Accred Qual Assur (2014) 19:329–342
DOI 10.1007/s00769-014-1076-1

GENERAL PAPER

"Degrees of equivalence" for chemical measurement capabilities: primary pH

David L. Duewer · Kenneth W. Pratt · Chainarong Cherdchu · Nongluck Tangpaisarnkul · Akiharu Hioki · Masaki Ohata · Petra Spitzer · Michal Máriássy · Leoš Vyskočil

Received: 22 August 2013 / Accepted: 12 August 2014 / Published online: 10 September 2014

Abstract The key comparison (KC) studies of the Consultative Committee for Amount of Substance—Metrology in Chemistry help ensure the reliability of chemical and biochemical measurements relevant to international trade and environmental-, health-, and safety-related decision making. The traditional final evaluation of each measurement result reported by a KC participant is a "degree of equivalence" (DEq) that quantitatively specifies how consistent each individual result is relative to a reference value. Recognizing the impossibility of conducting separate KCs for all important analytes in all important sample matrices at all important analyte levels, emphasis is now shifting to documenting broadly applicable critical or "core" measurement competencies elicited through a series of studies. To better accomplish the necessary synthesis of results, data analysis and display tools must be developed for objectively and quantitatively combining individual DEqs. The information detailed in the 11 KCs of primary method pH measurements publically available as of 2013 provides an excellent "best case" prototype for such analysis. We here propose tools that enable documenting the expected primary pH measurement performance of individual participants between pH 1 and pH 11 and from 15 °C to 37 °C. These tools may prove useful for other areas where the uncertainty of measurement is a predictable function of the measured quantity, such as the stable gases. That results for relatively simple measurement processes can be combined using relatively simple analysis and display methods does not ensure that similarly meaningful summaries can be devised for less well understood and controlled systems, but it provides the incentive to attempt to do so.

Keywords CCQM · EAWG · Degree of equivalence · Key comparison · Leave-one-out strategy · Parametric Bootstrap Monte Carlo analysis · pH

Abbreviations

APMP	Asia Pacific Metrology Programme
CCQM	Consultative Committee for Amount of Substance—Metrology in Chemistry
CIPM	Comité International des Poids et Mesures
DEq	Degrees of equivalence
DL	DerSimonian–Laird
EAWG	Electrochemical Analysis Working Group
ESM	Electronic supplementary material
EUROMET	European Collaboration in Measurement Standards
GD	Graybill–Deal

Electronic supplementary material The online version of this article (doi:10.1007/s00769-014-1076-1) contains supplementary material, which is available to authorized users.

D. L. Duewer (✉) · K. W. Pratt
National Institute of Standards and Technology (NIST), 100 Bureau Drive, Gaithersburg, MD 20899-8390, USA
e-mail: david.duewer@nist.gov

C. Cherdchu · N. Tangpaisarnkul
National Institute of Metrology (Thailand) (NIMT), 3/4-5 Moo 3, Klong 5, Klong Luang, Pathum Thani 12120, Thailand

A. Hioki · M. Ohata
National Metrology Institute of Japan (NMIJ), 3-9 Tsukuba Central, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8563, Japan

P. Spitzer
Physikalisch-Technische Bundesanstalt (PTB), Bundesallee 100, 38116 Brunswick, Germany

M. Máriássy · L. Vyskočil
Slovenský metrologický ústav (SMU), Karloveská 63, 842 55 Bratislava, Slovakia

GUM	Guide to the expression of uncertainty in measurement
KC	Key comparison
KCRV	Key comparison reference value
LOO	Leave one out
MRA	Mutual recognition arrangement
NMI	National metrology institute
PBMC	Parametric Bootstrap Monte Carlo
RMO	Regional metrology organization
RV	Reference value
SI	International system of units

Functions

MAD	Median absolute deviation of a set of values from their median value
MAX	Maximum value of a set of values
MEDIAN	Median value of a set of values
$N(\mu,\sigma^2)$	Normal (Gaussian) distribution having mean μ and standard deviation σ
PTILE(p,d_{MC})	The p percentile of the set of all d_{MC} values
$\sum$	Summation of a series of values
$\cup$	Union of two or more sets of values

