Evaluation of CRM homogeneity in cases of insufficient method repeatability: Comparison of Bayesian analysis with substitutes for ANOVA based estimates

Thomas P.J. Linsinger

Joint Research Centre, Retieseweg 111, Geel, Belgium

ABSTRACT

Insufficient method repeatability is a problem characterising the evaluation of certified reference materials (CRMs). In investigating the homogeneity studies of 216 certified parameters from 36 CRMs released by the European Commission’s Joint Research Centre (JRC) over the last four years, it was found that in 1/3 of the cases, the method repeatability ($s_r$) was too high to calculate the standard deviation between units ($s_{bb}$) by classical analysis of variance (ANOVA). It was also found that the application of the repeatability requirement stated in the ISO Guide 35:2017 is not feasible since it would require unrealistically low repeatability standard deviations or an impossibly high number of replicates per unit.

Evaluation of the uncertainty of homogeneity ($u_{bb}$) as evaluated by ANOVA using both the maximum of $s_{bb}$ and 0, the maximum of $s_{bb}$ and $u_{bb}$, the uncertainty hidden by method repeatability, the maximum of $s_{bb}$ and $s_{bb}/\sqrt{n}$ and Bayesian analysis, using both informative and diffuse priors, as well as the standard deviation of the unit means, were compared using simulated homogeneity studies with repeatabilities of 1–8% and $s_{bb}$ between 0.2 and 2.8%. It was found that using the maximum of $s_{bb}$ and $s_{bb}/\sqrt{n}$ as an estimate of $u_{bb}$ guards against severe underestimation but usually results in a severe overestimation of the between-unit variation. Using the maximum of ($s_{bb}$, 0) shows the least average bias but results in a severe underestimation of $u_{bb}$ in a high fraction of cases. Using the maximum of ($s_{bb}$, $u_{bb}$) limits, but does not completely eliminate cases of a severe underestimation. Also, it leads to average results biased towards high values. For the range of $s_{bb}$ and $s_r$ investigated, Bayesian analysis performed worse than max ($s_{bb}$, $u_{bb}$) in limiting severe underestimation of $u_{bb}$, but limited the average bias towards high results.

A risk-based approach to cases of insufficient method repeatability is proposed where, after evaluating the other contributions to the uncertainty of certified values, the effect of severe over- and underestimation of $u_{bb}$ is evaluated, and an appropriate approach is chosen based on this analysis.

© 2020 The Author. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Homogeneity of selected parameters within a batch of material is one of the two defining properties of reference material (RM) [1]. Although ISO Guide 30 does not indicate any specific degree of homogeneity, it, however, states that the degree of homogeneity must be “fit for its intended use in a measurement process”. For certified reference materials (CRMs), ISO 17034 [2] requires the inclusion of an uncertainty contribution from between-unit homogeneity, except when negligible compared to the other uncertainty contributions [3]. This between-unit variation, together with uncertainty contributions from characterisation and stability, leads to the assigned uncertainty of the certified value. Therefore, a quantitative assessment of the variation between different units of a CRM is necessary to either determine the variation directly or to be able to demonstrate that the variation is indeed negligible compared to other uncertainty contributions.

A homogeneity study aims to determine the variation of the certified property between the different units of a CRM batch. A quantitative assessment of homogeneity usually comprises one or more replicate measurements on a representative number of units of a material. The material tested is usually the candidate material itself but can also be a very similar material which can, depending...
on the validated processing procedures and material properties, serve as a proxy for the homogeneity of the candidate material. The variation of the certified property is a property of the CRM and independent of the chosen measurement method. The variation of the certified property can have different forms like a trend over the whole batch, a few units with outlying values, a sudden change in the certified property due to some change in the manufacturing process etc. However, in many cases, the variation of the average value of each item across the batch can be modelled as Gaussian distribution around the certified value ($\mu_{\text{CRM}}$). In this model, the average properties of the different units $\mu_i$ follow a normal distribution around $\mu_{\text{CRM}}$ with the variance $\sigma^2_{\text{bb}}$ (Equation 1).

$$\mu_i - \text{N}(\mu_{\text{CRM}}, \sigma^2_{\text{bb}})$$  \hspace{1cm} \text{Equation 1}$$

The aim of the homogeneity assessment is to estimate the true between-unit variation $\sigma^2_{\text{bb}}$. (Note that ISO Guide 35 [3] refers to the between-unit variation as "between-bottle variation", thus using the index "bb". For consistency reasons, this notation is used in this manuscript). A typical homogeneity study consists of performing $n$ replicate measurements of the certified property on $N$ units of the CRM. Any individual measurement $x_{ij}$ on unit $i$ of the CRM will give a result that depends on the true average of the batch ($\mu_{\text{CRM}}$), a potential constant method bias ($\epsilon_m$), the deviation of the value of the unit tested from the average of the batch ($\epsilon_i$) and a random error reflecting method repeatability ($\epsilon_{ij}$) (Equation 2).

$$x_{ij} = \mu_{\text{CRM}} + \epsilon_m + \epsilon_i + \epsilon_{ij}$$  \hspace{1cm} \text{Equation 2}$$

Usually, only one method is used for all measurements in a homogeneity study. This makes $\epsilon_m$ constant for the given study. The grand mean of each homogeneity study ($\mu_*$) can therefore be written as $\mu_{\text{CRM}} + \epsilon_m$, which simplifies Equation (2) to Equation (3).

$$x_{ij} = \mu_* + \epsilon_i + \epsilon_{ij}$$  \hspace{1cm} \text{Equation 3}$$

It is often assumed that the random error $\epsilon_{ij}$ is normally distributed with a mean of zero and the repeatability variance $\sigma^2_{\text{r}}$ [3]. In many cases also, $\epsilon_i$ is assumed to follow a normal distribution with an average of zero and a variation of $\sigma^2_{\text{bb}}$. The variation between units is usually small, so the repeatability variance is the same for all units, making $\epsilon_i$ statistically independent from $\epsilon_{ij}$.

The units tested are regarded as a typical sample of the whole batch and the between-unit effect $\epsilon_i$ is modelled as a random factor, which means that the basic model, which is given in Equation (3) is a random effects model. In this case, the observed variation of the measurement results ($s^2_{\text{bb}}$) comprises the variation between units ($s^2_{\text{bb}}$) and the measurement variation ($s^2_{\text{r}}$) as shown in Equation (4). In this case, $s^2_{\text{bb}}$ is an estimate of $\sigma^2_{\text{bb}}$ and $s^2_{\text{r}}$ is an estimate of $\sigma^2_{\text{r}}$.

$$s^2_{\text{bb}} = s^2_{\text{r}} + s^2_{\text{bb}}$$  \hspace{1cm} \text{Equation 4}$$

From the data of the homogeneity study, $s_{\text{bb}}$ is calculated and is then used as estimate of the uncertainty due to between-unit variation ($s_{\text{bb}}$). It should be noted that a constant method bias is irrelevant, as it affects all measurements in the same way and will therefore not contribute to the observed variation of results.

