Physikalisch-Technische Bundesanstalt
Braunschweig and Berlin
National Metrology Institute

CCM

# Degree of equivalence for KCs – past practice in WGFF and actual considerations.

**Dr.-Ing. B. Mickan (PTB)**
**Gerd Wübbeler (PTB)**
**Mengna Li (NIM)**

**20th meeting of CCM – 26-27 June 2025**
Technical workshop: Session II on Key Comparisons and Digitalization

- Introduction to the practise in WGFF
  *how KCs are evaluated and why we have 1 < En < 1.2 as a warning level*

- Basic idea of Bayesian Null Hypothesis Testing

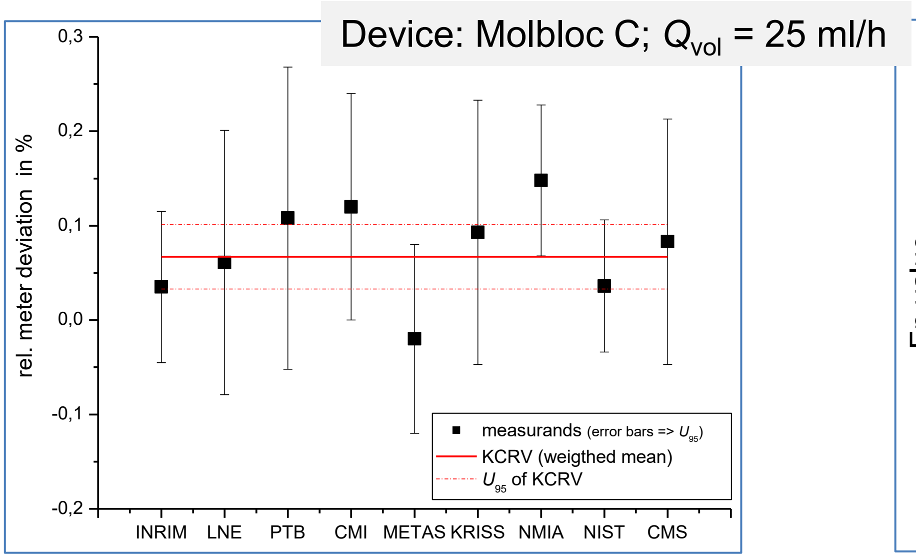
- Some results of its application

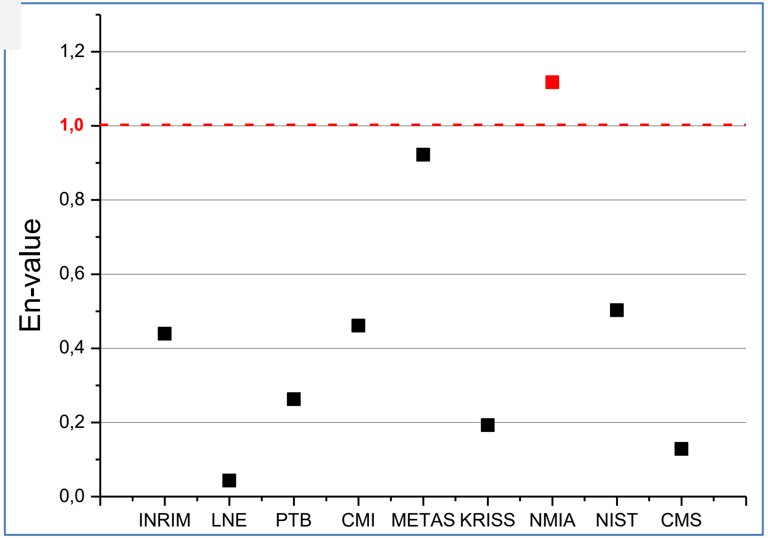
- Conclusions and outlook

# Some Notes on Comparisons in WGFF

- main purpose is to approve CMCs
- high value of DoF => $k$ = 2 commonly in use
- only a few technologies for realisation/dissemination of units are in use
  => uncertainty sources quite good known
  => underrated uncertainties due to specific situations in an individual Lab
  => uncertainties prone to be overrated due to conservative estimates
  => no common dark uncertainty
- transfer standard uncertainty plays specific role
  => shall be determined appropriately (see CCM-Webinar 5th Feb 2025)