Symbols

d	DEq for a single reported result for a specific NMI for a specific buffer
d_{MC}	PBMC estimate of d
D	Combination of available d over temperature for a specific NMI for a specific buffer
$\boldsymbol{Đ}$	Combination of available d over temperature and buffers for a specific NMI
k_{95}	Coverage factor providing a 95 % level of confidence coverage interval
MC	Subscript designating a relationship to PBMC analysis
n	Number of x in a given set
n_{MC}	Number of PBMC samplings of a complete set of data
N	Number of temperature-specific d available for estimating D or the number of buffer-specific D available for estimating $\boldsymbol{Đ}$
p	Probability expressed as a percentage (i.e., on the range 0–100)
pa^0	The acidity function at zero added chloride
ρ	Correlation between two quantities
s	Standard deviation
s_{GD}	Graybill–Deal weighted standard deviation (also called "external consistency")
S	Subscript designating a successor KC
t	Subscript designating a particular result in a series of evaluation temperatures
T	Number of evaluation temperatures for a given buffer
u_{GD}	Standard uncertainty estimated as a GD weighted standard deviation
u_{MAD}	Standard uncertainty estimated from the MAD
u_{SD}	Standard uncertainty estimated from s and $\bar{u}$
u	Standard uncertainty
$\bar{u}$	Pooled value of a set of u; i.e., the square root of the mean of the squared u values
U_{95}	One-half of a 95 % level of confidence symmetric coverage interval
$^{-}U_{95}$	Lower bound of a 95 % level of confidence asymmetric coverage interval
$^{+}U_{95}$	Upper bound of a 95 % level of confidence asymmetric coverage interval
V_{KC}	Reference value for the root KC
V_R	Reference value estimated from anchor participant results in the root KC
V_S	Reference value estimated from anchor participant results in successor KCs
x	Reported value
x_{DL}	DerSimonian–Laird weighted mean of a set of x
x_{GD}	Graybill–Deal weighted mean of a set of x
x_{mean}	Arithmetic mean of a set of x
x_{median}	Median of a set of x
x_{adj}	Value reported in a successor study re-centered onto the reference value of a given earlier study

Introduction

The Comité International des Poids et Mesures (CIPM) is responsible for the conduct of international key comparison (KC) studies that enable national metrology institutes (NMIs) and related organizations to document measurement capabilities relevant to international trade and environmental-, health-, and safety-related decision making. The technical supplement to the 1999 Mutual Recognition Arrangement (CIPM MRA) [1] establishes the process by which NMIs demonstrate the "degree of equivalence" (DEq) of national measurement standards. The CIPM MRA states that (1) KCs lead to reference values, (2) a key comparison reference value (KCRV) is expected to be a good indicator of an international system of units (SI) value, (3) DEqs refer to the degree to which a national measurement standard is consistent with the KCRV, and (4) DEqs for measurement standards are expressed quantitatively by the deviation from the KCRV and the uncertainty of this deviation at a 95 % level of confidence.

The Working Groups of the Consultative Committee for Amount of Substance—Metrology in Chemistry (CCQM)

Table 1 pH-related key Comparisons

Buffer	Designation	Year	Results reported as	Number of results used in KCRV or RV			Original estimators	
				15 °C	25 °C	37 °C	V_{KC}	V_R
Phosphate	CCQM-K9	1999	pH & pa^0	9	9	8	x_{GD}, u_{GD}	
Phosphate	CCQM-K9.1	2000	pa^0	1	1	1		x, $u(x)$[a]
Phosphate	CCQM-K9.2	2006	pa^0	2	2	2		x_{mean}, u^c_{SD}
Phosphate	APMP-K9	2009	pa^0	3	3	3		x_{mean}, u^c_{SD}
Phthalate	CCQM-K17	2001	pH	11	11	11	x_{GD}, u_{GD}	
Phthalate	EUROMET.QM-K17	2003	pH	1	1	1		x, $u(x)$[a]
Carbonate	CCQM-K18	2006	pa^0	0[b]	12	0[b]	x_{median}, u_{MAD}	
Carbonate	CCQM-K18.1	2007	pa^0	1	1	1		x, $u(x)$[a]
Borate	CCQM-K19	2005	pa^0	10	11	11	x_{median}, u_{MAD}	
Borate	CCQM-K19.1	2010	pa^0	3	3	3		x_{mean}, u^c_{SD}
Tetroxalate	CCQM-K20	2007	pa^0	9	10	10	x_{GD}, u_{GD}	

[a] Linkage to V_{KC} through the result of the single anchor participant

[b] This temperature was not included in the KC design

[c] Linkage to V_{KC} through the mean, standard deviation, and pooled $u(x)$ of two or more anchor participants

are responsible for selecting and overseeing the operation of KCs that address chemical (and biochemical) measurements. Few such measurements directly realize an SI unit: a mole of one chemical analyte may have no physiochemical properties in common with a mole of another beyond containing the same number of entities. Further, with a few exceptions such as atmospheric ozone [2], the higher order chemically related measurements made by an NMI do not reflect "national measurement standards" but rather the organization's measurement capabilities at a given time. However, until recently most CCQM-sponsored KCs have attempted to keep as closely as possible to the philosophy of the CIPM MRA as described above by estimating a separate DEq for each reported result in each KC.