Traditionally, $s_{\text{bb}}$ is calculated from the observed variation of all results using analysis of variance (ANOVA) [3] as shown in Equation (5). The basic assumptions of ANOVA are homoscedasticity (homogeneity of variances in each group), normality (values within each group are drawn from a normal distribution) and randomness and independence (the results are random and the units are independent of each other).

$$s_{\text{bb}} = \sqrt{\frac{MS_{\text{between}} - MS_{\text{within}}}{n}}$$  \hspace{1cm} \text{Equation 5}$$

With $MS_{\text{between}}$ and $MS_{\text{within}}$ being the mean squared errors between and within groups from ANOVA and $n$ the number of replicate measurements performed per unit, $s_{\text{bb}}$ is calculated as $\sqrt{MS_{\text{within}}}$. Equation (5) is derived from an expression that substitutes the estimates $MS_{\text{between}}$ and $MS_{\text{within}}$ for the expected values of these mean squares in the equation that relates the expected variances $\sigma^2_{\text{bb}}$ and $\sigma^2_{\text{r}}$ with the mean squares. This generally works as long as $\sigma^2_{\text{bb}}$ is large compared to $\sigma^2_{\text{r}}/n$. For very homogeneous materials ($\sigma^2_{\text{bb}}$ is very small) and/or methods with a high repeatability variance ($\sigma^2_{\text{r}}/n$ is large), the argument in the equation can become negative, which makes a calculation of $s_{\text{bb}}$ impossible. Several solutions have been suggested in this case, either by replacing the estimate for $s_{\text{bb}}$ from Equation (4) with a non-negative value, by selecting a method with a sufficiently small $\sigma^2_{\text{r}}$ or by performing a sufficiently high number of replicate measurements $n$ per unit that ensures that the argument in the square root does not become negative in the first place or recently by using Bayesian analysis.

1.1. Approach 1: choosing a non-negative value

The classic response in frequentist statistics to a negative argument in Equation (5) is to set $s_{\text{bb}}$ to zero. However, if the measurement method lacked sufficient precision, this can lead to a severe underestimation of the between-unit variation. It has, therefore, been argued that the relation between $s^2_{\text{bb}}$ and $s^2_{\text{r}}$ is of no interest in reference material production [4]. Non-negative estimators for the between-group variance have been proposed by Federer [5]. Alternatively, an attempt has been made to find an approximate estimate for the inhomogeneity that could be hidden by the method repeatability (named $u^*_{\text{bb}}$). This led to an equation based entirely on the estimated method repeatability and its degrees of freedom. The degrees of freedom, in turn, depend on the total number of units tested $N$ and the number of and the number of replicates performed per unit ($n$) [6] (Equation 6).

$$u^*_{\text{bb}} = \sqrt{\frac{MS_{\text{within}}}{n}} \cdot \left(2 \cdot \frac{1}{N(N-1)}\right)^{1/2}$$  \hspace{1cm} \text{Equation 6}$$

It was suggested to use the larger value of $s_{\text{bb}}$ and $u^*_{\text{bb}}$ as an estimate of the uncertainty of homogeneity [6]. This approach is given as an option in the ISO Guide 35 [3] and is followed by several reference material producers. The rationale for the procedure is that if no inhomogeneity is detected, one should rather use the potentially undetected inhomogeneity as a conservative estimate for the undetected inhomogeneity. This approach has recently been criticised not only on fundamental terms but also as being an underestimation of the inhomogeneity that could really be hidden [7].

As an alternative and a more conservative approach, using the larger of $s_{\text{bb}}$ and $u^*_{\text{bb}}/\sqrt{n}$ as an estimate of the between-unit variation has been proposed [8]. In the case of a perfectly homogeneous material ($s^2_{\text{bb}} = 0$), the variation between unit means will entirely be determined by the random variation $s_{\text{r}}/\sqrt{n}$. In the absence of any detected variation, the variation of unit means of perfectly homogeneous material is taken as an estimate of the between-unit variation.

It should be noted that regardless of the value chosen, replacing the nominal value for $s_{\text{bb}}$ with a non-negative value introduces a bias. The rationale for using these estimates, nevertheless, is that...
such a positive bias is preferable to the possibility of a severe underestimation of the heterogeneity.

1.2. Approach 2: ensuring that the argument does not become negative

ISO Guide 35:2017 [3] and ISO 13528 [9] suggest a different solution to this problem by setting minimum requirements for the repeatability standard deviation ($s_{p}$) of the methods used for homogeneity assessment. If these requirements are met, and no variation can be calculated, $s_{p}$ is set at zero. The rationale is the minimum requirements should ensure that $MS_{within}$ - $MS_{between}$ in Equation (5) does not become smaller than zero. If $MS_{within}$ - $MS_{between}$ in Equation (5) becomes smaller than zero nevertheless, then the $s_{p}$ is so small that the very fact of a negative argument in the square root proves that the uncertainty contribution from homogeneity is negligible compared to other uncertainty contributions. To this end, ISO Guide 35 recommends that $s_{i}/\sqrt{n}$ should be smaller than 1/3 of the target uncertainty of the certified value. ISO 13528 has an equivalent criterion, albeit calculated for $n = 1$. This approach is much more established in proficiency testing, where already, the 2005 edition of ISO 13528 stipulated the same criterion but is new for reference materials, where it was first formalised in the 2017 version of ISO Guide 35. The reason might be that the target is easier to reach in proficiency testing than for CRMs: for proficiency tests, the target uncertainty is usually the size of the reproducibility standard deviation of the results submitted by the participants or larger, which makes it relatively easy to find methods that fulfill this criterion.

In contrast, for CRMs, in an ideal case, the uncertainty of the certified value is negligible compared to the measurement uncertainty of the measurement result obtained by the user of the CRM. This leads to much smaller target uncertainties than in proficiency testing. The applicability of this approach for CRMs has, therefore, not yet been tested.

1.3. Approach 3: Bayesian analysis

A third and completely different approach to solving the problem of negative arguments in the square root of ANOVA is to abandon the calculation algorithm of the classical ANOVA and use Bayesian analysis instead. It should be emphasised that the basic approach is much more established in proficiency testing than for CRMs: for proficiency tests, the target uncertainty is usually the size of the reproducibility standard deviation of the results submitted by the participants or larger, which makes it relatively easy to find methods that fulfill this criterion.

In contrast, for CRMs, in an ideal case, the uncertainty of the certified value is negligible compared to the measurement uncertainty of the measurement result obtained by the user of the CRM. This leads to much smaller target uncertainties than in proficiency testing. The applicability of this approach for CRMs has, therefore, not yet been tested.

