Euromet project 691 Calibration Inter-comparison of a 5 litre Volume Standard

Peter Lau

SP Technical Research Institute of Sweden

Measurement Technology

SP Report 2007:10

Euromet project 691 Calibration Inter-comparison of a 5 litre Volume Standard

Peter Lau

Abstract

Euromet project 691 Calibration Inter-comparison of a 5 litre Volume Standard

A calibration was performed on a 5 litre volume standard made of glass in form of an inter-comparison between 20 European laboratories. This device is usual in many laboratories. The purpose was to validate the calibration services by looking on the agreement between the participants and comparing the degree of equivalence to both the laboratories stated measurement uncertainty in this calibration and their declared calibration measurement capability claims for this type of work. Although the agreement in volume data is fully acceptable some of the specified uncertainties are not consistent with the outcome and seem too optimistic. There are obvious differences in the experimental skill and in the judgement of measurement uncertainty.

Key words: Calibration, volume standard, inter-comparison, measurement uncertainty.

SP Sveriges Tekniska Forskningsinstitut
SP Technical Research Institute of Sweden

SP Report 2007:10
ISBN 91-85533-77-7
ISSN 0284-5172
Borås 2007

Contents

Summary

The defined inter-comparison task was to perform a volume calibration of a circulated volume standard with a ring mark in the neck. From 20 participating laboratories 18 results were received for the contained volume below the ring mark and another 16 for the volume that can be delivered after filling. From those 34 results only one lay just outside a 95 % confidence range of the corresponding reference value. The largest deviation from the reference value, but still within the 95 % confidence range, was 0,024 %. Despite this agreement clear differences are seen between the laboratories and some of the results are not consistent when uncertainties are considered. Eleven results deviated more from the respective reference value than what should be expected from the stated uncertainties in the measurement and 5 results lay outside the calibration measurement capability claims of the laboratories. There is a potential to harmonize the uncertainty estimations, not only in size but also in structure.

Six laboratories have performed this exercise on the same standard several years earlier. At average their repeatability is better now. Their inter-laboratory spread has decreased for one, but increased for the other calibrated volume. But the agreement concerning the uncertainty statements has become clearly better. The agreement is also considerably better between those than between the other laboratories, of which several have not taken part in an inter-comparison before.

1 Introduction

This inter-comparison with the Euromet project number 691was initiated during the Euromet Flow meeting 2002 in Prague. The proposal was sent to all members of the TC-flow and from start 16 participants were interested to participate. Another four laboratories willing to join this inter-comparison after its official start were accepted and could be added at the end so that all together 20 laboratories from 19 countries could take part.

This project is a repetition of one of the first inter-comparisons within Euromet (project 51 from 1988) and of the current participants six laboratories had made a calibration on this object before. This inter-comparison was performed in close cooperation with the Euromet-project 692 that was started simultaneously concerning the volume inter-comparison of a 100 mL pycnometer.

2 The Measurement Task

The purpose of the inter-comparison was to verify experimental methods and results in volume calibration using a representative object for different kinds of glassware. In contrast to key-comparisons, where the technical protocol is of crucial importance for a reference value close to the definition of a quantity, here the task was “only” to perform the calibration as usual. But the outcome is of course to be seen on the background that all laboratories were in the state of specifying their CMC-tables, i.e. their Calibration Measurement Capability.

For several participants this was their first participation in an inter-comparison. This exercise fills the gap between two volume standards a 100 mL pycnometer and a 20 L pipette in a key comparison CCM FF:K4.

2.1 The Transfer Standard

The transfer standard is a 5 litre glass flask with a ring mark shown in figure 1. It is a representative for different sizes of glassware usually used in laboratories for producing liquids with a certain concentration of a substance and constitutes a usual calibration object for several, but not all of the participating laboratories.

In contrast to other glassware like pipettes and pycnometers used in laboratories this standard can be characterized by two volumes

- the “contained” or “dry” volume up to the ring mark
- the “delivered” or “wet” volume that can be poured out

Of 20 participants 18 laboratories calibrated the contained and 16 the delivered volume. 14 laboratories determined both volumes.

Figure 1. The calibration object - a 5 litre volume standard of borosilicate glass

2.2 Questions directed to the Inter-comparison

The selected object is not a high precision standard and the expected uncertainties involved with its calibration do not allow fundamental metrological statements compared to a key-comparison. The purpose is a more practical one – to give some participants their first possibility to compare their calibration work and get a feed back how they accomplish in the comparison with other national laboratories. The volume determination of a glass flask has many similarities with larger volume standards using sight glasses. The findings gained here can therefore have relevance for volume calibrations in general and to some extend for volumetric flow calibration. The questions this experiment is aimed to answer can be formulated in the following way.

1 Do all laboratories manage to perform a calibration as expected by demanding customer?
2 Are the results comparable in the sense that the specified volumes experimentally define the same quantity? (definitely not self evident)
3 Do the laboratories differ in the used method or in the experimental skill?
4 Are the different results equivalent concerning the inter-laboratory spread?
5 Are the uncertainty claims comparable in size?
6 Do the laboratories differ in their way to estimate their measurement uncertainty?
7 Do they have different perspectives which the important uncertainties are?
8 Which are the dominating sources for the stated uncertainties?
9 Are the stated uncertainties consistent with the delivered results?
10 Are the CMC-claims supported by the outcome?
11 Do we need to reconsider some of our CMC-statements concerning uncertainty in volume determination.
12 Is the standard stable?
13 How do the laboratories performing the calibration for the second time reproduce their earlier results?
14 Is there possibly a better overall agreement between those laboratories than earlier?

2.3 The Calibration Preconditions

The outcome of an inter-comparison depends on the care with which such a calibration object is designed and fabricated. Its stability is of course of equal importance. A further influence can arise from the given instructions that accompany the transfer standard as it probably will limit the way the participants will perform the calibration task. The more detailed the higher the degree of equivalence in the result that can be expected. It might be a natural ambition for a pilot laboratory to reach a good agreement between the results of different laboratories. Too well defined instructions however, will not necessarily support a desired independence between the laboratories. In this inter-comparison the ambition was rather to give as little guidance as possible and instead provide necessary details to the participants on demand. Thus the only help was in form of a spread sheet protocol (excel) giving space for two volume determinations with 10 repeated measurements and a basic form to specify the accounted uncertainty contributions. A filled in example of this protocol, which was worked out by one of the participants, IPQ, is shown in appendix A. The idea was to let all participants act on their own relying on their own experience. Thus no help was given specifying a draining or dripping time when emptying the standard. And no recommendation was made as to the eventual cleaning, although those influences on the calibration results are well known from an earlier exercise [1, 2]. Only a few obvious uncertainty factors were listed and complemented with a simple calculation structure, which however, was to be filled in by the participating laboratories.

2.4 Participants and Schedule

The participating laboratories, the calibration period, the contact persons and which volume was determined are given in table 1 containing 20 laboratories from 19 countries.

Table 1 (**C** = Contained volume; **D** = Delivered volume)

Country	Laboratory				Responsible	Contact
Austria Jan-2004	BEV		**C**	**D**	Wilhelm Kolaczia	Tel: +43 1 49110 509 e-mail: w.kolaczia@metrologie.at
Belgium Aug-2003	SMD		**C**	**D**	Daniel Robert	Tel: +32 22479630 e-mail: daniel.robert@mineco.fgov.be
Bulgaria July-2004	NCM		**D**	**D**	Mariana Miteva	Tel: +359 2 873 52 88 e-mail:ncm@sasm.orbitel.bg
Czech Republic Nov-2002	CMI		**C**		Tomas Valenta	Tel: +420 40 6670728 e-mail: tvalenta@cmi.cz
Denmark Feb-2003	FORCE * (Dantest)	❖	**C**	**D**	Lene S. Kristensen	Tel: +45 432 67 106 e-mail: lsk@force.dk
France Jan-2003	LNE/ CMSI		**C**		Tanguy Madec	Tel: +33 1 40 433934 e-mail: tanguy.madec@lne.fr
Germany Jan-2004	PTB			**D**	Helmut Többen	Tel: +49 531 592 3110 e-mail: helmut.toebben@ptb.de
Great Britain Aug-2003	NWML	❖	**C**	**D**	Chris B Rosenberg	Tel: +44 20 8943 7255 e-mail: helmut.toebben@ptb.de
Greece Jan-2004	EIM		**C**	**D**	Zoe Metaxiotou	Tel: + 30 2310 569962 e-mail: zoe@eim.org.gr
Hungary Nov-2003	OMH		**C**	**D**	Csilla Vámossy	Tel: 36-1-45-85-947 e-mail: c.vamossy@omh.hu
Italy Dec-2003	INRIM * (IMGC)	❖	**C**	**D**	Giorgio Cignolo	Tel: + 39 011 3977448 e-mail: g.cignolo@inrim.it
Netherlands May-2003	NMi	❖	**C**	**D**	Erik Smits	Tel: +31 78 633 2201 e-mail: fmsmits@nmi.nl
Poland Apr-2004	GUM			**D**		
Portugal Aug-2002	IPQ		**C**	**D**	Elsa Batista	Tel: +35 1212948167 e-mail: ebatista@mail.ipq.pt
Slovakia May-2003	SLM		**C**		Miroslava Benkova Ivana Kianicova	Tel: +420 2 60294202 e-mail: Benkova@smu.gov.sk
Slovakia May-2003	SMU		**C**		Miroslava Benkova Mišovich	Tel: +420 2 60294202 e-mail: Benkova@smu.gov.sk
Spain Aug-2003	CEM		**C**	**D**	Antonio Puyuelo	Tel: +34 91 8074700 e-mail: puyuelo@mfom.es
Switzerland Dec-2002	METAS * (EAM)	❖	**C**	**D**	Bruno Kaelin	Tel: +41 31 32 33 243 e-mail: bruno.kaelin@metas.ch
Sweden July-2002	SP	❖	**C**	**D**	Peter Lau	Tel: +46 33 165462 e-mail: Peter.lau@sp.se
Turkey June-2003	UME		**C**	**D**	Ûmit Akcadag	Tel: +90 262 646 6356 Umit.akcadag@ume.tubitak.gov.tr

❖ Participants for a second time: NWML (Great Britain), Dantest* (Denmark), NMi (Netherlands), EAM* (Switzerland), IMGC* ((Italy), SP (Sweden) – DFM (Norway), TTK (Finland) – 8 laboratories from 8 countries. * same institution – changed name.