Recognizing the impossibility of conducting separate KCs for all important chemically related analytes in all important sample matrices (and the ever-increasing resource burdens placed on the world's NMIs by attempting to address even a tiny subset of these measurands), several of the Working Groups within the CCQM are now using KCs to evaluate a series of critical or "core" measurement competencies. While continuing to provide DEqs for the results reported in individual KCs, the overall assessment of an NMI's measurement capabilities may require combining DEqs for several different measurands that may be estimated in different KCs and at separate times.

The KCs conducted by the CCQM Electrochemical Analysis Working Group (EAWG) and two regional metrology organizations (RMOs) on primary pH-related measurements are an excellent, and prescient, model for such studies. Initiated in 1999, to date results are publicly available for 11 KCs involving five buffer systems, with all but one of these systems characterized at 15 °C, 25 °C, and 37 °C (see Table 1). While individual NMIs routinely if informally assess their primary pH measurement capabilities by qualitative comparison of the various DEqs for different temperatures and buffers, no formal mechanism currently exists for quantitatively summarizing such results.

We here propose quantitative data analysis methods for combining individual DEqs from multiple KCs to estimate an NMI's measurement capabilities for particular measurement areas. We will show that the various primary pH measurements can be combined to document the expected measurement performance for primary pH measurements from pH 1 to pH 11 and from 15 °C to 37 °C. These data analysis methods represent a first step in the development of tools for assessing NMI measurement capabilities from less coherent evidence.

Data

Sources

The data used in this study are the results of primary method pH measurements as provided in the published Final Reports [3–13] of the KCs listed in Table 1. All of the primary pH measurement data given in these reports are listed in Tables S1.a to S5.a of the electronic supplementary material (ESM), with the exception of values that (1) were identified in the KC's final report as technically flawed and as such were excluded from the reference value (RV) estimation process for that KC and (2) are not the most recent primary pH measurement in that buffer system for the NMI that submitted the excluded result. Table 2 lists the number of DEq estimates available for each NMI

for each buffer system. As the focus of this report is the process of combining results rather than particular outcomes for these data, each NMI is designated as a single-letter alphabetical code.

The 11 KCs considered include five "root" comparisons of pH measurements made in different buffer systems: CCQM-K9 (phosphate), CCQM-K17 (phthalate), CCQM-K18 (carbonate), CCQM-K19 (borate), and CCQM-K20 (tetroxalate). These root KCs were activities of the EAWG. The remaining studies, formally differentiated as "Subsequent KCs" and "Regional KCs" but here referred to as "successor" KCs, are each linked to one or another of the roots through the use of in-common measurement protocols and qualitatively similar buffer solutions. The four successor studies CCQM-K9.1, -K9.2, -K18.1, and -K19.1 were activities of the EAWG; the integer part of the label designates the root KC and the decimal designates the temporal order of the successor KC relative to its root. The APMP.QM-K9 and EUROMET.QM-K17 (also termed EUROMET Project 696) KCs were activities of the Asia Pacific Metrology Programme and the European Collaboration in Measurement Standards RMOs, respectively, both in collaboration with the EAWG. All of the successor studies were designed to enable additional NMIs to demonstrate newly acquired pH measurement capabilities and/or to allow participants in earlier studies to document improved capabilities.

The KCs examined in this study, all with completion dates ranging from 1999 to 2010, constitute the initial cycle of primary pH KCs. The recently completed CCQM-K91 (phthalate) [14] is the first KC of the second cycle and is not included in this study. CCQM-K91 and the other pH studies currently in progress or planned are designed as fresh root comparisons rather than maintaining linkages to the earlier studies.

Primary method pH measurements

All of the data considered here are the primary pH measurements reported by KC participants for a buffer solution prepared and distributed by the coordinator of each KC. The direct result of the primary measurement itself is pa^0, the acidity function at zero added chloride. Depending on the KC design, pa^0 determinations were made at one or more specified temperatures. The metrological basis for the primary measurement of pH is discussed in detail elsewhere [15–17]. In essence, the pa^0 is a function of the potential of a specified type of electrochemical cell, commonly referred to as the Harned cell.