For real materials, as the $s_{p}$ is unknown, one can only compare whether one approach gives a higher estimate for $s_{p}$ than another. The situation is even more complicated in cases where no homogeneity is detected and where it is not clear whether this is because an approach underestimates the inhomogeneity or whether the other approach overestimates it. Simulated data where an artificial introduced variation between units is known seems an effective way to circumvent this problem: Apart from mere comparison of the results of the different approaches relative to each other, comparison of the estimates with the true values also allows a more solid statement whether an approach over- or underestimates the uncertainty.

In this manuscript, two questions in connection with homogeneity studies of certified reference materials are investigated. The first part uses real data to check whether there is a significant number of homogeneity studies where $MS_{between}$ - $MS_{within}$ is negative and whether a repeatability limit as proposed by ISO Guide 35 is realistic and implementable. The second part compares different approaches to dealing with cases where $s_{p}$ is equal or smaller than $s_{i}/\sqrt{n}$, i.e. cases where it might be impossible to estimate $s_{p}$ from classical ANOVA because the data contain little information on the between-unit variation. To this end, three non-negative replacements for $s_{p}$ and the results of Bayesian analysis with informative and diffuse priors are compared simulated homogeneity studies with known between-unit variations.

2. Data and data analysis

2.1. Need for an uncertainty estimation for alternative homogeneity evaluations and practicality of a repeatability limit

Homogeneity studies of 36 different certified reference materials comprising 216 certified parameters released by the JRC between 2014 and 2018 were investigated. These materials cover a wide range of matrix/measurand combinations. They include mass fractions of trace elements in chocolate, algae and copper, particle sizes of nano- and microparticles, mass fractions of polycyclic aromatic hydrocarbons and polybrominated flame retardants in water and sediment, enzyme activities of pure enzyme
preparations and protein mass fractions in serum. A complete list of the CRMs used, the measurands investigated, and a link to the certification reports containing the data is given in the supplementary electronic materials. Most of these materials are matrix CRM, and the measurement of the analytes in question usually involves more or less complex sample preparation. Each study consists of two to four determinations of the measurand in question in 10−20 units. 117 of these studies concerned element analysis, 56 small organic molecules, 19 physical properties, 9 operationally defined properties, 11 proteins and 4 measurands based on polymerase chain reaction (PCR). In the first step, 38 studies that could not be evaluated by standard ANOVA because they contained outliers or trends in the filling sequence were excluded. While trends in the filling sequence could be included both in frequentists and Bayesian models, an investigation of such models is beyond the scope of this manuscript. The exclusion of these resulted in 181 studies that could be evaluated by standard ANOVA. The distribution of the selected studies was virtually unchanged by this exclusion: 101 elements, 43 small organic molecules, 13 physical properties, 8 operationally defined properties, 10 proteins and 4 PCR-based measurands.

For each of the studies selected, \( s_{bb} \) and \( s_r \) as calculated from ANOVA were taken from the certification reports, as were \( u_{s_{bb}} \) and \( u_{s_r} \). The number of cases where \( s_{bb} \) could not be calculated as \( MS_{between} \), was smaller than where \( MS_{within} \) was counted and the number of replicates that would have to be performed to fulfil the criterion \( s_{bb}/\sqrt{n} \leq u_{s_{bb}} \) with \( u_{s_{bb}} \) being the uncertainty of characterisation. Subsequently, these data were summarised as median values to minimise the effect of single extreme values.

2.2. Analysis of simulated homogeneity data

2.2.1. Data generation for simulated homogeneity studies

Sets of 20 homogeneity studies comprising 15 units analysed in triplicate were generated. The simulated \( s_r \) were 1%, 3%, 5% and 8% and the applied \( s_{bb} \) were 0.2%, 0.5%, 1% and 2.8%. Sets of 900 normally distributed random numbers with an average of 100 and a standard deviation of 1%, 3%, 5% and 8% were generated using the website www.random.org. These 900 numbers were grouped in 300 sets of three results, each set representing the three replicates of one unit. 300 random numbers with an average of 1 and standard deviations of 0.002, 0.005, 0.010 and 0.028 were generated using the website www.random.org. These random numbers represent the deviation of each unit from the average measured level of 100. Each of the 900 random numbers of each set was then multiplied by the variation factor to simulate the measurement results on this unit (Equation 7).

\[
x_i = z_i \times y_i \quad \text{Equation 7}
\]

\( x_i \) simulated result
\( z_i \) individual random number with \( i = 1 \) to 300 (unit number) and \( j = 1 \) to 3 (replicate number)
\( y_i \) with \( i = 1 \) to 300: fraction of each unit mean of the average of the study

The results from \( i = 1 \) to \( i = 15 \) are homogeneity study 1, \( i = 16 \) to \( i = 30 \) are homogeneity study 2 etc.

As seen in Equation (7), a multiplicative bias is applied rather than an additive one (the factor for the between-unit variation is multiplied with the random numbers for the replicates). This was done in line with the observation that for many methods, the relative variations are constant. Due to the relatively small bias (0.2−2.8% between-unit variation), the differences between multiplication and addition are negligible.

The complete set of random numbers is given in the supplementary electronic material, and an overview of the calculation is given in Fig. 1. The combination of \( s_r \) and \( s_{bb} \) used are shown in Table 1.

2.2.2. ANOVA and non-negative estimates of \( s_{bb} \)

The simulated data were evaluated using classical one-way ANOVA calculating \( S_{bb} \) and \( u^*_{s_{bb}} \). Three approaches to dealing with cases where \( MS_{between} < MS_{within} \) were applied. In one case, \( u^*_{s_{bb}} \) was estimated as the maximum of \( S_{bb} \) and 0, i.e. setting \( u_{s_{bb}} \) equal to zero in case \( MS_{between} < MS_{within} \). The second case was to set \( u_{s_{bb}} \) as the maximum of \( S_{bb} \) and \( u^*_{s_{bb}} \), i.e. setting a lower limit and the third case was to set \( u_{s_{bb}} \) as the maximum of \( S_{bb} \) and \( s_r/\sqrt{n} \).

Throughout the manuscript, the following codes are used:

\( "A-\max(S_{bb},0)" \) for evaluation by one-way ANOVA where \( S_{bb} \) was set to zero for all cases where \( MS_{between} < MS_{within} \).
\( "A-\max(S_{bb}, u^*_{s_{bb}})" \) for evaluation by one-way ANOVA where \( S_{bb} \) was set to the maximum of \( S_{bb} \) and \( u^*_{s_{bb}} \) as proposed in Ref. [6].
\( "A-\max(S_{bb}, s_r/\sqrt{n})" \) for evaluation by one-way ANOVA where \( S_{bb} \) was set to the maximum of \( S_{bb} \) and \( s_r/\sqrt{n} \) as proposed in Ref. [8].