3 Experimental Procedure

There is a standard, ISO 4787 for this calibration of glassware, but several of the laboratories use their own derived formula for calculating the contained or delivered volume. The gravimetric method comparing the filled and empty standard to mass standards was preferred by all participants. For the determination of delivered volume two possibilities exist. It can be derived from the difference between the contained and the remaining liquid after emptying (wet surface). Or it can be calculated from the poured liquid that is fetched and weighed in a separate container. Due to evaporation during the pouring process this can render two different definitions of delivered volume. Some laboratories stated their pouring and tripping time (no times were prescribed). Some laboratories mentioned that they performed a cleaning others didn't. All these differences in handling have of course some effect on the result. It was, however, the "service to calibrate" without restrictions that was the object for comparison, not a detailed prescribed procedure.

4 Results

4.1 Reported Laboratory Results

The primary information from this volume calibration inter-comparison of the 5 litre volume standard is summarized in table 2. The values here were rounded to 2 decimals by the pilot laboratory. The uncertainty data are understood on a 95 % coverage level (k=2).

Table 2. Reported values for contained and delivered volumes and claimed uncertainty.

Laboratory	**Contained volume**	Expanded uncertainty	number of tests	**Delivered volume**	Expanded uncertainty
	[mL]	[mL]		[mL]	[mL]
IPQ	4 999,99	0,14	10	4 997,52	0,2
SP	5 000,13	0,46	10	4 998,09	0,67
FORCE	4 999,79	0,37	10	4 997,55	0,40
CMI	4 999,72	0,46	10		
METAS	4 999,86	0,35	11/15	4 996,98	0,34
BNM-LNE	4 999,47	0,13 (0,35)*	4		
SMU	4 999,60	0,53	10		
SLM	4 999,91	0,29	8		
PTB			10	4 998,34	0,16
NMi	5 000,14	0,35	10	4 997,52	0,37
UME	4 999,36	0,37	15	4 996,67	0,40
CEM	4 999,42	1,1	10	4 996,57	1,1
BEV	4 999,73	0,45	10	4 997,42	0,45
SMD	4 999,17	0,38	10	4 997,93	0,45
NWML	5 000,30	0,35	10	4 998,81	0,41
OMH	4 999,76	0,22	10	4 998,45	0,6
IMGC	5 000,06	0,26	10	4 997,38	0,53
GUM			15	4 996,92	0,31
EIM	4 999,82	0,7	10	4 998,29	0,8
NCM	4 999,97	0,22	10	4 997,76	0,34

* Note: The higher value takes into consideration the uncertainty in meniscus adjustment (a component that routinely was added first after the measurement in 2003).

The majority of the data was delivered via e-mail in the prepared excel form, which was the fashion the pilot preferred. The pilot had insight in details of the measurement data, the judgements and calculations. In one case only a summary was sent in on paper. No attempts were undertaken to receive more information. Two laboratories send exhaustive reports on their calibration work along with the excel-form (IMGC, EIM) and one laboratory (NWML) gave a short presentation.

Figure 2 and figure 3 display the calibration results in table 2 in a graphical form. For comparison the mean, weighted mean and median values for respective series of measurement results is shown too. Two dashed lines symmetric to the chosen reference were drawn to indicate a 2-sigma band. This choice is presented in 4.2.

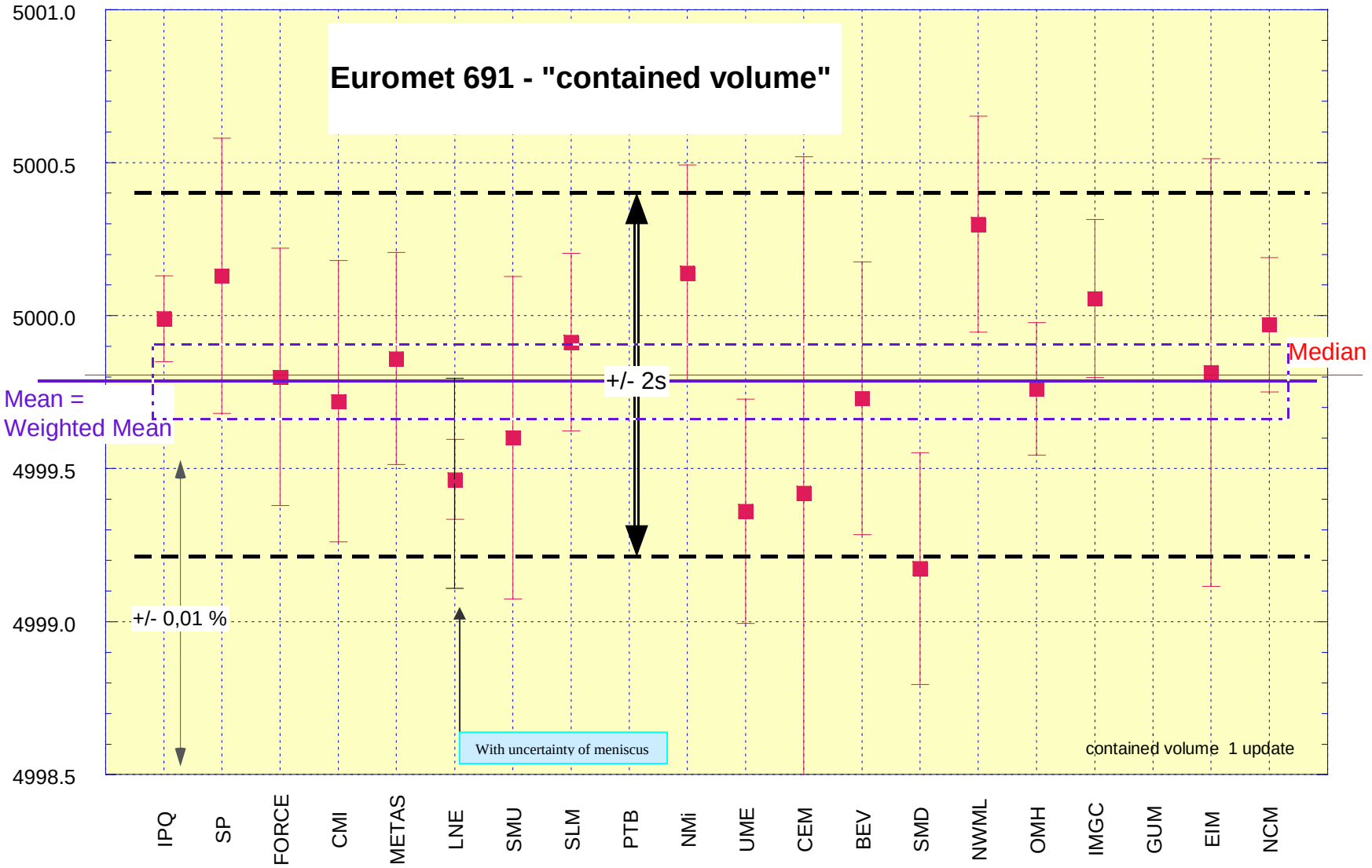


Figure 2. Graphical result for the contained volume. The mean equalling the weighted mean and the median are shown as a reasonable reference value. The dashed lines indicate a 95 % coverage interval based on the inter-laboratory standard deviation. A symmetric uncertainty band concerning the chosen reference is indicated. All laboratories are listed for easier comparison with figure 3. LNE

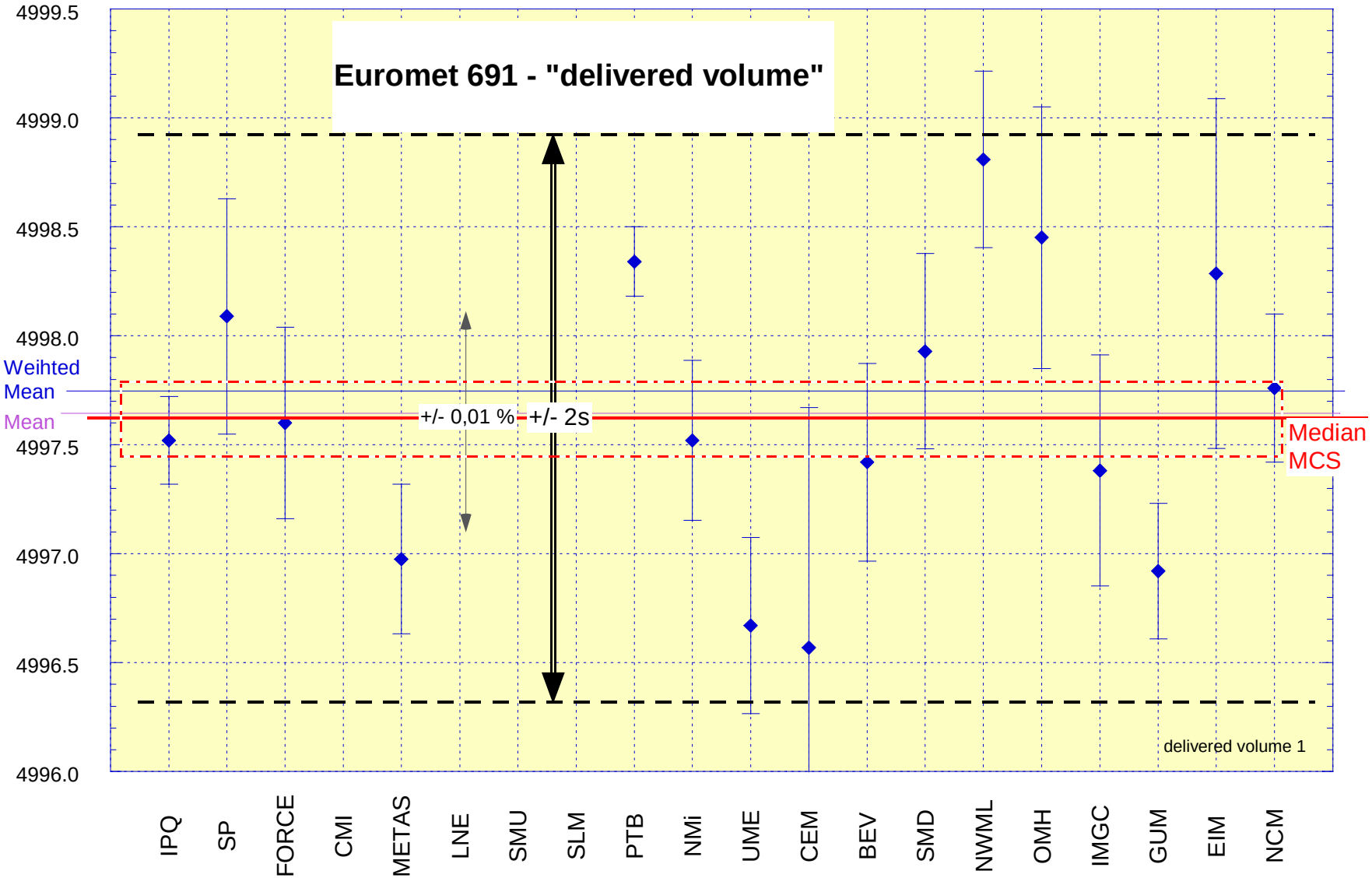


Figure 3. *Graphical result for the delivered volume. Three possible reference values are given with an uncertainty band symmetrically around the selected one (median). The dashed lines indicate a 95 % coverage interval based on the inter-laboratory standard deviation*

From these two graphs some immediate observations can be made:

- The overall impression of this inter-comparison is well satisfying.
- Almost all results fall inside a 95 coverage band around the reference value.
- No result is a clear outlier.
- All results seem consistent with each other at first glance following a normal distribution.
- The claimed uncertainties differ by factor 8 from the lowest 0,13 to the largest 1,1 mL (0,0026 % to 0,022 %). This numbers can be compared with the tolerance of 1,2 mL (0,024 %) with which this volume device is produced. LNE/CMSI announced after the first draft of this report that since 2004 an uncertainty is accounted for the meniscus adjustment (0,16 mL at k=1), which would give a total uncertainty of 0,35 instead of the reported 0,13 mL.
- However, some of the volume results (5 in each series) and their belonging uncertainties are not totally consistent with the reference value and need a closer investigation (see 4.3).