# Introduction: Example out of CCM.FF-K6.2017



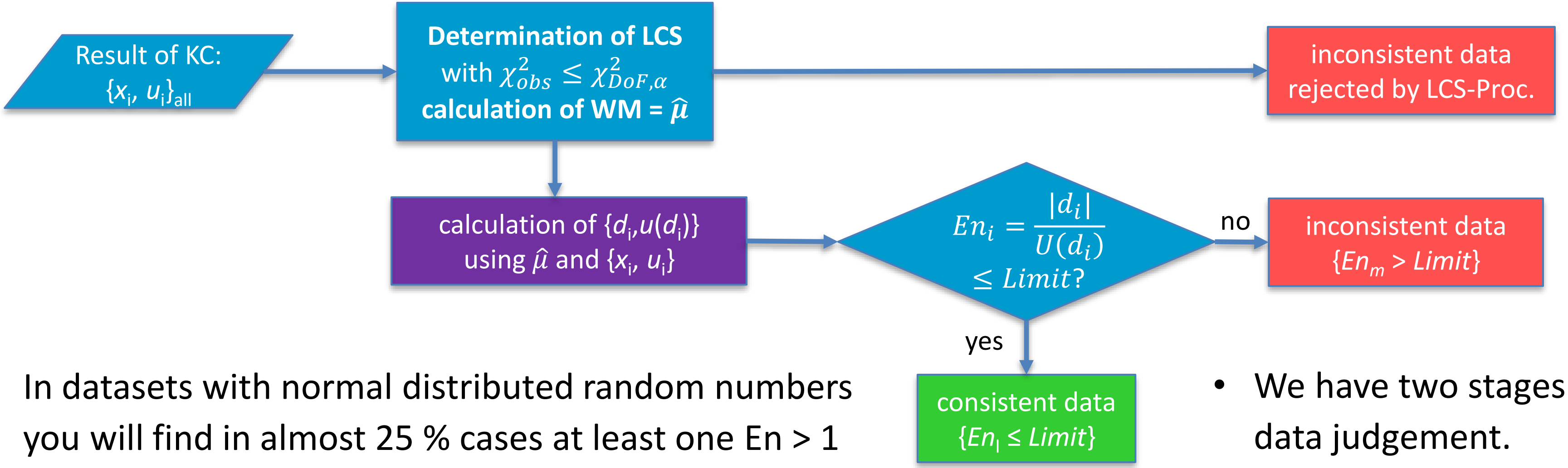

$$\frac{1}{u_{ref}^2} = \sum_i \frac{1}{u_i^2} \qquad KCRV = u_{ref}^2 \sum_i \frac{x_i}{u_i^2}$$

$$E_{n,i} = |x_i - x_{ref}| / \sqrt{U_i^2 - U_{ref}^2}$$

$$\chi_{obs}^2 = \sum_i \frac{(x_i - x_{ref})^2}{u_i^2} = 9.8 \leq \chi_{8,0.05}^2 = 15.5$$

- Shall NMIA be urged to revise its CMC?
- Or, shall NMIA ask for another comparison?

# Conventional Approach used in WGFF




- In datasets with normal distributed random numbers you will find in almost 25 % cases at least one En > 1 even if $\chi^2_{obs} \leq \chi^2_{DoF,\alpha}$
- We are using $1 < E_n \leq 1.2$ as a warning level to reduce the risk that data are declared as inconsistent when they already passed the $\chi^2$-test .

- We have two stages of data judgement.
- In many cases, we will include data in the KCRV which finally are declared as non-reliable.

# EN ISO 17043: Definition and Usage of Zeta- and En-Score

**DIN EN ISO/IEC 17043:2010-05**
**EN ISO/IEC 17043:2010 (E)**

d) The zeta score, $\zeta$, is calculated using Equation (B.4), where calculation is very similar to the $E_n$ number [see e) below], except that standard uncertainties are used rather than expanded uncertainties. This allows the same interpretation as for traditional $z$ scores.

$$\zeta = \frac{x - X}{\sqrt{u_{\text{lab}}^2 + u_{\text{av}}^2}} \tag{B.4}$$

where

$u_{\text{lab}}$ is the combined standard uncertainty of a participant's result;

$u_{\text{av}}$ is the standard uncertainty of the assigned value.

e) $E_n$ numbers are calculated using Equation (B.5):

$$E_n = \frac{x - X}{\sqrt{U_{\text{lab}}^2 + U_{\text{ref}}^2}} \tag{B.5}$$

# EN ISO 17043: Definition and Usage of Zeta- and En-Score

**DIN EN ISO/IEC 17043:2010-05**
**EN ISO/IEC 17043:2010 (E)**

1) for $z$ scores and zeta scores (for simplicity, only "$z$" is indicated in the examples below, but "$\zeta$" may be substituted for "$z$" in each case):

— $|z| \leqslant 2{,}0$ indicates "satisfactory" performance and generates no signal;

— $2{,}0 < |z| < 3{,}0$ indicates "questionable" performance and generates a warning signal;

— $|z| \geqslant 3{,}0$ indicates "unsatisfactory" performance and generates an action signal;

2) for $E_n$ numbers:

— $|E_n| \leqslant 1{,}0$ indicates "satisfactory" performance and generates no signal;

— $|E_n| > 1{,}0$ indicates "unsatisfactory" performance and generates an action signal.