2.2.3. Bayesian analysis

2.2.3.1. Definition of the model

Each study was evaluated using WinBUGS 1.4.3 (Imperial College and MRC, UK) using the hierarchical WinBUGS program given in Table 2. The model itself follows the model described in the introduction:

Each of the units means \( i (\mu_i) \) in the Winbugs code follows a normal distribution around the grand mean \( \mu \) (gm in the Winbugs code) with a variance \( \sigma^2 \) (Equation (11)). The individual measurement results \( z_i \) in the Winbugs code follow a normal distribution around the respective unit mean with a standard deviation of \( s_r \) as Winbugs has no integrated function for the Cauchy distribution, the fact that \( (x_i - N(0, B)) \) is a normal distribution on a single day may be worse, so the limit was increased. The prior itself is of little importance,
as sufficient data points are available to provide a good estimate of $s_r$.

$s_{bb}$: Half Cauchy distributions were used as prior for $s_{bb}$, as suggested in Ref. [12,13] and applied in Ref. [10]. Three variants of this prior were applied. The first one was a diffuse prior with a scale parameter lambda of 25. This prior allows virtually all values for $s_{bb}$ even to very high ones. Furthermore, two informative priors were applied. For nominal $s_{bb}$ of 0.2%, 0.5% and 1%, a scale parameter of 1 was applied, which severely limits the probability of high $s_{bb}$. For nominal $s_{bb}$ of 2.8%, a scale parameter of 5 was applied, which means that most values chosen for $s_{bb}$ are <5. The relevant parts of the probability density function of half Cauchy distributions with scale parameters of 1, 5 and 25 are shown in Fig. 2.

For each study, 100000 iterations were performed. A thinning of 20 was applied to reduce the autocorrelation in $s_{bb}$, resulting in 5000 valid iterations per study. In all cases, the median of the 5000 $s_{bb}$ was used as an estimate of $u_{bb}$. The choice of the median rather than the average as an estimate for $s_{bb}$ results in slightly lower values, especially in cases where $s_{bb}$ is equal or smaller than $s_r$ because in these cases, skewed distributions for the posterior are obtained. It was found that the difference is about 3 relative % in case $MS_{between} = MS_{within} = 3$ and about 20 relative % in the most extreme case where $s_{bb} = 0.2%$ and $s_r = 8%$.

Throughout the manuscript, the following codes are used:

"B-uninf" for the results of the Bayesian analysis applying a scale parameter of 25 for the half-Cauchy distribution of $s_{bb}$

"B-inf" for the results of the Bayesian analysis with an informative prior for $s_{bb}$ using scale parameters of 1 (for nominal $s_{bb}$ of 0.2%, 0.5% and 1%) and 5 (for nominal $s_{bb}$ of 2.8%), respectively.

2.2.3.2. Validation of the model

The basic model was validated by comparing the data for $s_{bb} = 2.8$ and $s_r = 1$ to the results from classical ANOVA. This is a situation where, for all data used, $s_{bb}$ can be evaluated by ANOVA, and the Bayesian analysis should give the same result. For each of the 20 tests, the ratio between the $s_r$ obtained by Bayesian analysis and by ANOVA was calculated. This ratio was on average, 1.03 reflecting the difference between median and mean.

Similarly, for each of the tests, the ratio between the $s_{bb}$ obtained by Bayesian analysis and $s_{bb,real}$ was calculated. This ratio was on average, 0.98 for the uninformative prior and 1.01 for the informed prior. This comparison shows that the model and its implementation are an appropriate implementation of the hierarchical model chosen.

2.2.3.3. Investigation of the influence of the priors

The influence of the scale parameter of the half Cauchy distribution for $s_{bb}$ is part of the manuscript, as one of the datasets generated is a diffuse prior with a scale parameter of 25 and one is an informative prior with scale parameters of 1 and 5.

The goal of this manuscript is to investigate cases where the repeatability standard deviation is larger than $s_{bb}$. The influence of the other priors was therefore investigated on a dataset with $s_{bb} = 2.6$ and $s_r = 5$, as well as on a dataset with $s_{bb} = 1%$ and $s_r = 5%$ for standard deviations of the grand mean of 10 and 3.1, including a less informative prior for $s_r$, namely a uniform distribution between 0 and 20. The results are shown in Table 3.

The results show that a less informative prior for $s_r$ has no influence on the result, neither for $s_r$ nor for $s_{bb}$. This is an expected

---

**Table 1**

| $s_r$ | $s_{bb}$ | Ratio $s_r/s_{bb}$ |
|-------|----------|---------------------|
| 1%    | 0.2%     | 5                   |
| 1%    | 0.5%     | 2                   |
| 1%    | 1%       | 1                   |
| 2.8%  | 0.2%     | 15                  |
| 0.5%  | 1%       | 6                   |
| 1%    | 2.8%     | 1                   |
| 5%    | 0.2%     | 25                  |
| 0.5%  | 5%       | 10                  |
| 1%    | 1%       | 5                   |
| 2.8%  | 2.8%     | 1.67                |
| 5%    | 0.2%     | 40                  |
| 0.5%  | 5%       | 16                  |
| 1%    | 1%       | 8                   |
| 2.8%  | 2.8%     | 2.67                |

---

Fig. 1. Overview of the data generation of the simulated homogeneity studies.
outcome since the number of individual results is high enough to dominate the posterior. Also, a reduction of the standard deviation of the grand mean does not change the result, as long as the standard deviation, which is chosen is larger than the combination of $s_{bb}$ and $s_r/\sqrt{n(\frac{u_{cc}}{4}u_{cc}$ of Equation(4)): A standard deviation of 10 does not change the results but a standard deviation of 3.1, in combination of an $s_{bb}$ of 2.6 and an $s_r$ of 5 would result in a lower estimate for $s_{bb}$.

Based on this, it is concluded that the priors are adequate for the purpose.

### 2.2.4. Comparison of approaches

Real between-unit standard deviation ($s_{bb,real}$): For each simulated study, the real between-unit standard deviation of the study, defined as the standard deviation of the 15 $y_i$ of the respective study was calculated. This standard deviation was multiplied with 1.018, the correction factor for 15 data points to obtain an unbiased estimate [14].

For each study, the ratio between the estimated $s_{bb}/s_{bb,real}$ was calculated, and the following parameters were derived:

- **Average estimate**: For each combination of $s_{bb,real}$ and $s_r$, the average of the 20 ratios $s_{bb}/s_{bb,real}$ was plotted against the ratio $s_r/s_{bb,real}$. For a perfect evaluation method, this ratio would be one for all ratios $s_r/s_{bb,real}$—in practice, even for a completely unbiased evaluation, this ratio will scatter around 1 due to the effect of $s_r$ on the estimate. Systematic deviations from unity indicate a bias in the evaluation approach.

- **Maximum overestimation**: For each combination of $s_{bb,real}$ and $s_r$, the upper single-sided 95% confidence limit was calculated from the 20 ratios $s_{bb}/s_{bb,real}$. This parameter accounts for the fact that users are affected by an unnecessarily large uncertainty of the CRM they purchased. Whether an evaluation performs well on average is of minor importance. The maximum overestimation was plotted against the ratio $s_{bb,real}/s_r$.