4.2 Reference Values for the Inter-comparison

For key comparisons the obligatory choice of reference value should be the weighted mean - if the results are consistent, Cox [3].

$$x_{ref} = \frac{\sum_{i=1}^{n} \frac{x_i}{u^2(x_i)}}{\sum_{i=1}^{n} \frac{1}{u^2(x_i)}} \qquad (1)$$

$$u(x_{ref}) = \left(\sum_{i=1}^{n} u^2(x_i) \right)^{-1} \qquad (2)$$

x_{ref}:	definition weighted mean
x_i:	volume reported by laboratory i
$u(x_i)$:	uncertainty belonging to volume x_i
$u(x_{ref})$:	uncertainty belonging to reference volume x_{ref}

If we assume all provided results are good ones, and if we further believe all the claimed measurement uncertainties are credible, then we should allow a result x_i with a very low uncertainty $u(x_i)$ to have a strong influence on the reference value. And vice versa a high uncertainty should imply a smaller influence on the reference as equation (1) points out.

If on the other hand, the uncertainties are not reliable, i.e. some laboratory has underestimated whereas another one has overestimated its uncertainty, the arithmetic mean would give all results the same influence on the reference value. In situations with few inter-comparison results (< 8) the median would be less sensitive to extreme values and to possible outliers. Table 3 below illuminates the situation for the three measures in the two reported volumes in this inter-comparison.

Table 3. Different reference values [mL] – bold figures indicate the chosen reference.

Type of Reference value	"Contained" volume (18 Labs)	$U(x_{ref})$ (k=2)	"Delivered" volume (16 Labs)	$U(x_{ref})$ (k=2)
Mean	4999,789	± 0,141	4997,640	± 0,338
Weighted mean	**4999,789**	**± 0,128**	4997,750	± 0,204
Median	4999,808 *		**4997,618 ****	**± 0,175**
Range in reference values	3,8 ppm		26 ppm	
Range in reported volume	225 ppm		448 ppm	

****** With and * without respect to reported uncertainties.

Concerning the contained volume comparison the mean and the weighted mean happen to result in the same value. The range for the different possible reference values is only 3,8 ppm compared to a range of 225 ppm in the 18 laboratory results (60 times larger spread). The conclusion can be drawn that the weighted mean is a good representative. The uncertainty $U(x_{ref})$ (k=2) connected to the reference value is given as well (see equation (2)).

For the delivered volume the range of reference values is seven times larger than for the contained volume (26 ppm). The range of reported volume results is 448 ppm, which is twice as large as the corresponding figure for the contained volume. As there are more uncertainty contributions such a result is reasonable. As shown below (4.3.3) and in contrast to the first impression, the results in the delivered volume are not completely consistent. The recommended technique in such a case is to use the median as estimator for a Monte Carlo Simulation involving the uncertainties around the reported volumes and using the average over many medians as the most reliable reference value (see appendix 4 and 5).

4.3 Consistency in the Results

As mentioned above the two measurement series seem to be consistent as long as one does not incorporate the belonging uncertainties. Cox suggested a simple hypothesis test using a chi squared test. The idea is to build the individual differences to the weighted mean, square them and "normalize" them using their claimed uncertainty squared. The sum of all these terms, the observed χ^2_{obs} -value (see equation (3)), is compared with the chi squared distribution, which tells if the sum of these differences are generated by chance, given the belonging degree of freedom (i.e. ν =17 and 15 for n=18 and 16 results per respective series). If the probability for this test value $\chi^2(\nu)$, that can be calculated using excel or looked up in a table, to be larger than the observed one is less than 5 %, the used significance level, then one should accept the hypothesis that these normalized differences are randomly distributed and thus accept the weighted mean as reference value. Otherwise the hypothesis is rejected and the weighted mean is not accepted as reference.

4.3.1 Test Procedure for Consistency

The observed χ^2_{obs} –value is calculated by the following equation with n=18 or 16 for the contained and delivered volume respectively.

$$\chi^2_{obs} = \sum_{i=1}^{n} \frac{(x_i - x_{ref})^2}{u^2(x_i)} \qquad (3)$$

The test criteria for rejection of the hypothesis of consistency is

$$Pr\{\chi^2(\nu) > \chi^2_{obs}\} < 0{,}05 \qquad (4)$$

4.3.2 Consistency in Results concerning Contained Volume

For the contained volume the test shows clear consistency giving a probability of 37,8 %, which is clearly larger than 5 %. Thus the weighted mean is a good reference.

The consistency check recommended by Cox does not take into consideration the uncertainty in the reference value itself, which would be worthwhile. Consistency is easier achieved if many laboratories are involved and if some of them specify large uncertainties. To judge how reasonable this test works, we can look what would happen if we lower the highest uncertainty statement (1,1 mL) to an average one (0,357 mL). This lowers the probability to 32,3 %, but does not change the statement of consistency.

Table 4. The Chi-squared test result at three conditions.

χ^2_{obs}	$Pr\{\chi^2(17) > \chi^2_{obs}\}$	Conditions for test
18,18	0,378	consistent - with uncertainties as reported
19,10	0,323	with U(CEM) reduced to average uncertainty
89,45	0,00000	with lowest uncertainty 0,13 mL assumed for all laboratories

But one has to keep in mind that 17 degrees of freedom is a relatively large number that would allow some results to be departed from the reference value as long as there are enough results with large uncertainties. Assuming the lowest claimed uncertainty (0,13 mL) to be valid for all results would definitely lead to a judgement of non consistency.

4.3.3 Consistency in Results concerning Delivered Volume

For the delivered volume the Chi-squared test confirms a non-consistency (probability less than 5 %) among the reported results on delivered volume. This is not easily detected in figure 3. The number of results is lower and the spread is larger. Most of all, however, the relation between the deviation from the reference value and the claimed uncertainty is critical.

Table 5. The Chi-squared test result at three conditions.

χ^2_{obs}	$Pr\{\chi^2(17) > \chi^2_{obs}\}$	Conditions for test
46,11	0,00005	non-consistent with uncertainties as reported
27,09	0,0188	non-consistent with x(PTB) removed giving the largest value for $(x_i-x_{ref})^2/u^2_i$
16,2	0,239	consistent with further x(NWML) removed giving the second largest value for $(x_i-x_{ref})^2/u^2_i$

If we first remove the result with the largest contribution to the observed Chi-squared value, we find that the remaining results still are not consistent as 0,0188 is less than 5 %. First if we also remove the second largest observed Chi-squared component building up the weighted mean a consistent result can be reached. Again this result is also dependent on all other claimed uncertainties. In the given situation the weighted mean would be reduced from 4997,71 to 4997,39 as both rejected values are larger. This is however not the way the reference value is defined in this case. Because of the inconsistency the reference value is determined by a procedure presented in 4.3.4.

4.3.4 Monte Carlo Simulation Generated Reference Value

The alternative technique proposed by Cox was conducted. The principle is the following. From each of the 16 laboratories a randomly selected value lying within the range of the possible outcome is selected. For the random value of an arbitrary laboratory we assume a normal distribution around the reported average with half of its stated uncertainty as the

belonging standard deviation. The Monte Carlo Simulation depicts thus a value from 16 normal distributions characterizing all possible results. For such a data set of 16 randomly chosen values the median is calculated. A great number (preferably 10^6) of medians, all calculated on random selections are then averaged to represent the best possible reference value. The multitude of median values form a normal distribution and the uncertainty (k=2) in the reference value then is given by the 95 % confidence interval of this distribution.

A Monte Carlo Simulation is normally run with suitable programs. In this project the simulation was performed using Excel and an Add-in called PopTools generating the random numbers and the summary statistics. Instead of many thousand runs in one stage repeated runs of a size of 5000 medians, were performed, which is easier to handle with Excel. From these repeated runs several averages were collected showing very small differences (range <0,6 ppm). An example is given in appendix 5. This was considered stable enough for the reference value, which is given in table 3 and figure 3. The median in figure 2 is calculated without respect to reported uncertainties.

4.4 Uncertainty Considerations

In a perfect metrological world we would expect that claimed low uncertainties would go along with results having a small deviation to the reference value and vice versa. Figures 2 and 3 prove that we have not come so far. There is no straight relation between low uncertainty and closeness to the reference. The highest uncertainty statement of 1,1 mL (0,022 %) is 8 times larger than the smallest one 0,13 mL (0,0026 %).The average value is 0,4 mL (0,008 %).

The corresponding figures for the delivered volume are generally higher with an average of 0,465 mL (0,009 %) and a factor 7 between maximum (1,1 mL) and minimum (0,16 mL). Most laboratories performing both measurements have increased their claims for the delivered volume, which should be expected. With the findings of 4.3.3 one can conclude that the minimum value seems too low and that even the average should be somewhat higher. With those differences it might be interesting to compare how the laboratories have estimated their most important contributions. This is done graphically in figure 6.

The uncertainty connected to the chosen reference value is given in table 3 on a 95 % coverage level. In the case of the weighted mean (contained volume) it is given by equation (2). In the case of the arithmetic mean it is given by a two sigma interval based on the standard deviation of the mean between the laboratory results. In the case of the median (delivered volume) the 95 % confidence level of the Monte Carlo simulation is used (see appendix 5).