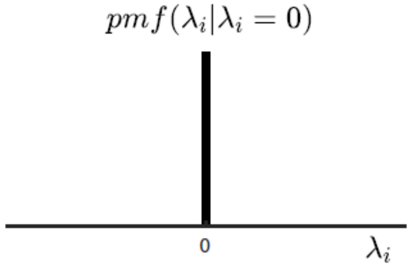
*Please note: In ISO 17043:2023, this explicit „warning" related to z- or Zeta-score is removed;*
*but the new clause 7.7.2 is „***The proficiency testing provider shall select, justify and document appropriate methods and performance criteria for evaluation of participant performance***."*

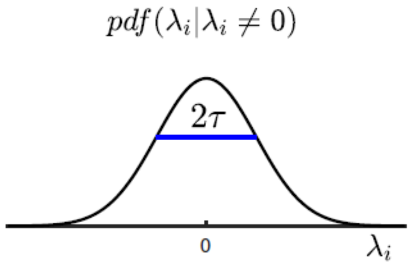
# Bayesian Testing a Point Null Hypothesis

$$x_i = \mu + \epsilon_i + \lambda_i$$

- $\mu$: (unknown) true value
- $\epsilon_i$: $N(0, u_i^2)$, $u_i^2$ => CMC
- $\lambda_i$: possibly overlooked effect
  - $H_{0,i}$: $\lambda_i = 0$; Prior $P_0(\lambda_i = 0) = \pi_0$
  - $H_{1,i}$: $\lambda_i \neq 0$; Prior $P_1(\lambda_i \neq 0) = 1 - \pi_0$

$P_0$ expresses prior belief about the probability of a lab-effect of zero.



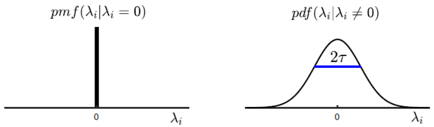

Alternatevely ($\lambda_i \neq 0$), a Gaussian prior is used.

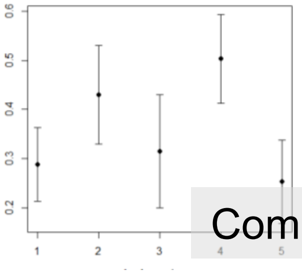
The hyperparameter $\tau$ expresses the belief about the size of lab-effect.

# Bayesian Testing a Point Null Hypothesis

$$x_i = \mu + \epsilon_i + \lambda_i$$

- $\mu$: (unknown) true value
- $\epsilon_i$: $N(0, u_i^2)$, $u_i^2$ => CMC
- $\lambda_i$: possibly overlooked effect
  - $H_{0,i}$: $\lambda_i = 0$ — Prior $P_0(\lambda_i = 0) = \pi_0$
  - $H_{1,i}$: $\lambda_i \neq 0$ — Prior $P_1(\lambda_i \neq 0) = 1 - \pi_0$



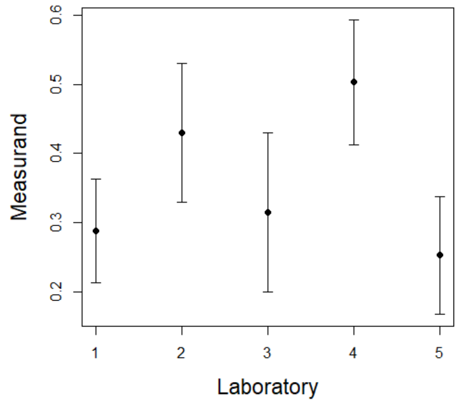

Comparison Data → $P_{0,posterior}(H_{0,i}|\boldsymbol{x}, \pi_0, \tau)$ ← Bayesian Analysis Tool

[Wübbeler et al., 2016]

# Bayesian Testing a Point Null Hypothesis

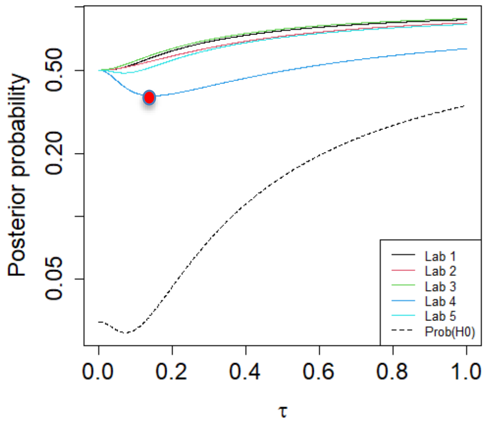
$$x_i = \mu + \epsilon_i + \lambda_i$$

**prior** comparison:
$P_0(\lambda_i = 0) = 0.5$




[Wübbeler et al., 2016]