---

**Table 2**

WinBUGS® Program for the evaluation of homogeneity studies.

```plaintext
model
{
  for (i in 1:15) # i = unit number from 1 to 15
  {group.mu[i] ~ dnorm(gm, tbb) # gm = grand mean; tbb is precision parameter for sbb - 1/sbb^2}
  for (i in 1:45) # for each of the 45 results of each study
  {z[i] ~ dnorm(mu[i], prec); # individual results follow a normal distribution around the mu[i] <- group.mu[unit[i]]}
}
# the priors
sr ~ dunif(0,10) # sr= repeatability standard deviation; between 0 and 10 %
gm ~ dnorm(100,0.001); # grand mean (gm) is 100
prec <- pow(sr, 2)
x ~ dnorm(0,tau.xi) # half Cauchy distribution as multiplication of a normal distribution and a gamma
distribution (0.5,0,5)
tau.xi <- pow(A,-2) # A is the scale factor of the half Cauchy distribution
tau.eta ~ dgamma(0.5,0.5)
sbb <- abs(xi)/sqrt(tau.eta) # sbb is the relative between unit standard deviation
tbb <- pow(sbb,-2)
}
z[1] unit[] # list of results z = result unit = unit number
93.43698896 1
99.11677964 2
97.84209916 2
97.37720806 15
END
```

---

**Table 3**

Results of the influence of various priors. $SP$ = scale parameter of the half Cauchy distribution. In both cases the informative priors were used. $s_{GM}$ = standard deviation of the grand mean.

| $s_{bb}$ | Prior $s_{GM}$ | Prior $s_r$ | $s_{bb}$ | $s_r$ |
|---------|----------------|-------------|----------|-------|
| $s_{bb}$ = 2.8 and $s_r$ = 5 | 5 | 0–10 | 3.272 | 5.961 |
| | 10 | 0–10 | 3.246 | 5.971 |
| | 3.1 | 0–10 | 3.154 | 5.990 |
| | 31 | 0–20 | 3.271 | 5.967 |
| | 31 | 0–10 | 6.114 | 1.243 |
| | 10 | 0–10 | 6.119 | 1.213 |
| | 3.1 | 0–10 | 6.108 | 1.210 |
| | 31 | 0–20 | 6.117 | 1.189 |

---

**Fig. 2.** Probability density functions of half-Cauchy distributions with a scale parameter of 1 (solid line) 5 (dashed line) and 25 (dotted line).
Maximum underestimation: For each combination of $s_{bb,real}$ and $s_r$, the lower single-sided 95% confidence limit based was calculated from the 20 ratios $u_{bb}/s_{bb,real}. As nobody will ever introduce a negative between-unit standard deviation, the lower confidence interval was capped at zero, meaning that values < 0 were replaced by 0. The rationale for this parameter is the same as for the maximum overestimation, namely that a user is affected by an underestimation of the uncertainty of a specific CRM bought and not by an average bias. The maximum underestimation was plotted against the ratio $s_{bb,real}/s_r$.

Fraction of significant underestimations: For each combination of $s_{bb,real}$ and $s_r$, the fraction of the studies with $u_{bb} < 0.7s_{bb,real}$ was calculated. Whereas the maximum overestimation and maximum underestimation investigate extreme events, this parameter quantified the frequency where $u_{bb}$ is severely (here defined as more than 30%) underestimated.

3. Results and discussion

3.1. Need for an uncertainty estimation for alternative homogeneity evaluations and practicality of a repeatability limit

The results of the compilation of homogeneity studies are given in Table 4. In 13–46% of the cases, it was impossible to calculate $s_{bb}$ as $M_{between}/M_{within}$. For method defined properties and PCR based methods, $s_{bb}$ could not be calculated in 13 and 17% of the cases, respectively. For all the other fields, the fraction of studies where $s_{bb}$ could not be calculated ranged from 30 to 46%. This is not because of the selection of inferior methods that show high repeatability standard deviations. The median repeatability standard deviation for elements is 3.4% and for small organic molecules, 5.0%, which is close to the optimum that good laboratories can achieve for a larger number of samples. Smaller repeatability standard deviations for elements can be achieved by, e.g. thermal ionisation isotope-dilution mass spectrometry of elements. However, this method is generally not suitable for the determination of the 20–40 samples required for homogeneity studies. The reason for not being able to calculate $s_{bb}$ is rather the good homogeneity of the materials, where median $s_{bb}$ ranges from 0.5 to 1.8%. This means that the problem of negative arguments under the square root in ANOVA is real and present in a large fraction of homogeneity studies of matrix materials.

Despite the good repeatabilities of the methods used in the studies, very few of them would have met the criterion for method repeatability stipulated in ISO Guide 35. The median number of replications to fulfill this requirement for chemical analyses ranges from 11 to 40. Physical properties form an outlier on the low side, as the methods often are exceptionally repeatable and PCR-based measurement forms an outlier on the high side, as here, the assigned values often are not based on PCR itself. Consequently, the requirement combines rather small $u_{char}$ with the large standard deviations from PCR. Clearly, performing 10 to 40 replicate measurements on each unit in a homogeneity study is not practical and often impossible due to the limitation of sample per unit. Even if the amount of sample per unit was sufficient to perform this number of measurements and also, if one could commit the resources needed for this large number of measurements (several hundred measurements per study), it is easily possible that the criterion would still not be fulfilled as instrument drift, the time required for this number of sample preparations etc. could increase the observed repeatability standard deviation.

These data show that the repeatability criterion stated in ISO Guide 35 is not achievable for a wide range of matrix CRMs, as either unrealistically small repeatability standard deviations would have to be achieved or an impractically high number of replicates would have to be performed.

3.2. Comparison of the different evaluation approaches on the simulated homogeneity studies

3.2.1. Average estimate

The plot of the average estimates of the five evaluation approaches is shown in Fig. 3. The graphs show that using $A_{max}(s_{bb}, s_r/\sqrt{n})$ is for high $s_{bb,real}$ ratios, which is significantly biased towards high results. This is expected and known, especially for higher $s_{bb,real}$ ratios where the standard deviation of the means only reflects the random variation of the averages due to repeatability. Also, as expected, $A_{max}(s_{bb,0})$ shows a very low average bias. This is because for a perfectly homogeneous material ($s_{bb,real} = 0$) in half of the cases, $M_{between} < M_{within}$ and the value 0 is in these cases the correct estimate. Using Bayesian statistics with an informative prior ($B$-inf) performs on average, similarly well as $A_{max}(s_{bb,0})$. The on the average high positive bias of Bayesian analysis with a diffuse prior is somewhat surprising. On average, the positive bias is comparable to $A_{max}(s_{bb,0}u_{*bb})$, which only reflects measurement repeatability and the study setup. As is evident from the equations, the average positive bias for $A_{max}(s_{bb,0}u_{*bb})$ is smaller than for $A_{max}(s_{bb,0})$ due to the inclusion of the term reflecting the degrees of freedom of $s_r$.