4.5 Calibration Details in Comparison

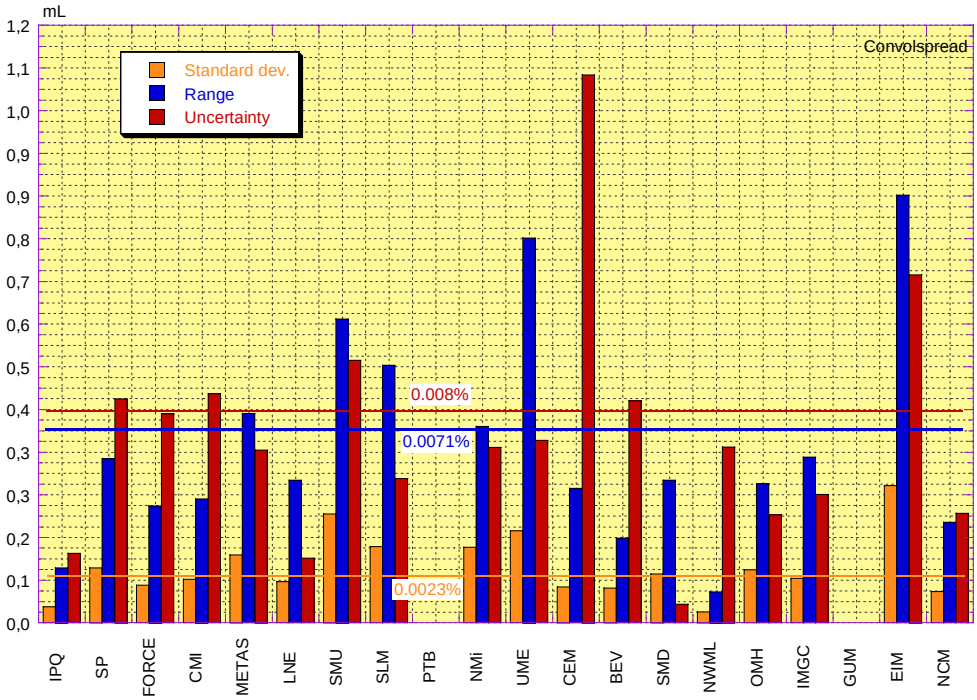


Fig. 4. Comparison between different laboratories experimental skills in terms of in-series standard deviation, range and expected uncertainty for the measurement of the contained volume together with the respective average for all laboratories.

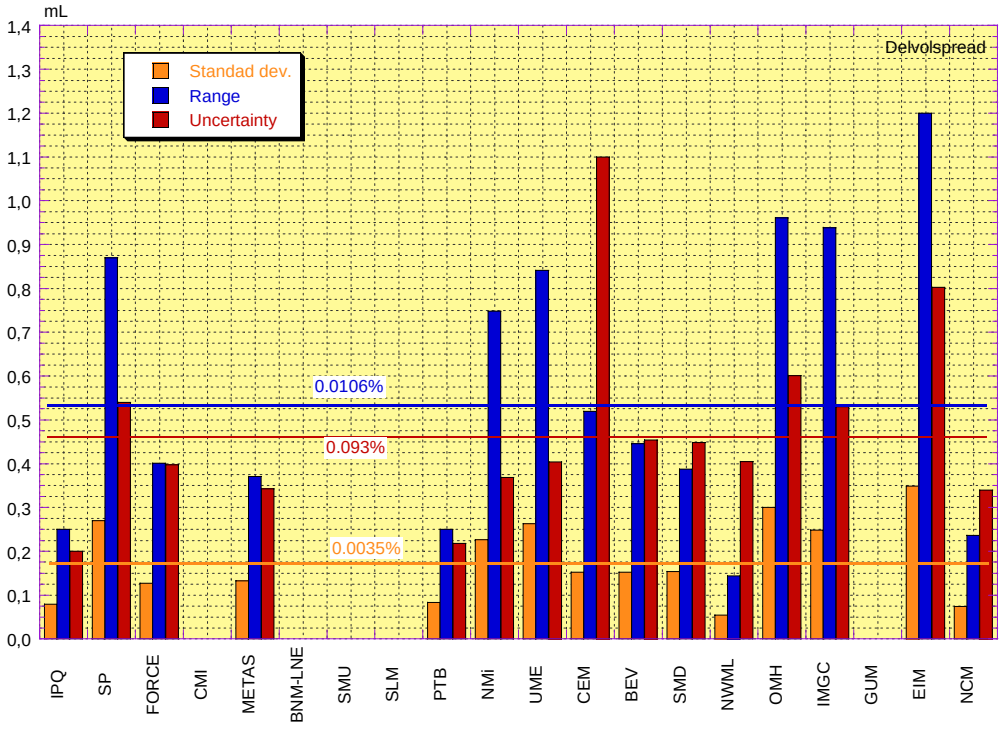


Fig. 5. Comparison between different laboratories experimental skills in terms of in-series standard deviation, range and expected uncertainty for the measurement of the delivered volume together with the respective averages for all laboratories.

For several of the participating laboratories this inter-comparison is their first one and some did not have this calibration as a routine service before. Therefore it might be interesting to analyse more than just the reported volumes. One aspect is to compare the repeatability in this rather manual work, both in terms of in-series standard deviation and range, i.e. the difference between the maximum and minimum in a measurement series. This is done in figure 4 and 5 above.

The highest and lowest standard deviation for the two series differ by a factor 12 and 9 respectively. For the range a factor 14 and 8 apply between highest and lowest spread. Average values for the standard deviation, the range and the uncertainty statement are indicated by a line for comparison. The accompanying number indicates the relation to the actual volume in percent. It can also be noted that the average uncertainty is larger than the average range for the contained volume, whereas the opposite is valid for the delivered volume. The above figures clearly show that there is a potential to improve the calibration work.

For the experimenters in an inter-comparison it is of special interest how the others have judged their uncertainty. The following figures in parenthesis displays the numbers of uncertainty contributions that the experimenters from the different laboratories have listed: CEM (8), EIM (11), SMU (7), CMI (8), SP (9), BEV (7), FORCE (8), SMD (5), UME (6), NWML (5), NMi (7), METAS (6), GUM (9), SLM (8), IMGC (6), NCM (8), OMH (8), PTB (14), IPQ (7), LNE/CMSI (7;8). Figure 6 below tries to visualize the first four of these on standard level (k=1) just linearly "stacked". In practice this sets the level.

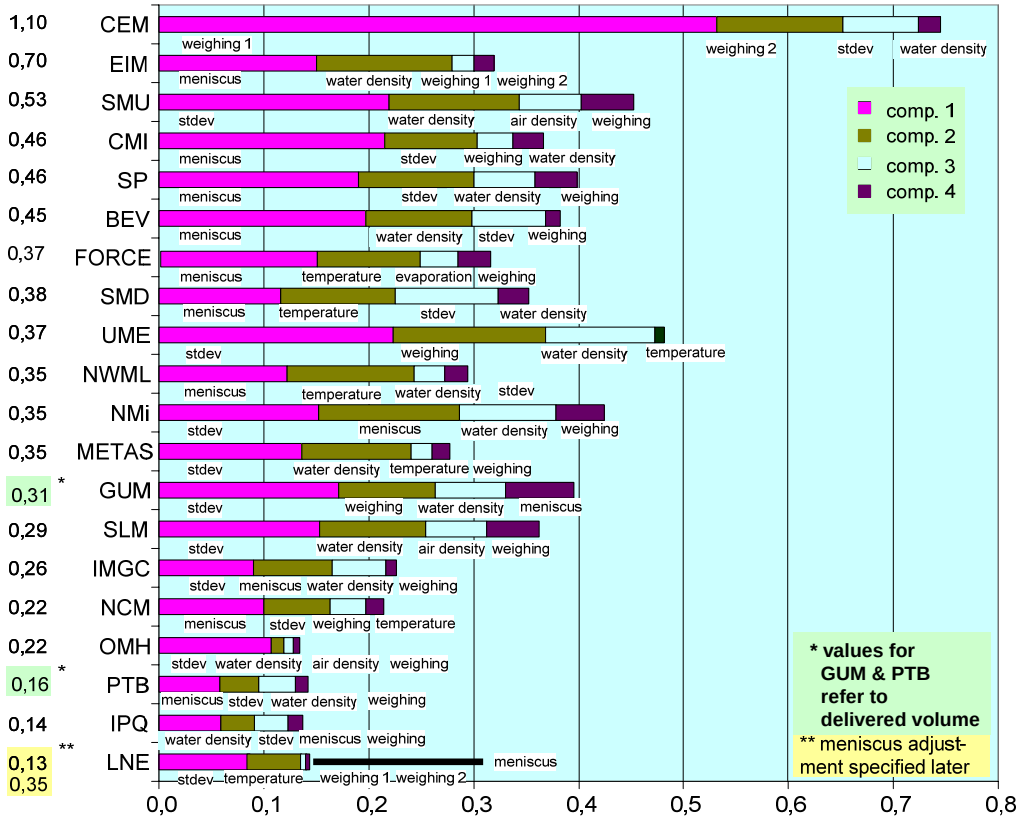


Fig. 6. Comparison of the four dominating components in the uncertainty estimation. (contained volume). The laboratories are sorted in order of falling expanded uncertainty claims in the left given in mL. LNE updated after draft 1 of this report.

4.6 Degree of Equivalence

An overall picture for the whole comparison is presented in figure 7 below. For all laboratories having reported two volume results their deviation from the corresponding reference value is plotted in a cross correlation plot. This represents the degree of equivalence in a graphical way. The six laboratories that reported just one volume (4 laboratories for the contained and 2 for the delivered volume) are plotted on the “zero-line” of the not reported volume, thus showing only the equivalence with respect to the volume reported. Only one result falls just outside a 95 % confidence area.

