than letting any value, however big, be equally likely *a priori*. Let us therefore assume that  $\pm 10$  would be appropriate vague lower and upper bounds for the quantities  $\mu_1, \dots, \mu_n$ , and that the analyst chooses a uniform distribution over that interval. It happens that the results are just the same as those obtained by the GUM-S1 evaluation and that they are even insensitive when doubling or halving the width of the uniform prior. Hence, Bayesian inference in combination with vaguely informative priors can encounter

<sup>2</sup> Under mild conditions, the posterior will be dominated completely by the data when the number of observations grows without bound (see, e.g. [17]). However, such asymptotic behavior is usually not relevant in metrology, where often small samples sizes are chosen.

exactly the same pitfalls as when using non-informative priors; thus, vaguely informative priors will in general not enable a transferable uncertainty calculus either.

In fact, this conclusion can be widened to any Bayesian inference for which some vagueness remains in the selection of a prior. The reason is that one cannot be sure about the consequences of this vagueness upon subsequent re-use of the resulting posterior. Even when the posterior for a chosen vague prior is insensitive under reasonable variations of the prior, one cannot conclude that this holds for any subsequent re-use of the posterior in other models.

### 5.3. Fully subjective Bayesian inference

In a subjective Bayesian inference, the analyst specifies their prior degree of belief completely, and the resulting posterior distribution expresses their degree of belief after accounting for the information contained in the data. While the analyst will be as careful as possible when forming their prior belief, there is no guarantee that the opinion they have formed is consistent with the ground truth (see section 2.4 in [23]). Once the data are observed, the degree of belief of the analyst will be updated following probability calculus, irrespective how consistent or inconsistent data and prior belief are. Each single Bayesian inference represents a unique task characterized through the particular prior knowledge that is available for it, and the question of long-run success does not pose itself, and especially not the ‘performance’ under repeated sampling. Actually, long-run success would in a sense measure how consistent prior knowledge and underlying truth are; yet, in a real application, the truth is unknown, and the best the analyst can do is to account for all they know about the task. The observed data will then teach how their state of knowledge is updated. Again, the question of long-run success does not pose itself.

Subjective Bayesian inference provides a transferable uncertainty calculus. An assessment between degree of belief and ground truth is not seen as a well-posed question; thus, the observed shortcomings from section 2 would not arise. The posterior simply describes the degree of belief of the analyst who does their best when forming their prior belief. The posterior captures the final result without the possibility of deriving the separate influence the employed prior and the observed data had on it. From what has been said above about the lack of transferability when using vaguely informative priors, the analyst should not introduce any vagueness into their prior but rather express precisely their degree of belief. Transferability throughout metrology then requires that any analyst adopts the degree of belief of any previous analyst. Note that, in many cases, this situation already applies today for results based on type B evaluations.

## 6. Conclusions

Supplement 1 to the GUM has shifted the type A uncertainty evaluation of the GUM towards a Bayesian uncertainty

analysis. However, rather than using available prior knowledge the prior underlying GUM-S1 is an improper non-informative prior. We have demonstrated in terms of examples that, as a result of such an implicit choice of prior, the transferability of results achieved with the GUM-S1 type A evaluation approach is challenged. The lack of transferability can be generalized to Bayesian type A uncertainty evaluations utilizing other non-informative or vaguely informative priors.

## References

- [1] Joint Committee for Guides in Metrology 2008 *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement* (Sèvres: International Bureau of Weights and Measures (BIPM)) (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections)
- [2] Joint Committee for Guides in Metrology 2008 *Evaluation of Measurement Data—Supplement 1 to the ‘Guide to the Expression of Uncertainty in Measurement’—Propagation of Distributions Using a Monte Carlo Method* (Sèvres: International Bureau of Weights and Measures (BIPM)) (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008)
- [3] Elster C, Wöger W and Cox M G 2007 Draft GUM supplement 1 and Bayesian analysis *Metrologia* **44** L31
- [4] Elster C and Toman B 2009 Bayesian uncertainty analysis under prior ignorance of the measurand versus analysis using the supplement 1 to the Guide: a comparison *Metrologia* **46** 261
- [5] Kyriazis G A 2008 Comparison of GUM supplement 1 and Bayesian analysis using a simple linear calibration model *Metrologia* **45** L9
- [6] Lira I and Grientschnig D 2010 Equivalence of alternative Bayesian procedures for evaluating measurement uncertainty *Metrologia* **47** 334
- [7] Forbes A and Sousa J 2011 The GUM, Bayesian inference and the observation and measurement equations *Measurement* **44** 1422–35
- [8] Gleser L J 1998 Assessing uncertainty in measurement *Stat. Sci.* **13** 277–90
- [9] White D R 2016 In pursuit of a fit-for-purpose uncertainty guide *Metrologia* **53** S107
- [10] Hall B D 2008 Evaluating methods of calculating measurement uncertainty *Metrologia* **45** L5
- [11] Giaquinto N and Fabbiano L 2016 Examples of S1 coverage intervals with very good and very bad long-run success rate *Metrologia* **53** S65
- [12] Huang H 2019 Why the scaled and shifted t-distribution should not be used in the Monte Carlo method for estimating measurement uncertainty? *Measurement* **136** 282–8
- [13] O’Hagan A and Forster J 2004 *Kendalls Advanced Theory of Statistics: Bayesian Inference* vol 2B, 2nd edn (Oxford: Oxford University Press)
- [14] Lira I 2008 On the long-run success rate of coverage intervals *Metrologia* **45** L21
- [15] Kay S M 1993 *Fundamentals of Statistical Signal Processing* (Englewood Cliffs, NJ: Prentice Hall)
- [16] Gelman A, Carlin J B, Stern H S, Dunson D B, Vehtari A and Rubin D B 2013 *Bayesian Data Analysis* (London: Chapman and Hall)
- [17] Van der Vaart A W 2000 *Asymptotic Statistics* vol 3 (Cambridge: Cambridge University Press)

- [18] Berger J and Bernardo J M 1992 On the development of reference priors *Bayesian Statistics* vol 4, ed J M Bernardo *et al* (Oxford: Oxford University Press) pp 35–60
- [19] Berger J and Bernardo J M 1992 Ordered group reference priors with application to the multinomial problem *Biometrika* **79** 25–37
- [20] Berger J, Bernardo J M and Sun D 2009 The formal definition of reference priors *Ann. Stat.* **37** 905–38
- [21] Bernardo J M 1979 Reference posterior distributions for Bayesian inference *J. R. Stat. Soc. B* **41** 113–28
- [22] Goldstein M 2006 Subjective Bayesian analysis: principles and practice *Bayesian Anal.* **1** 403–20
- [23] Berger J 2006 The case for objective Bayesian analysis *Bayesian Anal.* **1** 385–402
- [24] Kass R E and Wasserman L 1996 The selection of prior distributions by formal rules *J. Am. Stat. Assoc.* **91** 1343–70
- [25] Datta G S and Ghosh J K 1995 On priors providing frequentist validity for Bayesian inference *Biometrika* **82** 37–45
- [26] Irony T Z and Singpurwalla N D 1997 Non-informative priors do not exist A dialogue with Jose M Bernardo *J. Stat. Plan. Inference* **65** 189
- [27] Mana G and Palmisano C 2014 Interval estimations in metrology *Metrologia* **51** 191
- [28] Attivissimo F, Giaquinto N and Savino M 2012 A Bayesian paradox and its impact on the GUM approach to uncertainty *Measurement* **45** 2194–202
- [29] van der Veen A M H 2018 Bayesian methods for type A evaluation of standard uncertainty *Metrologia* **55** 670