Bayesian Analysis Tool



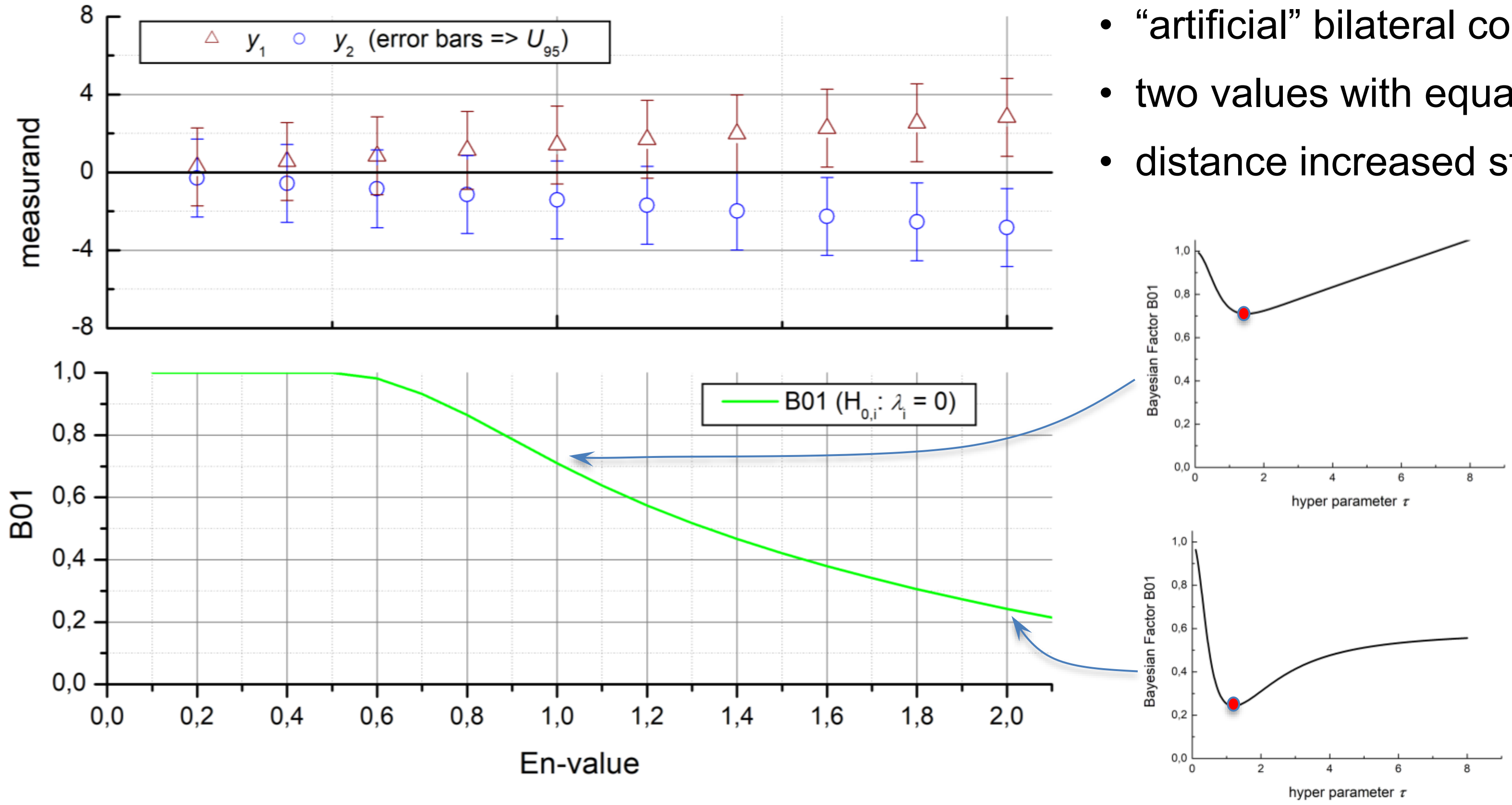

$$B01 = \frac{P_{0,post}(\lambda_i = 0)_{min}}{P_{1,post}(\lambda_i \neq 0)_{max}} \cdot \frac{P_{1,prior}}{P_{0,prior}}$$

$$\mathbf{B01} = \frac{0.3}{1-0.3} \cdot \frac{0.5}{0.5} = \mathbf{0.42}$$

**posterior**:
$P_{0,post}(\lambda_i = 0)_{min} = 0.3$

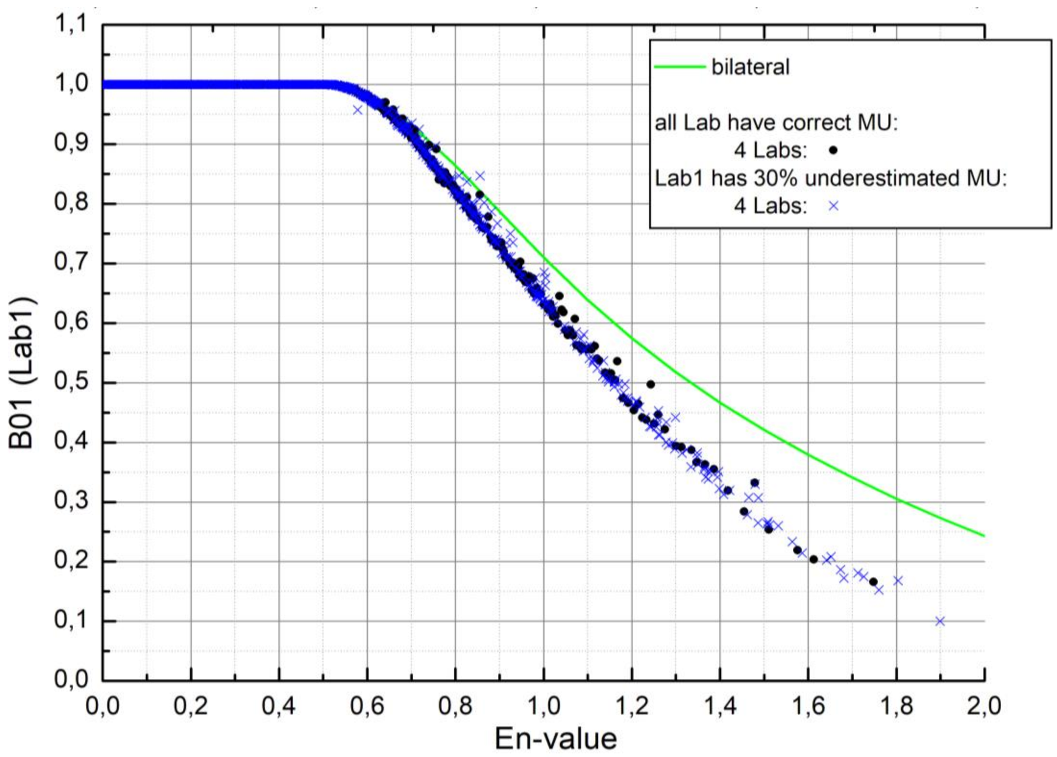
The tool gives the individual Bayes factors B01 for each Lab based on the data determined in the comparison.