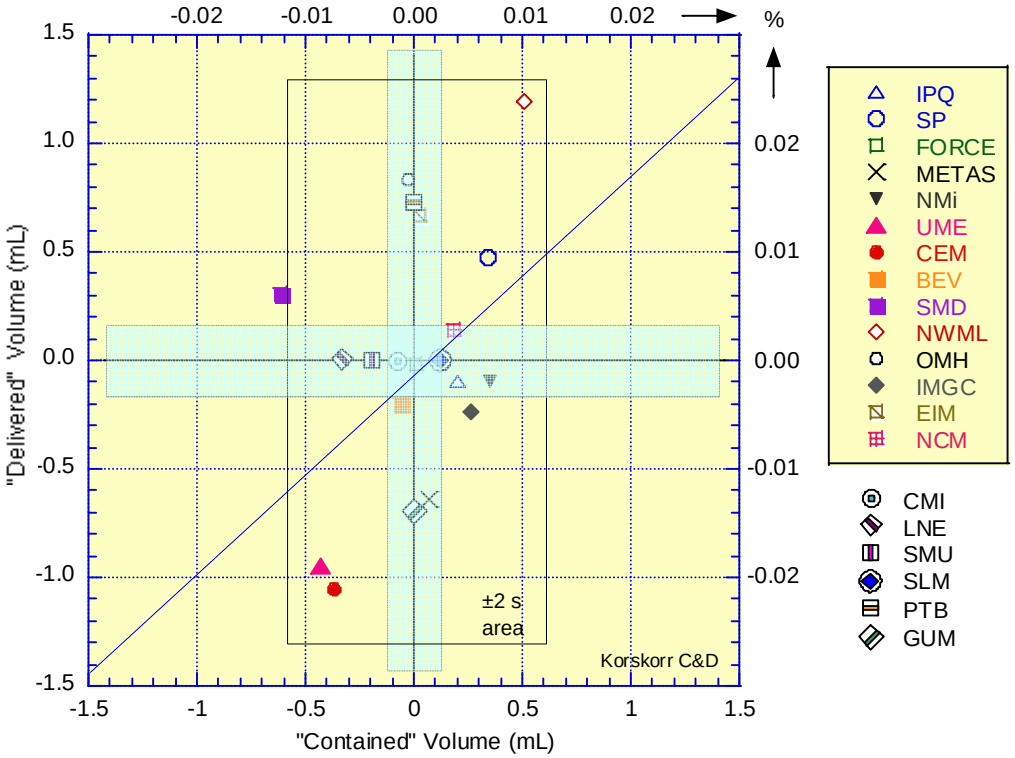


Fig. 7. Cross correlation plot for all results with respect to the two reference volumes. The shadowed fields indicate the uncertainty in the respective reference value. The large rectangle covers a ±2 sigma area. The equivalence/deviation is given both in absolute (mL) and relative (%) figures.

5 Discussion of Results

The inter-comparison results presented can be considered typical for a volume calibration of glassware. No further instructions or details were given how to do the job. Thus six laboratories only determined the one volume they regarded relevant although both volume definitions are used in practice in laboratories.

A lower value for the delivered volume was expected, likewise a worse repeatability and a higher variability or inter-laboratory spread between the laboratories. The emptying procedure can generate additional random effects, but also lead to systematic shifts due to differences in cleaning, in pouring and tripping time, which was not prescribed. An analysis of how much those aspects matter for the individual results and their variation was not the objective of this inter-comparison and no detailed information was inquired.

5.1 Systematic Effects in Volume Calibration

A vital motive conducting comparison is to detect eventual systematic effects that are difficult to resolve otherwise. Two calibrated volumes render this possibility. If in the scatter of data both values are high or low compared to the respective reference value, then a used weight, thermometer, the assumed water density or some aspect in the calculation model might suffer from a systematic error. Also a personal preference in meniscus adjustment might add a systematic influence. A further possibility lies in the effect of cleaning or not cleaning the glassware prior to calibration, which might lead to a somewhat larger wetting mass in the neck and a more complete draining. Figure 7 with 14 real pair-wise results only indicates a tendency for four laboratories (NWML, SP, UME, CEM) to have both volume results high or low respectively. Otherwise there is only a weak correlation for the data series (the thin straight line indicates a linear least square fit for 14 laboratories with a correlation coefficient 0,46). The most probable reason for these four results is a systematic over- or underestimation of the correct meniscus setting. Important to notice is that only the first two of these laboratories mention the meniscus setting as the most important uncertainty component – see fig 6. The meniscus setting was placed by 8, the repeatability (in-series standard deviation) by 9 laboratories as the dominating component. This arguing perhaps can be understood in that way that the repeated measurements would cope with the imperfection in adjusting the meniscus correctly every time.

Based on earlier experience [1, 2] the author believes that there is a measurable preference between different persons to adjust the meniscus and should be treated as a systematic effect by a type B contribution. Several experimenters seem not to share this view. And the central part of figure 7 could support this opinion.

In figure 8 below it is tried to correlate the results from the 5-L flask with those of the almost parallel measured 100 mL pycnometer. To achieve this with those 14 laboratories that delivered results in both inter-comparisons, the relative differences to respective reference value were plotted over each other for comparison.

However, if we compare the relative standard deviation (0,006 %) for the 18 results of the contained volume of the 5 L standard with the same measure for the parallel experiment with a 100 mL pycnometer (0,0039 %) Euromet project 692 [4] we find a 1,5 times larger dispersion in the results here. A reasonable explanation should be the difficulty to adjust a precisely defined meniscus compared to a pycnometer filling.

The cross correlation plot (see figure 8) does not show any tendency for a systematic behaviour, which would result in a pattern grouping the data along a 45 °-line. What can be said in comparison is that we find a larger variance for the 5 L standard, which might be due to the greater difficulty to fill the standard to the ring mark. The non-symmetry to the zero-line concerning the 100 mL pycnometer is due to the selection of the reference value excluding one non-consistent result from the formation of the reference value.

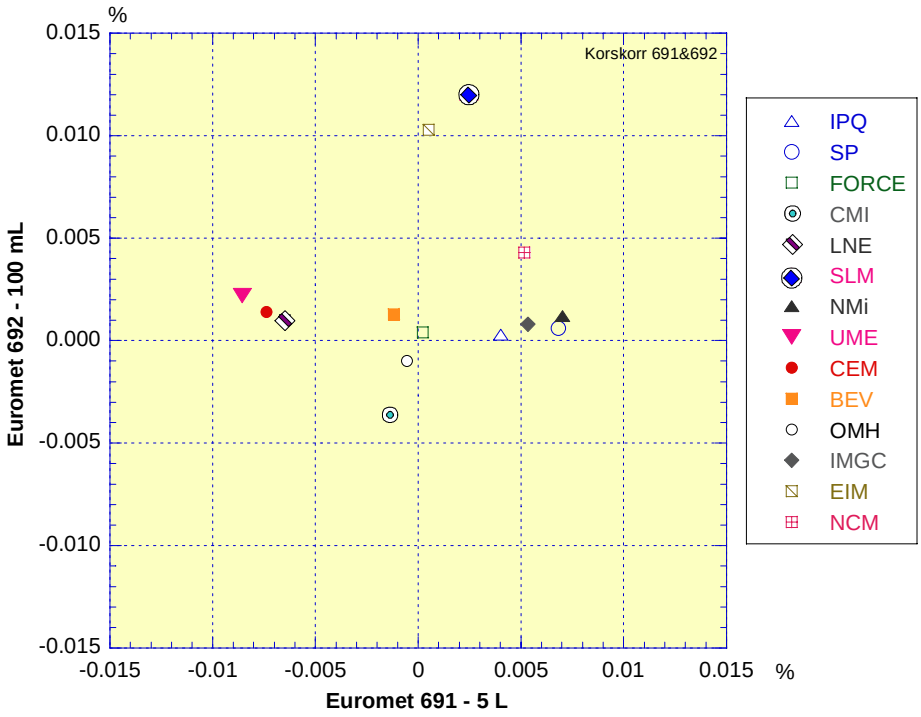


Fig. 8. Cross correlation plot for contained volume representing two volume inter-comparison projects 691 and 692 performed in parallel.

5.2 Experimental Differences

The two figures 4 and 5 reveal that the experimenters probably work at different conditions. Some laboratories exhibit both a very low standard deviation and a small range in their results (IPQ, NWML, NCM, PTB) others show a considerably higher spread (SMU, EIM, SP). For the two last mentioned very low humidity and extremely high and unstable room temperature are most probably the direct causes.

Other differences concerning the quality of water, its temperature and the corresponding density, the used water tables, the quality of weights and weighing instruments probably also have some influence. Those differences are collected and presented in the parallel comparison project 692 [4].

5.3 Differences in Uncertainty Estimation

Besides the agreement in the calibration results the conformity in uncertainty statement is the most important issue. Compared to the field of flow calibrations with a variety of calibration resources or methods and considering a much worse stability of the calibration object one would expect to find a lower variety in uncertainty claims when it comes to

volume calibration. The type-B evaluation is obviously the critical part and the corresponding estimations differ more than expected.

Are the claimed uncertainties realistic? At least for the contained volume the presented results were consistent. The consistency is however guaranteed by large uncertainties for some laboratories. In both series we could find 5 results that with their uncertainty margins (95 % coverage) did not overlap the respective reference value. This is more clearly evaluated using the En-value for each participant (see appendix 4) and gives an indication of an underestimation of the measurement uncertainty.

A speculation of the author to explain this fact is that uncertainty can be understood both in a local and a global view. From a local perspective the experimenter, having good control of all experimental conditions, good repeatability and reproducibility, feels confident that his method and equipment produces a statistically predictable range of results. From a global perspective one must be aware that a different experimenter even with an equivalent method and equipment and a stable test object might get unexplained shifts. If we accept the idea that our colleagues perform an equally good job, then we also must allow for uncertainty contributions beyond our own preferences. The cleaning or not, the correction with a possibly different expansion coefficient, personal preferences or techniques to adjust the meniscus are those kinds of systematic effects for which we might reserve an additional uncertainty to our own intra-laboratory judgement. In most laboratories this calibration is probably always performed by the same person. The uncertainty that a laboratory claims for a calibration should include a within laboratory reproducibility. A possible customer of the calibration service perhaps would expect that a laboratories uncertainty specification should overlap at least 50 percent of the inter-comparison results.

5.3.1 Relation between Type-A and Type-B Estimations

Beyond what was mentioned in 4.5 there are some further aspects of uncertainty statements worth mentioning. Half of the laboratories used the standard deviation of the mean, but without any Student t-factor. The other half used the series standard deviation directly as the type-A contribution. The first group is marked with an * in figure 9 and 10.

The different view between the participants is not casually. Statisticians tell that from a probability point of view the average always is the best representative in a distribution. They call a measurement for an observation and assume 10 repeated volume determinations mean identical measurements. In the opinion of the author putting a weight 10 times on a balance within a minute or so is as close an idealized observation one can get. To adjust the volume to the ring mark several times during two days probably is not the “same” measurement, even if we correct for all temperature effects and instrument drift that might arise. The object itself also can change in terms of pouring and the laboratory conditions might change in a sense that we do not perform the “same” measurement. The standard deviation of the mean implies a reduction by a factor of more than 3 in a type-A estimation if the laboratory is marked with * in the figures 9 and 10.

Standard deviation of the mean s_m and in series s $\quad s_m = s/\sqrt{10}$

One laboratory has performed 12 measurements removed the maximum and minimum value and used the reduced in-series standard deviation as type-A estimation. An other laboratory certainly calculated the standard deviation of the mean but did combine the in-series standard deviation with the contributions estimated according to type-B.