The finding that the results from Bayesian analysis (both using diffuse and informative priors) are lower than when using the maximum of $s_{bb}$ and $u_{*bb}$ seemingly contradicts [10]. This publication stated, for the examples investigated that "Bayesian analysis gives a value for the between-unit standard deviation, which is not even close to $u_{*bb}$. Several factors explain this apparent contradiction: On the one hand, the ratios of $s_r/s_{bb}$ explored in Ref. [10] range from 0.4 to 3.4 which is significantly lower than the ranges explored here and in the majority (7 of 9) cases, $s_{bb}$ was lower than $s_r$. In addition, 6 replicates per unit were performed. This means, as shown by the fact that $s_{bb}$ could be evaluated in all cases by ANOVA, that there was not a single case where $u_{*bb}$ was larger than $s_{bb}$. A comparison of the estimate for $s_{bb}$ and $u_{*bb}$ is not relevant, as $u_{*bb}$

### Table 4

Summary of homogeneity studies investigated.

| Measurand             | Total number of studies | Number of studies where $s_{bb}$ could not be calculated | Median $s_r$ | Median $s_{bb}$ | Median $u_{*bb}$ | Median $u_{char}$ | Median n required for the ISO Guide 35 criterion |
|-----------------------|-------------------------|----------------------------------------------------------|-------------|-----------------|------------------|------------------|-----------------------------------------------|
| Element               | 101                     | 39                                                       | 3.40        | 1.00            | 0.84             | 2.30             | 22                                            |
| Small organic molecule| 43                      | 14                                                       | 5.04        | 1.80            | 1.51             | 4.33             | 11                                            |
| Physical properties   | 13                      | 6                                                        | 0.31        | 0.49            | 0.12             | 0.57             | 3                                             |
| Method defined        | 8                       | 1                                                        | 0.88        | 0.58            | 0.29             | 0.65             | 12                                            |
| Biomolecule           | 10                      | 3                                                        | 2.34        | 0.96            | 0.62             | 0.66             | 40                                            |
| PCR                   | 4                       | 0                                                        | 11.57       | 4.84            | 2.26             | 4.64             | 111                                           |
| Sum                   | 179                     | 63                                                       | 3.28        | 1.11            | 0.90             | 2.71             | 19                                            |
should not be used as an estimate of $u_{bb}$ in such cases.

3.2.2. Maximum overestimation

In general, both CRM producers and users are less interested in the average bias than in the bias introduced for a specific CRM. The evaluation of the maximum overestimation by the five approaches is shown in Fig. 4. The general trend for the maximum overestimation is the same as for the average estimate discussed before: $A_{\text{max}}(s_{bb}, s_r/\sqrt{n})$ results in the highest and $B_{\text{inf}}$ in the lowest maximum overestimation. However, it is noticeable that the differences between the various approaches is less pronounced than in the average estimation. Especially for $s_r/s_{bb}$ ratios below 6, the results from the various approaches are rather close together. This can be explained by the fact that $s_r$ also generates a random variation between units. Whenever this variation coincides with the real between-unit variation (i.e. the average of units with a higher average increases even more by random fluctuations), a too high estimate for $s_{bb}$ will be given and cannot be reduced by any evaluation approach. The severity can only be mitigated by performing more measurements per unit or by measuring more units: more measurements per unit makes the measured unit mean a better estimate of the real unit mean while measuring more units reduces the probability that this enhancement occurs for all units.

3.2.3. Maximum underestimation

The maximum underestimation observed for the homogeneity studies investigated is shown in Fig. 5. In this graph, values below 1 show that a particular approach is expected to underestimate $s_{bb,\text{real}}$ for a certain $s_r/s_{bb,\text{real}}$ combination in 5% of the cases. Values above 1 show that for a given $s_r/s_{bb,\text{real}}$ combination $u_{bb}$ should in 95% of the cases overestimate $s_{bb,\text{real}}$. The graph shows that using $A_{\text{max}}(s_{bb}, s_r/\sqrt{n})$ is of all the investigated approaches, the least
prone to severe underestimations of $s_{bb,real}$. The minimum ratio $u_{bb}/\sqrt{s_{bb,real}}$ was 0.61, i.e. an underestimation by only 39%. This safe estimate, of course, comes at a price: Not only is the average result biased (see above), but also from $s_{bb,real}$ ratios of 1 or below was still overestimate the real value still overestimate $s_{bb,real}$ by a factor of about two or more.

The other extreme is as expected, A-max($s_{bb,0}$): Only for $s_{bb,real}$ of 1 or below was $MS_{between}$ for all 20 simulated studies, which are larger than $MS_{within}$, leading to small to modest underestimations. However, at higher ratios for $s_{bb,real}$, $MS_{between}$ was at least in one of the 20 simulated studies smaller than $MS_{within}$, which resulted in an estimate of $u_{bb}$ of 0 and hence a complete underestimation of $s_{bb,real}$. This results in lower confidence limits of 0 for all $s_{bb}/s_{r}$ ratios above 1.7.

The lower confidence limit for A-max($s_{bb,0}$) is between 0.4 and 0.6 ratios of $s_{bb,real}$ below 5, indicating that underestimations by 40–60% are possible. The notable exception is the combination, $s_{bb} = 3$ and $s_{r} = 8$, where the lower confidence limit is only 0.07, corresponding to an underestimation of 93%. However, it should be pointed out that a single overestimation of $s_{bb}$ by a factor of 3 causes this lower confidence limit, which increased the standard deviation and hence resulted in broad confidence bands. The lowest value observed was 0.64 (corresponding to an underestimation of 36%). In most of the $s_{bb,real}$ combinations, the calculated $s_{bb}$ was still larger than $u_{bb}^*$. This means the underestimation was caused by the compensation of the between-unit variation involving the random fluctuations due to measurement repeatability. In all of these cases, the underestimation was still slightly less severe than when applying Bayesian statistics (see below). The findings that $u_{bb}^*$ sometimes underestimates $s_{bb,real}$ partly supports the criticism of the approach raised in Ref. [7] that $u_{bb}^*$ may underestimate the potentially hidden homogeneity. While these simulated data confirm this worry, they also indicate that this is only a problem for low $s_{bb,real}$ combinations, i.e. for either very inhomogeneous materials or very repeatable methods. The maximum underestimation observed in these cases was 50%, which can have a substantial impact on the assigned uncertainty (see section “Potential impact of the under/overestimation of $u_{bb}$”). On the other hand, for all $s_{bb,real}$ ratios larger or equal than 8, even the lower confidence limit of A-max($s_{bb,0}$) is an overestimation of the $s_{bb,real}$, which is driven by $u_{bb}^*$. This indicates that for higher ratios of $s_{bb,real}$, $u_{bb}^*$ is a conservative estimate of $s_{bb,real}$.