# Most reduced case: bilateral comparison




- "artificial" bilateral comparison
- two values with equal MU
- distance increased stepwise

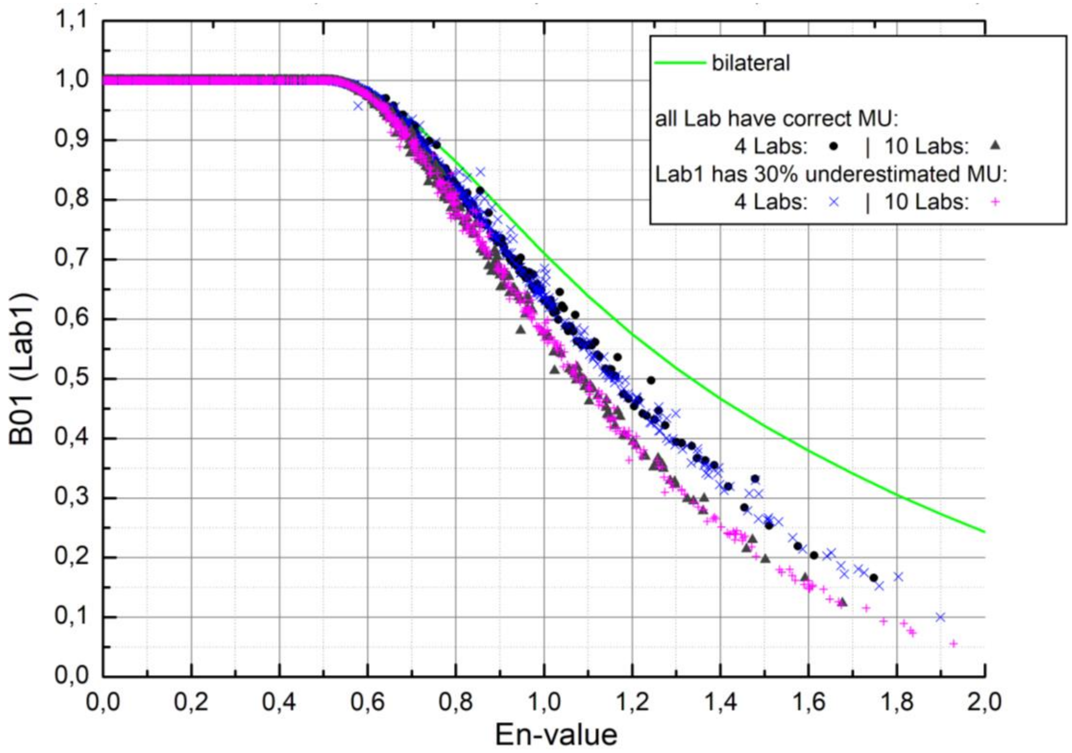
# MC-Results




- simulating comparison with <u>four</u> laboratories
- random-generated data of $N(0,u^2)$
- 10 000 trials (10% plotted)
- Case A: all laboratories have values according to their CMC.
- Case B: same as case A but Lab1 has 30% underrated uncertainty.

- The trend is similar for both cases
- There are more events at the right low tail in case 2 (underrated uncertainty of the Lab1)