Figures 9 and 10 display the relation between the statistically evaluated uncertainty contribution (Type-A) and the total amount of experience based uncertainty contributions (Type-B). The lower values for the *-marked participants are obvious. But one also can

observe pronounced differences between the first and the second group. The reported in-series standard deviations in group 2 are at average higher for both volumes than for group 1* (factor 1,57 and 1,44). And also the spread in the reported in-series standard deviations is larger for group 2 (factor 1,6 and 2,2). Thus not only has group 1* generally lower experimental spread its contribution to the uncertainty is also reduced in analysis.

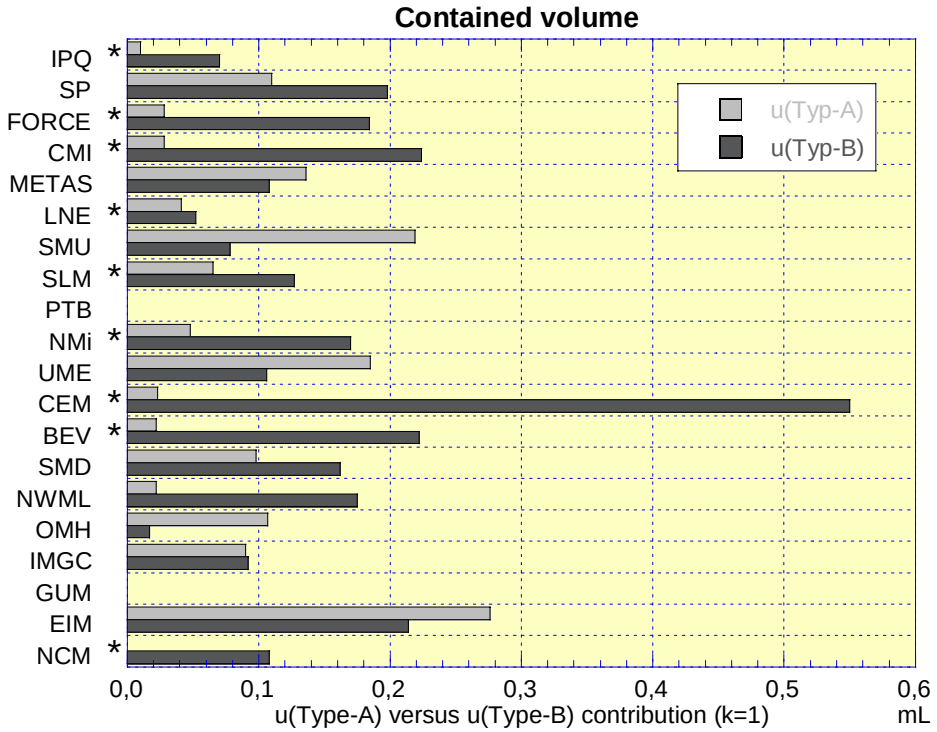


Fig 9. The horizontal bars indicate the estimated uncertainty contributions according to type-A and type-B evaluation connected to the contained volume.

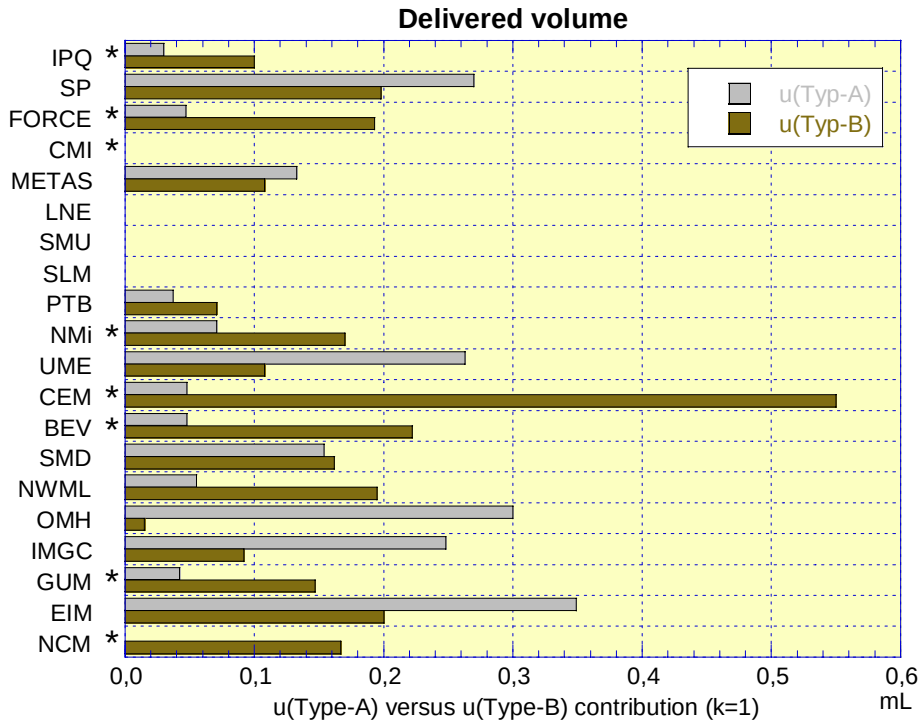


Fig 10. The horizontal bars indicate the estimated uncertainty contributions according to type-A and type-B evaluation connected to the delivered volume.

The second observation is that group 1* laboratories at average have larger type-B contributions than group 2 (factor 1,54 and 1,64 for the two volumes) and that their inter-laboratory spread between these estimations is larger (factor 2,16 and 2,21) than for group 2 laboratories.

A third observation is that 5 out of 18 laboratories state the experimental spread being a larger uncertainty contribution than all type-B evaluations combined. For the delivered volume 6 out of 16 laboratories came to this conclusion. The variability in relation between the two contributions is distinctly shown in the graphs of figure 9 and 10.

A final observation is that one laboratory did not combine all type-B components as listed and another one used the standard deviation from the weighing series directly, not the volume determination series, which would render a somewhat larger spread. But these aspects are marginal and do not change the overall results.

As a final comment one can state that we despite GUM and EA 02-4 still handle uncertainties not really harmonized and that this actually also influences our final statement.

5.4 Degree of Equivalence and CMC-claims

Although this experiment was not conducted as a regional key-comparison the outcome can be looked at as a verification of the laboratories CMC-claims [6].

Table 5 below displays the degree of equivalence (the deviation of each laboratory result from the chosen reference value) in percent. It is the same information as in figure 7, one value for each of the determined volumes. These data can be compared to the uncertainty claims for routine like calibrations but on an excellent level, which is what the calibration

Table 5 Degree of equivalence in comparison

Laboratory	**DoE** [%] ”contained” vol	**DoE** [%] ”delivered” vol	**CMC**-claim [%]	**U**($V_{contained}$) [%]	**U**($V_{delivered}$) [%]
IPQ	0,0040 *	-0,0020	**0,02**	0,0028	0,0040
SP	0,0068	0,0094	**0,01**	0,0091	0,0134
FORCE	0,0002	-0,0004	**0,007**	0,0084	0,0088
CMI	-0,0014		**0,01**	0,0092	
METAS	0,0014	**-0,0128 ***	**0,01**	0,0069	0,0069
LNE/CMSI	-0,0065 *[1]		-	0,0026	
SMU	-0,0038		**0,04**	0,0105	
SLM	0,0025			0,0058	
PTB		**0,0145 ***	**0,004**		0,0032
NMi	0,0070	-0,0020	**0,02**	0,0071	0,0074
UME	-0,0086 *	-0,0190 *	**0,02**	0,0073	0,0081
CEM	-0,0074	-0,0210	-	0,0220	0,0220
BEV	-0,0012	-0,0040	**0,01**	0,0089	0,0091
SMD	-0,0123 *	0,0062	-	0,0076	0,0090
NWML	**0,0102 ***	**0,0239 ***	**0,01**	0,0071	0,0081
OMH	-0,0006	0,0167 *	**0,02**	0,0043	0,0120
INRIM	0,0053 *	-0,0047	**0,01**	0,0052	0,0106
GUM		**-0,0140 ***	**0,01**		0,0062
EIM	0,0005	0,0133	-	0,0140	0,0160
NCM	0,0036	0,0028	**0,02**	0,0044	0,0068

* DoE > $U(x_i)$, ***DoE** > CMC-claim; *[1] LNE has increased U(V) later on by adding an uncertainty for the meniscus setting from 0,13 to 0,35 mL.

measurement capability (CMC) is about. The CMC-claims for comparison are taken from [6]. The red bolt numbers indicate equivalence values that are larger than the CMC-claim. Only one of these numbers concerns the contained value for which probably most of the claims are given. Four of equivalence values for the delivered volume are outside the claim and one is just on the border. Four countries have not declared or got acceptance for a claim for this service (SLM and SMU belong both to Czech Republic).

Does this mean several claims are too narrow? According to the opinion of the author the answer is yes. This position is confirmed studying the En-values given in appendix 4. One should, however, keep in mind that many laboratories interpret the uncertainties given in the CMC-tables as there best, i.e. the lowest uncertainty connected with a calibration. To this only the reference equipment, the measurement conditions and the method do contribute. In case the object contributes with some uncertainty, the repeatability for example, an almost ideal object is assumed. In a comparison like this one, where we had to deal with a definitely non-ideal object, all except one laboratory (FORCE) claimed even lower uncertainties then the relevant CMC-values. This is probably because we all took more care and made more repetitions than we normally would do. The meaning of an uncertainty statement should, however, not only be based on what we know for sure. It should also cover unknown, but possible measurement errors outside of our control.

Therefore the other possible comparison, the degree of equivalence in relation to the specified uncertainty in each volume determination, is even more interesting. Five (one on the border) and six values respectively are indicated with a star behind showing an equivalence value exceeding the stated uncertainty. This information is graphically well caught in figure 2 and 3 as well.

6 Conclusions

Concerning the large number of participants and the given preconditions, especially the object itself, the total outcome shows undoubtedly a good agreement between the laboratories. Only one result out of 34 is just outside a 95 % confidence level of the respective reference value. The answer to the first four questions in 2.2 is yes.

Despite the good agreement in the results there are relatively large differences in the repeatability between laboratories, which probably depend on experimental skill and laboratory conditions. And there are also clear differences how the participants asses and estimate their different uncertainty sources. The main sources are the spread and the adjustment to the ring mark. The varying proportions between the statistical and experience based contributions are, however, astonishing. Also the transformation of the experimental spread into an uncertainty component shows two principal views that should be harmonized.

The demonstrated results support most CMC-claims and even the lower uncertainty statements connected to this calibration exercise. But this is not valid for all results. If the CMC-claims are given to include all glassware, then some claims should be reconsidered. An important conclusion to draw is that we are not harmonized in our view on measurement uncertainty. And this concerns both treatment and estimations. Comparing the results of those laboratories that repeated the measurements with those that took part for the first time (see appendix 3) indicates a better agreement for the first group. This is most clearly seen in the uncertainty statements.