As the between-unit variation is in many of the investigated scenarios hidden by the random variation, the performance of Bayesian analysis with respect to the maximum underestimation is mixed. At $s_{bb}/s_{bb,real}$ ratios above 20, using the diffuse prior results in all cases in moderate overestimations, whereas the use of the informative prior does not lead to an overestimation. On the other hand, using a point estimate for $s_{bb}$ from Bayesian analysis cannot prevent significant underestimation of $u_{bb}$ either. In all cases, the maximum underestimations of $u_{bb}$ of the Bayesian approaches were worse than A-max($s_{bb,0}$): $u_{bb}$ was underestimated by up to about 65% by both informative and diffuse priors. The reason here is the opposite of the one discussed under “Maximum overestimation”: Whenever the random fluctuation of unit means, caused by method repeatability, compensates for the real variation of unit means, $u_{bb}$ will be underestimated. As in the case of maximum overestimation, also the occurrence of underestimation can only be reduced by either performing more measurements per unit or performing measurements on more units.

As noted above, in cases where $s_{bb}/s_{r} > 1$, the posterior distributions from Bayesian analysis are very broad, and it could be argued that using the median is a poor point estimate for $s_{bb}$. However, investigation of other point estimates is beyond the scope of this manuscript.

### 3.2.4. Fraction of significant underestimations

The fractions of cases where $u_{bb}$ was underestimated by more than 30% are shown in Fig. 6. As expected from the discussion above, the extremes are again formed by A-max($s_{bb,0}$) and A-max($s_{bb,0}$, $s_{r}/\sqrt{n}$): Using A-max($s_{bb,0}$, $s_{r}/\sqrt{n}$) resulted in none of the 320 simulated homogeneity studies in an underestimation of $u_{bb}$ by more than 30%. This confirms the suggestion made in Ref. [8] that the $s_{max}$ is a safe estimate of $u_{bb}$. On the other hand, using max($s_{bb,0}$) as an estimate for $u_{bb}$ resulted in half of the cases for $s_{bb,real}$ of 0.2 and 0.5 of underestimations of at least 30%. The outcome was slightly better for $s_{bb,real} = 1$, but even there, 35–45% of $u_{bb}$ were underestimated by more than 30% when $s_{r} \geq 3$. 

---

**Fig. 5.** Maximum underestimation of $s_{bb,real}$. Left: all $s_{bb,real}$ ratios. Right: only area for $s_{bb,real}$ up to 16. Numbers below 1 correspond to an underestimation of the real between-unit variation; values above 1 show that even the lowest estimate of $u_{bb}$ was above $s_{bb,real}$. 

---
This shows that using $\max(s_{bb}, 0)$ as an estimate for $u_{bb}$ results, in many cases, in significant underestimations of the uncertainty unless the requirement for method repeatability of ISO Guide 35 is met. It should be noted that the respective section of ISO Guide 35 (section 7.7.3) refers to this requirement, albeit only in a note. Seeing the high probability of underestimating the assigned uncertainty, making this requirement more prominent would seem a reasonable suggestion for a future revision of this guide.

A-$\max(s_{bb}, u_{*bb})$ resulted in 11 of the 320 simulated studies in underestimations of $u_{bb}$ by at least 30%. This low rate is because the studies investigated have rather high ratios of $s_{bb}/s_{bb,real}$ where at high $s_{bb}$, $u_{*bb}$ exceeds the assumed $s_{bb,real}$. The situation would have been less positive for lower ratios of $s_{bb}/s_{bb,real}$ where $s_{bb}$ can usually be calculated, but $u_{*bb}$ becomes too small to account for the potentially hidden inhomogeneity. This rather low fraction of severe underestimations relativizes the concern of $u_{*bb}$ not being a sufficiently conservative estimate of the potentially hidden inhomogeneity raised in Ref. [7]: while the data confirm that $u_{*bb}$ may underestimate the true hidden inhomogeneity, they also indicates that the underestimation is limited and rather rare for the conditions studied.

Using Bayesian analysis does not mitigate the risk of severely underestimating $u_{bb}$. There was hardly any difference in the percentage of significant underestimation between the use of informative and diffuse priors. For some combinations of $s_{bb}/s_{bb,real}$, $u_{bb}$ was underestimated by more than 30% in 40% and more of the studies. For all combinations of $s_{bb}/s_{bb,real}$ investigated, Bayesian analysis performed worse than A-$\max(s_{bb}, u_{*bb})$ using this measure, with underestimations in 26 (diffuse prior) and 54 (informative prior) of the 320 simulated studies.

### 3.2.5. Potential impact of the under/overestimation of $u_{bb}$

The discussion above focussed on over- and underestimation of $u_{bb}$. However, $u_{bb}$ in isolation is only relevant for non-certified reference materials where the material information sheet must indicate the degree of homogeneity. For certified reference materials, $u_{bb}$ is always combined with the $u_{char}$ and the uncertainties of long-term and short-term stability ($u_{LTS}, u_{UTS}$ estimated either from previous knowledge or from dedicated stability studies, as described for example in Ref. [15]) to obtain $u_{CRM}$, which is calculated as shown in Equation (8).

\[
u_{CRM} = \sqrt{u_{bb}^2 + u_{char}^2 + u_{LTS}^2 + u_{UTS}^2}.
\]  

\[\text{Equation 8}\]

Depending on the sizes of $u_{char}, u_{bb}, u_{LTS}$ and $u_{UTS}$ relative to each other, an over- and underestimation of $u_{bb}$ may have a significant effect or no effect at all on $u_{CRM}$. In investigating the effect of a false estimation of $u_{bb}$ on certified uncertainties in a real-life situation, the effect of an increase/decrease of the $u_{bb}$ used for the certified values investigated in the section “Need for an uncertainty estimation” by 30% and 50% was investigated. In this case, $u_{LTS}, u_{UTS}$ were not included in the calculation as different CRM producers use widely different approaches to account for these uncertainties. The results, depicted in Fig. 7, seems at first glance to indicate that the effects of overestimation of $u_{bb}$ are more pronounced than of underestimation. This is because in most cases, $u_{bb}$ is smaller than $u_{char}$ so a further reduction has hardly any effect. For the certified values investigated here, even changes of $u_{bb}$ by ± 50% have a rather small effect. For more than 75% of the values, the change of $u_{CRM}$ is below 10%. In reality, the effect is even smaller as the JRC has, in line with the requirements of ISO 17034, the policy to include quantitative estimates of $u_{bb}$ and $u_{LTS}$ into $u_{CRM}$. This seemingly higher importance of overestimation, however, needs to be contrasted with the rounding of certified uncertainties. The policy for CRMs produced at the JRC is never to round down uncertainties (e.g. 0.112 is rounded up to 0.12) and to round uncertainties such that the rounding error is within 3 and 30% of the uncertainty. This means that for the cases studied, an overestimation of $u_{bb}$ by 30% would virtually never have resulted in an overestimation of $u_{CRM}$ by more than the accepted rounding error. Also, an overestimation of $u_{bb}$ by 40% would only in a few cases have led to such an overestimation. On the other hand, underestimation of $u_{bb}$ even by 30%, would have in 12% of the cases resulted in underestimations by more than 15%.
4. Conclusions