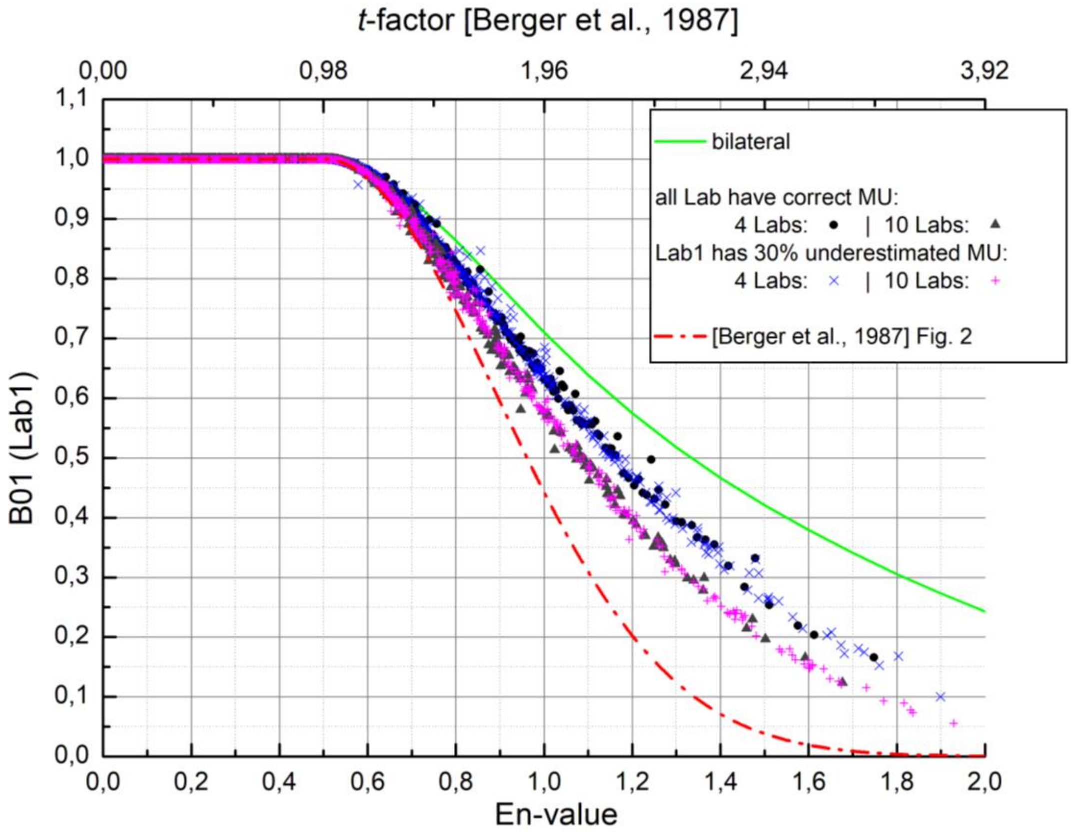
# MC-Results




- simulating comparison with 10 laboratories
- random-generated data of $N(0,u^2)$
- 10 000 trials (10% plotted)
- Case A: all laboratories have values according to their CMC.
- Case B: same as case A but Lab1 has 30% underrated uncertainty.

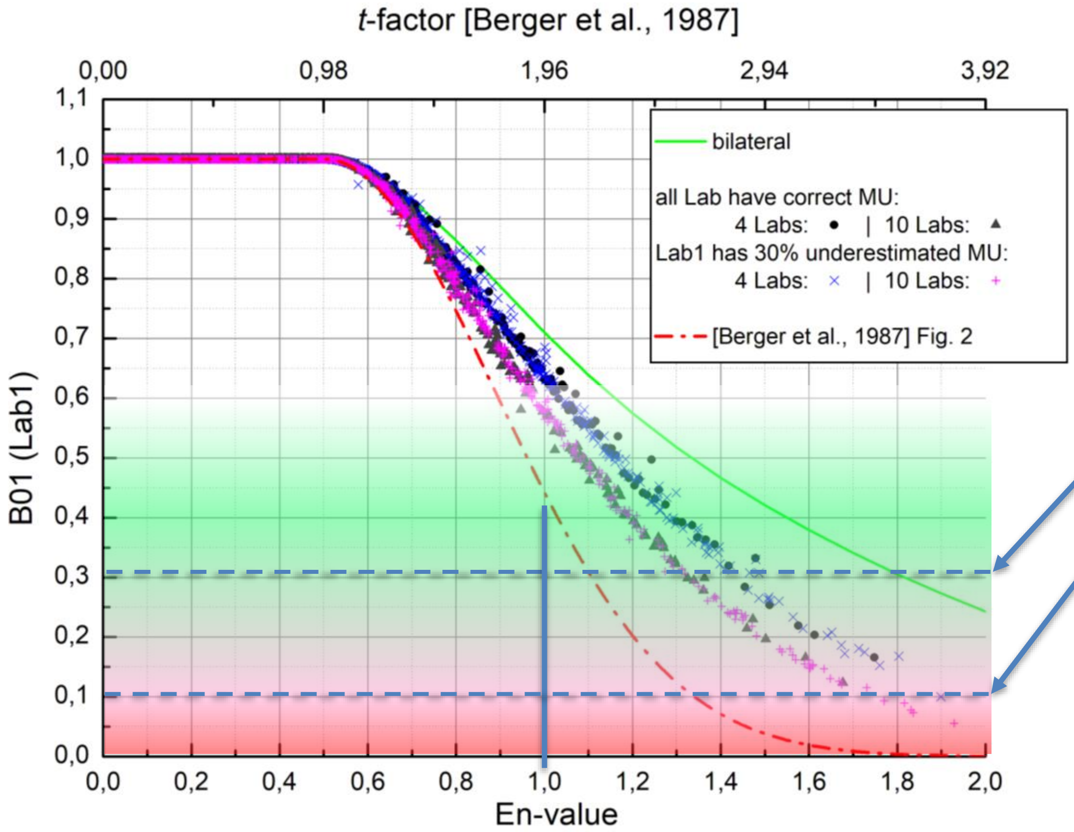
- Similar trend as in previous test
- Slightly steeper