7 References

[1] Lau P., and Waltersson D., Euromet project A88/51 Report on the Intercompariosn of the Calibration of a 5-litre Volume Standard, SP Report 1991:29.

[2] Lau P., Euromet Project A88/51 – Addendum to the Report on the Intercompariosn of the Calibration of a 5-litre Volume Standard, SP Report 1992:34.

[3] Cox M., "The evaluation of key comparison data", Metrologia, 2002, 39, 589-595

[4] Batista E., Inter-laboratory calibration comparison of the volume of a 100 mL Gay-Lussac Pycnometer, Euromet project no. 692.

[5] Batista E., Bilateral comparison of a 100 mL Gay-Lussac Pycnometer, Euromet project no. 793

[6] BIPM homepage
http://kcdb.bipm.org/appendixC/country_list.asp?Iservice=M/FF.9.1.2
and in
CMC-FLOW-table for EU TC F review June 2006.xls

Appendix 1 Report Form

EUROMET Project 691 "Calibration Inter-comparison of a 5-litre volume glass standard"

Data Form

General Information

Country		**Laboratory**	
Responsible		**Date**	

Equipment

	Type	Range	Resolution
Weighing instrument	Mettler PR 8002	0 - 8100 g	0,01 g
Thermometer	Testo 601	0 - 70 °C	0,1 °C
Barometer	Vallace-Tiernan Diptron 3 plu	0 - 1100 mbar	0,1 mbar
Hydrometer	Testo 601	0 - 100 %	0,10%
Quarz thermometer	Hewlett Packard 2801 A	0 - 30 °C	0,001 °C

Other Informations

	Type	Density reference
Water	deionizated water	2 solid density standards

	Type	Density(kg/m^3)
Mass standards	KERN E_2	7866

Used volume calculation formula:

$$V_{20} = (m_2 - m_1) \cdot \frac{1}{\rho_W - \rho_A} \cdot \left(1 - \frac{\rho_A}{\rho_B}\right) \cdot [1 - \gamma(t - 20)]$$

Cleaning and drying the flask: Cleaning: with dishwashing agent, destillated water then ethanol
Drying: with N_2 gas

Comments: The water density was determined by hydrostatic weighing method at 20 °C, and at 1013,25 hPa.

Signature:

Lab (Country) | **Volume contained**

ev. details

Measurements from to

Result

V = **4999,76** ml
U(V) = **0,217** ml
U(V)/V = **43,4** ppm

Air temperature (°C)	20,2
Pressure (hPa)	998,2
Humidity (%)	33
Air density (kg/m3)	1,1831
Density of mass pieces (kg/m3)	7866
Coef. of expansion (1/K)	1,0E-05

Test number	Water mass (g)	Water temperature (°C)	Water density (g/cm^3)	Volume (ml)
1	4985,150	20,6	0,99807740	4999,932
2	4985,520	20,1	0,99818590	4999,681
3	4985,710	19,8	0,99824310	4999,666
4	4985,230	20,6	0,99806820	4999,919
5	4985,250	20,4	0,99812620	4999,731
6	4986,780	18,7	0,99846670	4999,676
7	4986,200	19,2	0,99836050	4999,653
8	4986,020	19,5	0,99831270	4999,695
9	4986,660	18,9	0,99847700	4999,817
10	4987,290	18,3	0,99854400	4999,843
Mean value	4985,981	19,6	0,99828617	4999,761
Standard deviation	0,741	0,8	0,00017330	0,107

Uncertainty budget

Quantity	*xi*	Distribution	Standard uncertainty $u(x_i)$	Sensitivity coefficient c_i	Uncertainty $c_i \times u(x_i)$	Degrees of Freedom *vi*
Balance indication with air (g)	1371,582	normal	6,000E-03	1,003	0,0060	9
Balance indication with water (g)	6357,563	normal	6,000E-03	-1,003E+00	-0,0060	9
Mass of weights (g)	7729,000	normal	2,800E-03	1,003E+00	0,0028	50
Air density (g/ml)	1,1831E+00	rectangular	2,000E-06	4,379E+03	0,0088	50
Water density (g/ml)	0,99828617	normal	2,500E-06	5014	0,0125	29
Weight density (g/ml)	7,866	normal	5,000E-03	0,0956	0,0005	2
Coefficient exp. of glass (1/°C)	1,0E-05	rectangular	1,0E-06	-2039,9038	-0,0020	1000000
Water temperature (°C)	19,6	normal	0,05	0,050	0,0025	50

Volume (ml)	4999,761		
Type A uncertainty (ml)	0,107	**combined Type B uncertainty (ml)**	0,017
Combined uncertainty (ml)	0,109	**Expanded uncertainty (ml) (k=2)**	0,217

Lab (Country) | **Volume delivered**

ev. details

Measurements from to

Result

V =	**4998,45** ml	Pouring time	70 s
U(V) =	**0,601** ml	Draining time	30 s
U(V)/V =	**120,2** ppm		

Air temperature (°C)	20,6
Pressure (hPa)	995
Humidity (%)	25
Air density (kg/m3)	1,1765E+00
Density of mass pieces (kg/m3)	7866
Coef. of expansion (1/K)	1,0E-05

Test number	water mass (g)	Water temperature (°C)	water density (g/cm^3)	Volume (ml)
1	4983,880	20,7	0,99805670	4998,753
2	4984,330	20,2	0,99815470	4998,632
3	4984,390	20,5	0,99810330	4998,933
4	4983,870	20,6	0,99807040	4998,558
5	4983,840	20,8	0,99803960	4998,661
6	4983,940	20,3	0,99814780	4998,320
7	4983,680	20,4	0,99811270	4998,218
8	4985,070	18,9	0,99841300	4998,217
9	4984,990	19,0	0,99839990	4998,251
10	4984,410	19,3	0,99833700	4997,969
Mean value	4984,240	20,1	0,99818351	4998,451
Standard deviation	0,488	0,7	0,00014383	0,300

Uncertainty budget

Quantity	*xi*	Distribution	Standard uncertainty $u(x_i)$	Sensitivity coefficient c_i	Uncertainty $c_i \times u(x_i)$	Degrees of Freedom *vi*
Balance indication with air (g)	1468,810	normal	6,000E-03	1,003	0,0060	9
Balance indication with water (g)	6,4531E+03	normal	6,000E-03	-1,003E+00	-0,0060	9
Mass of weights (g)	7,9219E+03	normal	2,800E-03	1,003E+00	0,0028	50
Air density (g/ml)	1,1765E+00	rectangular	2,000E-06	4,378E+03	0,0088	50
Water density (g/ml)	0,99818351	normal	2,500E-06	5013	0,0125	29
Weight density (g/ml)	7,866	normal	0,005	0,0951	0,0005	2
Coefficient exp. of glass (1/°C)	1,0E-05	rectangular	1,0E-06	359,8872	0,0004	1000000
Water temperature (°C)	20,1	normal	0,05	0,050	0,0025	50

Volume (ml)	4998,451		
Type A uncertainty (ml)	0,300	**combined Type B uncertainty (ml)**	0,015
Combined uncertainty (ml)	0,300	**Expanded uncertainty (ml) (k=2)**	0,601

Appendix 2 Differences in results between project 691 and 51

Six laboratories (NWML, FORCE, METAS, NMi, INRIM(IMGC) and SP) have performed this volume inter-comparison two times. But the experimenters have partly changed. The table below collects some data from which possible changes can be deducted. The basis for the comparison is the arithmetic mean of the results of two groups. The first is the results in the previous project 691 and the second the results in project 51 from 1988.

	Project	Contained Volume [mL]	Stated uncertainty [mL]	Delivered Volume [mL]	Stated uncertainty [mL]
Average for 6 "old"	691	5000,05	0,37	4997,73	0,46
	51	4999,97	0,37	4997,62	0,43
Change to earlier results		0,08	0	0,11	0,03
		increase 16 ppm	-	increase 22 ppm	increase 7 %

	Project	Contained Volume [mL]	Stated uncertainty [mL]	Delivered Volume [mL]	Stated uncertainty [mL]
Inter laboratory reproducibility	691	0,187	0,069	0,640	0,122
	51	0,314	0,184	0,359	0,189
Change to earlier results		-0,127	-0,115	0,281	-0,067
		decrease 40 %	decrease 36 %	increase 78 ppm	decrease 35 %

	Project	Contained Volume [mL]		Delivered Volume [mL]	
Intra laboratory repeatability	691	0,10		0,18	
	51	0,15		0,24	
Change to earlier results		-0,05		-0,06	
		decrease 33 %		decrease 25 %	

All but one (FORCE/DANTEST), who had the highest results, received higher values for both volumes in the project 691 than in the old one 51. Thus the average result for both volumes seems to have slightly increased (16 and 22 ppm). With an average uncertainty twice as large (74 and 92 ppm) this is not really a stated change. The volume standard thus must be considered quite stable. Also the average uncertainty statement has been stable.

Looking to the spread in results between the six laboratories, above denoted as the inter-laboratory reproducibility, and measured as the standard deviation, it has decreased for the contained volume but increased remarkably for the delivered volume. This might be explained by the fact that no cleaning of the standard was prescribed. The percentage given in the table always refers to the earlier project 51. In total no improvement can be stated. However, a kind of harmonization may be detected in the statement of uncertainty as the spread between laboratories judgement has decreased by over 30 %.

The laboratories also seem to have worked with lower in-series spread during project 691. At average the in-series standard deviation, in the above table denoted as intra-laboratory repeatability, has decreased for both volumes about 30 %.

Appendix 3 Differences between repeating (old) and new Laboratories

The same comparison as above can be done between the group of six “old” laboratories in project 691 and the 14 or 12 laboratories respectively denoted “new” with the results of the “old” ones as the basis for reference. Again the comparison refers to the plain average. Thus one can state that the “new” laboratories at average got a some what lower volume. The difference amounts to 78 and 28 ppm for the contained and delivered volume. For the “new” group the average uncertainty statement is somewhat higher (9 and 5 %).

The two reported volumes differ 39 % and 8 % more between the “new” than between the “old” laboratories in project 691 (inter-laboratory reproducibility). But this spread is actually 18% lower for the contained but 92 % higher for the delivered volume than was the spread for the “old” laboratories during the project 51. At the same time it is important to say that some of the laboratories give low uncertainties and some large ones so that the spread in uncertainty statement is considerably higher (296 % and 135 %) than between the “old” ones. This means there is a real potential to come to a more common judgement in uncertainty declarations.