Given the results of the evaluation of real and simulate homogeneity studies, the following conclusions are taken: Even using state of the art methods, the actual homogeneity of matrix reference materials, combined with the limitations on the amount of material per unit and the resources that can be put into a single homogeneity study, estimation of \( u_b \) by one-way ANOVA will in many cases not be possible. Also, the limit for method repeatability as given in ISO Guide 35:2017 is, in many cases, impossible to meet. There is, therefore, a real need to find reliable approaches to deal with cases where \( MS_{\text{between}} < MS_{\text{within}} \). While all the approaches investigated in this work have their specific strengths and shortcomings, it must be borne in mind that this study deliberately investigated cases where the analytical noise overpowers the signal from the between unit variation, hence examining cases where there was not enough “data” in the numbers to perform good statistics.

Setting \( u_b \) zero when \( MS_{\text{between}} < MS_{\text{within}} \) results in many cases in severe underestimations of \( u_b \). It seems appropriate that future revisions of ISO Guide 35 should strongly emphasise the fact that this estimate is only valid if the repeatability limit is fulfilled, otherwise the maximum of \( (s_b)_0 \) is not a suitable estimate for the between-unit variation.

Estimating \( u_b \) as A-max\((s_b, s_i/\sqrt{n})\) was, as found in all cases investigated here, to guard against severe underestimations of \( u_b \). This, however, comes at the cost of severe overestimations at already rather moderate ratios of \( s_b/s_i \), which diminishes the value of the CRMs for the users.

Estimating \( u_b \) as the maximum of \( (s_b, u^{*}_b) \) was found in some cases, especially at low \( s_b/s_i \) ratios to underestimate \( s_b \), hence confirming the suspicion raised in Ref. [7]. On the other hand, it mitigated the occurrence and extent of severe underestimations. Also, this advantage comes at the price at on average positive biases towards too high values, albeit lower ones than seen for A-max\((s_b, s_i/\sqrt{n})\).

Using the median as point estimates from Bayesian analysis is not a universal solution to the problem of estimating \( s_b \) when \( s_i > s_b \), because neither using informative nor uninformative priors prevented severe underestimations of \( u_b \). It might be that other point estimates could distribution mitigate the risk of underestimations. This is an interesting option, which should be investigated further. The use of informative priors, however, limit

the positive bias of the estimation.

A solution to cases where \( s_i > s_b \) could be implementing a risk-based approach as advocated in ISO 17034: After estimating \( u_b \) (by whatever approach), \( u_{\text{bb}}, u_{\text{char}} \) and \( u_{\text{CRM}} \) it can be determined what the effect of a severe over- or underestimation of \( u_{\text{bb}} \) on \( u_{\text{CRM}} \) could have. If \( u_{\text{bb}} \) is a minor component, a large relative bias of \( u_{\text{bb}} \) will not have a significant effect on \( u_{\text{CRM}} \). If \( u_{\text{bb}} \) is significant, the reference materials producer can (preferably, considering past experience) decide which risk of over/underestimation the actual uncertainty is acceptable and choose an evaluation approach that limits this effect.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

**Thomas P.J. Linsinger:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.acax.2020.100049.

References

[1] ISO Guide 30:2015 Reference Materials – Selected Terms and Definitions, International Organization for Standardisation, Geneva, 2015.
[2] ISO 17034:2016 General Requirements for the Competence of Reference Material Producers, International Organization for Standardisation, Geneva, 2016.
[3] ISO Guide 35:2017 Reference Materials – Guidance for Characterization and Assessment of Homogeneity and Stability, International Organization for Standardisation, Geneva, 2017.
[4] J. Pauwels, A. Lamberty, H. Schimmel, Homogeneity testing of reference materials, Accred Qual. Assur. 3 (1998) 51–55.
[5] W.T. Federer, Non-negative estimators to variance components, J. Roy. Stat. Soc. C 17 (1968) 171–174.
[6] T.P.J. Linsinger, J. Pauwels, A.M.H. van der Veen, H. Schimmel, A. Lamberty, Homogeneity and stability of reference materials, Accred Qual. Assur. 6 (2001) 20–25.
[7] J. Pauwels, A. Lamberty, H. Schimmel, Homogeneity testing of reference materials, Accred Qual. Assur. 3 (1998) 51–55.
[8] W.T. Federer, Non-negative estimators to variance components, J. Roy. Stat. Soc. C 17 (1968) 171–174.
[9] T.P.J. Linsinger, J. Pauwels, A.M.H. van der Veen, H. Schimmel, A. Lamberty, Homogeneity and stability of reference materials, Accred Qual. Assur. 6 (2001) 20–25.
[10] S.L.R. Ellison, Homogeneity studies and ISO guide 35:2006, Accred Qual. Assur. 20 (2015) 495–500.
[11] S.L.R. Ellison, S. Burke, R.F. Walker, K. Heydorn, M. Mansson, J. Pauwels, W. Wegscheider, B. Njenhuis, Uncertainty for reference materials certified by interlaboratory study: recommendations of an international study group, Accred Qual. Assur. 6 (2001) 274–277.
[9] ISO 13528:2015 Statistical Methods for Use in Proficiency Testing by Inter-laboratory Comparison, International Organization for Standardisation, Geneva, 2015.

[10] A. van der Veen, Bayesian analysis of homogeneity studies in the production of reference materials, Accred Qual. Assur. 22 (2017) 307–319.

[11] ISO/IEC Guide 98-3:2008/Suppl 1:2008 Uncertainty of Measurement — Part 3: Guide to the Expression of Uncertainty in Measurement (GUM:1995) — Supplement 1: Propagation of Distributions Using a Monte Carlo Method, International Organization for Standardisation, Geneva, 2008.

[12] A. Gelman, Prior distributions for variance parameters in hierarchical models, Bayesian Anal. 1 (2006) 515–534.

[13] N.G. Polson, J.G. Scott, On the half-Cauchy prior for a global scale parameter, Bayesian Anal. 7 (2012) 887–902.

[14] B.W. Bolch, More on unbiased estimation of the standard deviation, Am. Statistician 22 (1968) 27.

[15] T.P. J Linsinger, J. Pauwels, A. Lamberty, H. Schimmel, A.M.H. van der Veen, L. Siekmann, Estimating the uncertainty of stability for matrix CRMs, Fresen. J. Anal. Chem. 370 (2001) 183–188.