# Comparing with findings in other publication




- In [Berger et al., 1987], basic work of Bayesian Evidence versus *p*-values was done.
- Calculation of Bayesian factors for one value under request against a known, independent reference.
- unimodal symmetric prior under $H_{1,i}$: $\lambda_i \neq 0$

# Bayesian Factors and Evidence




| B01 | Evidence for $H_1$ |
|---|---|
| 1 to 0.31 | not worth than a bare mention |
| 0.31 to 0.1 | substantial |
| 0.1 to 0.01 | strong |
| < 0.01 | decisive |

[Kass et al., 1995]

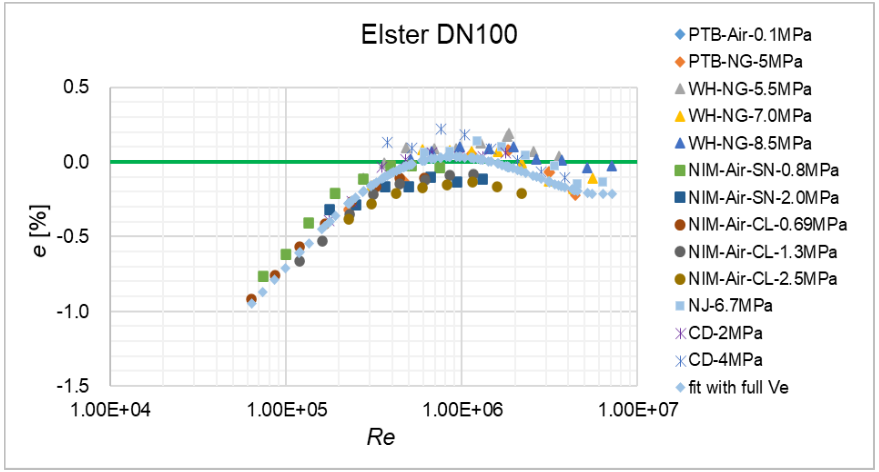
# Conclusions

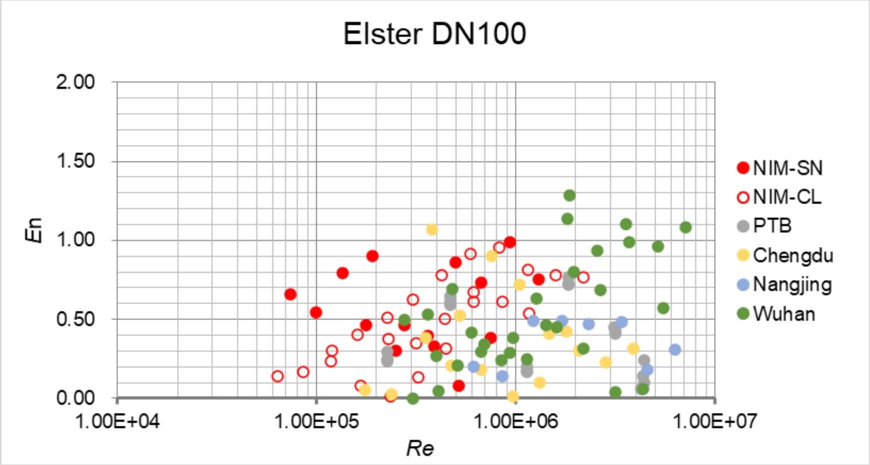
- The application of Bayesian hypothesis testing enables the usage of Bayesian factors as evidence indicators whether a claimed CMC is correct (*) or not.
- The ranges of Bayesian factors indicating sufficient evidence (see e.g. [Berger et al, 1987]) to reject the $H_0$-hypothesis (i.e. CMC is correct) correspond to En-scores larger than 1.
- The calculation scheme published in [Wübbeler et al., 2016] has been applied successfully to a large number of data out of key comparisons or to similar, simulated data.
- **The outcome of these calculations confirms that the "warning level" used in WGFF for measurement results with $1 < En \leq 1.2$ is in line with Bayesian hypothesis testing and is reasonable.**

(*) correct means here that the CMC-uncertainty covers sufficiently all potential effects, no overlooked effect exists

# Outlook

- **extending to curves**: (data example out of [Li et al., 2023])