“new” and “old” means first and second participation		Contained Volume [mL]	Stated uncertainty [mL]	Delivered Volume [mL]	Stated uncertainty [mL]
Average for new and old laboratories	“new”	4999,66	0,415	4997,59	0,482
	“old”	5000,05	0,37	4997,73	0,46
Difference between “new” and “old” laboratories		-0,39	0,035	-0,03	0,022
		78 ppm lower	9 % higher	28 ppm lower	5 % higher

		Contained Volume [mL]	Stated uncertainty [mL]	Delivered Volume [mL]	Stated uncertainty [mL]
Inter laboratory reproducibility	“new”	0,259	0,273	0,691	0,287
	“old”	0,187	0,069	0,640	0,122
Difference between “new” and “old” laboratories		0,072	0,204	0,051	0,165
		39 % higher	296 % higher	8 % higher	135 % higher

		Contained Volume [mL]		Delivered Volume [mL]	
Intra laboratory repeatability	“new”	0,12		0,173	
	“old”	0,10		0,18	
Difference between “new” and “old” laboratories		-0,02		-0,007	
		20 % higher		4 % lower	

It is also worth to notice that the reported in-series standard deviations show a 20 higher and 4 % lower spread in the “new” than in the “old” laboratories for respective volume determination. That means there is no distinct difference between “old” and “new”, but quite some difference between the participating laboratories. And the pure fact that the new ones are 14 whereas the old ones are 6 makes these relations reasonable.

Appendix 4 En-values as a different measure for equivalence

A different way to present the outcome of a comparison measurement is to relate the deviation from the reference value to the uncertainty in the deviation itself. This uncertainty is of course dependent not only on the specified uncertainty of the individual reported values, but also on the uncertainty in the reference value. The En-value thus is a normalized equivalence.

$$En_i = \left| \frac{x_i - x_{ref}}{\sqrt{U^2(x_i) + U^2(x_{ref})}} \right| < 1$$

A low En-value tells that the result is closer to the reference than the uncertainty would allow. The border is a value of En = 1. When taking uncertainty into consideration the border value means there is only a small statistical chance that the reported value and its stated uncertainty actually contain the reference value, which is the meaning of the stated uncertainty. For En-values larger than 1 this chance is definitely too small. The En-values calculated according to the above equation are listed in the table 6 and shown graphically in figure 11 below. For the calculations the data given in table 2 and 3 were used.

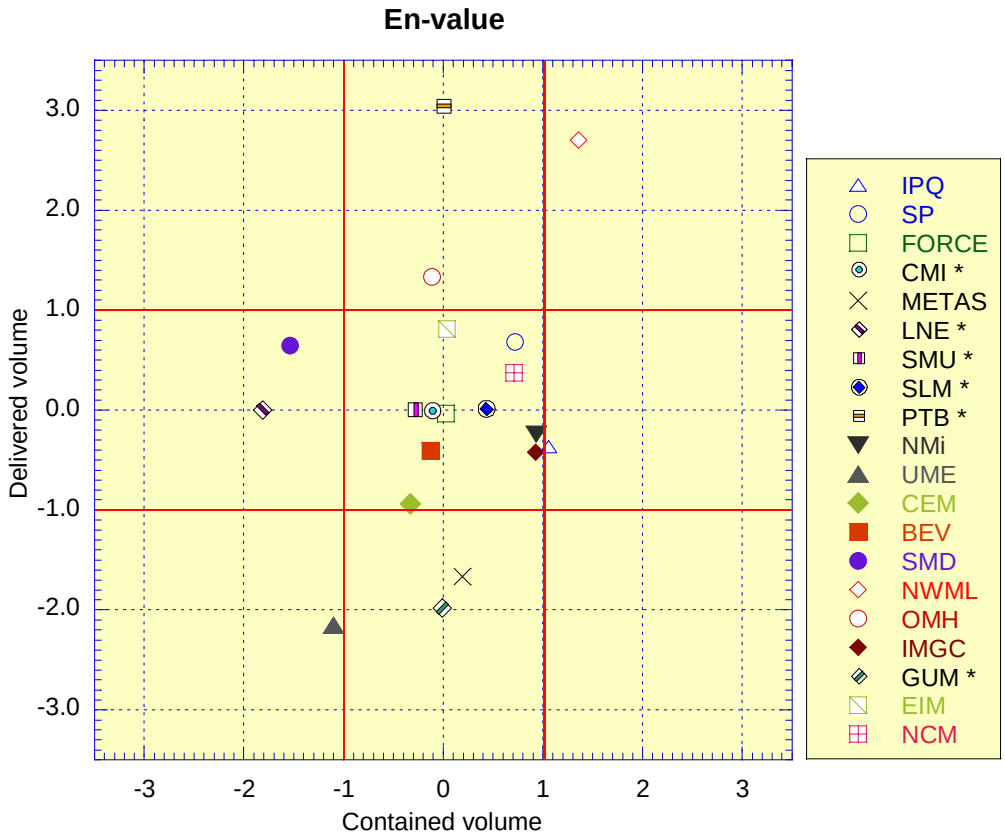


*Fig 11. The two En-values presented as a cross correlation plot. Laboratories marked with * having only one value are plotted on the "zero"-line for the missing data. The pairs of lines set out the border for an acceptable En-value. Here the sign is kept for convenience in presentation.*

The En-value is a reasonable indicator for the combined results in terms of reported value and stated uncertainty. The conclusion thus should be that all symbols outside the inner square of length 1 and the bold numbers in table 6 indicate a too optimistic uncertainty in one or both volumes as long as we accept the chosen references. The picture would of course look different if we had used different criteria to build the reference value, if we for instance had chosen to remove certain results from forming the reference.

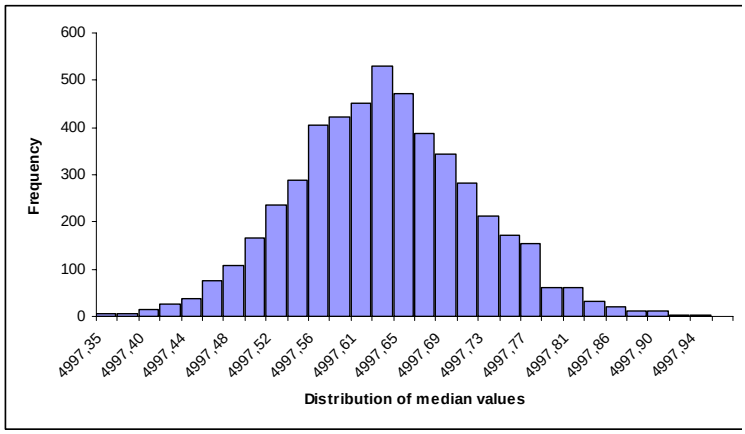
Table 6 Absolute En-values for the comparison

Participating laboratories	En-value for contained volume	En-value for delivered volume
IPQ	**1,060**	0,369
SP	0,720	0,682
FORCE	0,025	0,038
CMI	0,145	
METAS	0,192	**1,667**
LNE/CMSI	**1,776***	
SMU	0,347	
SLM	0,391	
PTB		**3,049**
NMi	0,935	0,240
UME	**1,102**	**2,153**
CEM	0,333	0,941
BEV	0,127	0,407
SMD	**1,537**	0,645
NWML	**1,358**	**2,702**
OMH	0,111	**1,331**
IMGC	0,927	0,423
GUM		**1,961**
EIM	0,037	0,813
NCM	0,711	0,371

* Calculated with the original uncertainty statement.
En=0,87 with added meniscus uncertainty.

Appendix 5 Result of a Monte Carlo Simulation

A typical result from 5000 random median calculations performed on the delivered volume using Excel and PopTools is shown in figure 12 below.



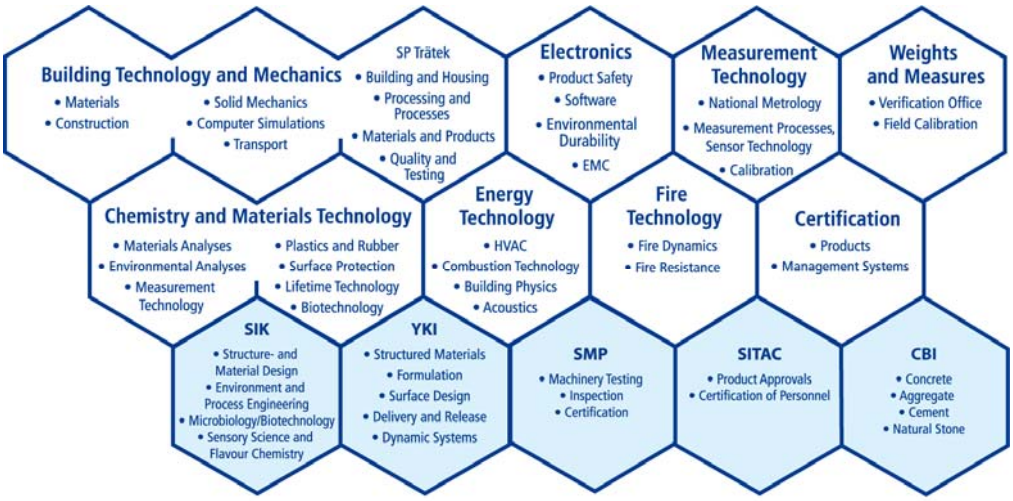
Mean	4997,618
Std dev	0,088754
Min	4997,332
Max	4997,941
Count	5000
Test value	4997,615
Exceed test	2518
Less than test	2482
Lower percentile	4997,447
Upper percentile	4997,798
Uncertainty k=2	0,175295

Fig. 12 Left: Graph over the distribution of 5000 randomly simulated median values calculated by PopTools. Right: The calculated summary statistics.

For repeated trials the mean does not shift more than ±0,003 mL. For the uncertainty in the reference value, i.e. the average over 5000 randomly selected median values, the difference between he lower percentile for 2,5 % and the upper percentile for 97,5 % is divided into a symmetric interval giving 0,1753 mL. This value is between 1,96 and 2 times the above standard deviation of all 5000 median values (0,174 and 0,1775 mL), i.e. corresponding to a 95 and 95,45 % confidence level.

SP Technical Research Institute of Sweden develops and transfers technology for improving competitiveness and quality in industry, and for safety, conservation of resources and good environment in society as a whole. With Sweden's widest and most sophisticated range of equipment and expertise for technical investigation, measurement, testing and certification, we perform research and development in close liaison with universities, institutes of technology and international partners.

SP is a EU-notified body and accredited test laboratory. Our headquarters are in Borås, in the west part of Sweden.



SP is organised into eight technology units and five subsidaries

SP Technical Research Institute of Sweden
Box 857, SE-501 15 BORÅS, SWEDEN
Telephone: +46 10 516 50 00, Telefax: +46 33 13 55 02
E-mail: info@sp.se, Internet: www.sp.se
www.sp.se

Measurement Technology
SP Report 2007:2007:10
ISBN 91-7848-91-85533-77-7
ISSN 0284-5172