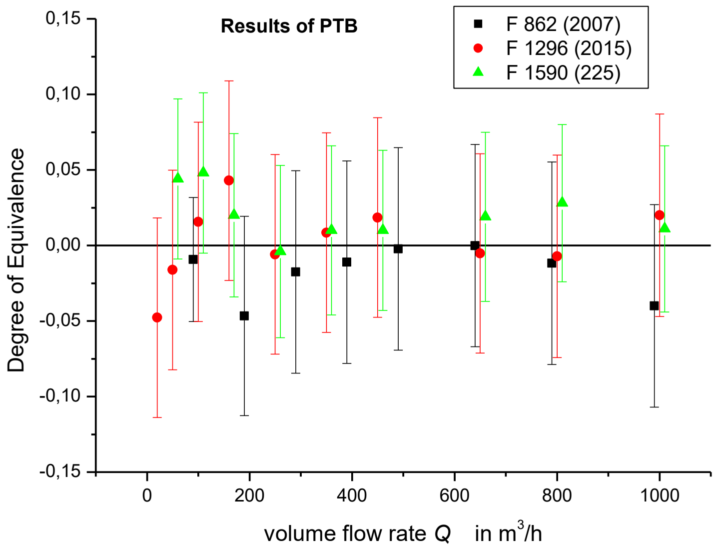

- No fixed flow point for testing defined in advance (only ranges and rough number of points).
- Curve (function) is defined by means of GLSF as KCRV.

=> much higher complexity for the algorithm because the reference value changes from scalar to vector (parameters of function)

# Outlook: investigating the usage of prior knowledge

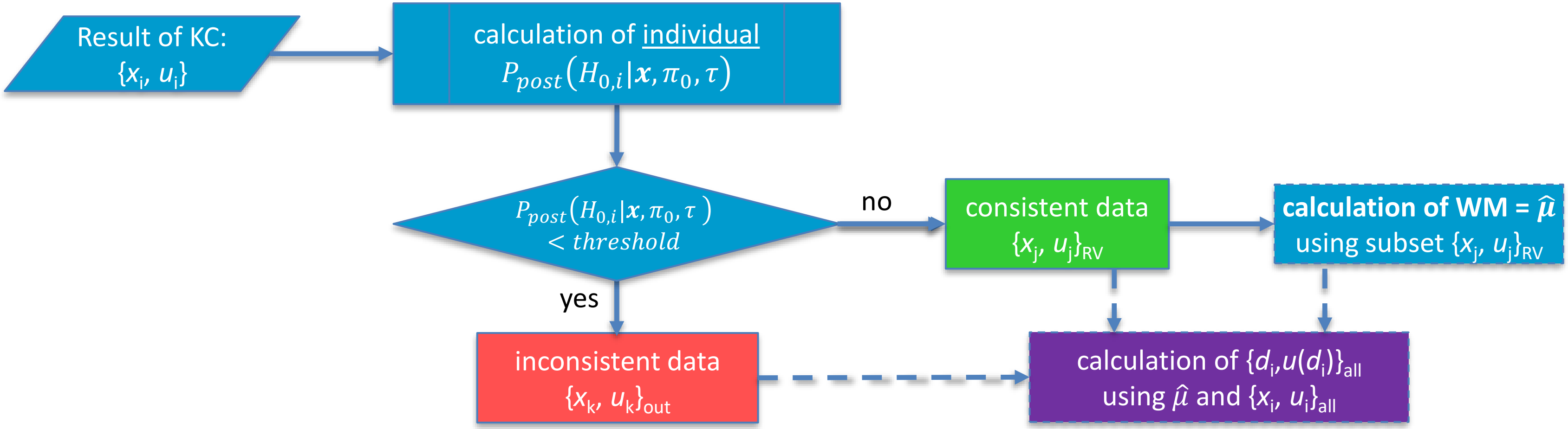
- extending to curves
- **investigating the usage of prior knowledge** (previous comparison results)

One good example will be the EURAMET-projects F 862/ F 1296/ F 1590 because comparison setting was almost the same and also the group of participants did not change so much.



# Outlook

- extending to curves
- investigating the usage of prior knowledge (previous comparison results)
- **establishing of clear rules for application**
- **looking for simplified approximation(s)** (making it easy for everyone)



# Reference


[Cox 2002] Cox, M.G. The evaluation of key comparison data. Metrologia **2002**, 39, 589–595.

[Cox 2007] Cox, M.G. The evaluation of key comparison data: Determining the largest consistent subset. Metrologia **2007**, 44, 187–200.

[ISO 17043] Conformity assessment — General requirements for the competence of proficiency testing providers.
ISO/IEC Edition 1, 2010 as well as ISO/IEC Edition 2, 2023

[Wübbeler et al., 2016] Wuebbeler, G., Bodnar, O. and Elster, C., Bayesian hypothesis testing for key comparisons. Metrologia **2016**, 53(4):1131.

[Berger et al, 1987] Berger, J. O. and Sellke, T. Testing a point null hypothesis: the irreconcilability of p values and evidence.
Journal of the American Statistical Association **1987**, 82(397):112–122.

[Kass et al., 1995] Kass, R. E. and Raftery, A. E. Bayes factors. Journal of the American Statistical Association **1995**, 90(430):773–795.

[Li et al., 2023] Li, C., Mickan, B. , Li, M., Ren, J., Wu, Y. and Xu, M.,
The comparison of the gas flow standards at high pressure. Measurement **2023**, Volume 223

**Physikalisch-Technische Bundesanstalt**
**Braunschweig and Berlin**
Bundesallee 100
38116 Braunschweig

Bodo Mickan
Telefon: 0531 592-1331
E-Mail: bodo.mickan@ptb.de
www.ptb.de