Table 2 Participation history

Code[a]	Number of DEq estimates					
	Phosphate	Phthalate	Carbonate	Borate	Tetroxalate	Total
A		3	1	3		7
B				3		3
C			1	2	2	5
D	3					3
E	3		1	3	3	10
F	3	3	1	3		10
G	3	3		3		9
H			1		3	4
I	2	3	1	3	3	12
J	3	3	1	3	3	13
K	3	3	1	3	3	13
L	3	3	1	3	3	13
M	3	3	1	3		10
N	3	3	1	3	3	13
O	3	3		3		9
P	3		1	3	3	10
Q	3	3		3	3	12
R	3	3	1		3	10
S	1					1
T			1		3	4
Total participants	15	12	14	15	12	68
Total DEq estimates	42	36	14	44	35	171

[a] Single alphanumeric character unique to each participating NMI

The pH is obtained from pa^0 by adding a constant term, defined by the Bates–Guggenheim convention, specific for a given buffer and temperature [15, 18]. Since the value of this term is invariant among the participants of each KC, all measurement-specific factors that affect the pa^0 affect the corresponding pH values (as well as any KCRV calculated from them) to the same extent. The uncertainty [15] of the Bates–Guggenheim convention is excluded from the reported uncertainties for the pH KCs. This exclusion avoids inflating the reported uncertainties for the pH KCs and ensures that the reported uncertainties relate to the measurement capabilities per se of the participants.

Measurements for the carbonate, borate, and tetroxalate buffer KCs are recorded in the Final Reports as the reported pa^0 values. Measurement results for some of the phosphate and phthalate buffer system KCs were recorded as pH values. We consider the recorded values for all of these KCs as being of the same kind: "primary pH".

Note that primary pH is a procedurally defined *kind-of-quantity* [19]. Since primary pH cannot be determined except through the measurement process itself, the KCRV for a primary pH KC must be estimated from the measurement results even though the study materials are prepared quantitatively from materials of established composition. This is in contrast to some chemical systems (such as synthetic gas mixtures and organic and inorganic calibration solutions) where materials can be prepared to have well-defined compositions that, with suitable verification, provide KCRVs that are independent of results reported by the study's participants.

Computation

All calculations used in this study were performed in a spreadsheet environment using a modern desktop computer. Purpose-built programs in languages native to this environment were used to automate repetitive computations. Versions of these tools are available on request from the corresponding author.

Results and discussion

"National standard" degrees of equivalence as currently estimated

As defined by the CIPM MRA, the DEq, d, for a particular KC result is estimated as

$$d = x - V_{\mathrm{KC}} \tag{1}$$

where x is the reported value and V_{KC} is the KCRV and is a close realization of an SI value as assigned by the sponsoring Working Group and approved by the Consultative Committee.

Using formal variance propagation, the uncertainty associated with d should be estimated as [20],

$$u(d) = \sqrt{u^2(x) + u^2(V_{\mathrm{KC}}) - 2\rho(x, V_{\mathrm{KC}})u(x)u(V_{\mathrm{KC}})} \tag{2}$$

where $u(x)$ is the standard uncertainty associated with x, $u(V_{\mathrm{KC}})$ is the standard uncertainty of the V_{KC}, and $\rho(x,V_{\mathrm{KC}})$ is the correlation between the reported value and the KCRV. Within at least the CCQM, except when the KCRV has been assigned using the Graybill-Deal estimator [21, 22], the $\rho(x,V_{\mathrm{KC}})$ term has generally been ignored—effectively asserting that $\rho(x,V_{\mathrm{KC}}) = 0$.

Since the MRA requires that uncertainties are to be specified at the 95 % level of confidence, standard uncertainties must usually be estimated from reported expanded uncertainties

$$u(x) = \frac{U_{95}(x)}{k_{95}}; \quad u(V_{\mathrm{KC}}) = \frac{U_{95}(V_{\mathrm{KC}})}{k_{95}} \tag{3}$$

where k_{95} is the coverage factor expected to yield an expanded uncertainty such that the interval $x \pm k_{95}u(x)$ includes the true value with a 95 % level of confidence. The desired 95 % level of confidence expanded uncertainty on d, $U_{95}(d)$, is likewise typically estimated as

$$U_{95}(d) = k_{95} \cdot u(d). \tag{4}$$

Again, within at least the CCQM, k_{95} has generally been asserted to be 2 regardless of how the various quantities are actually estimated.

"Measurement capability" degrees of equivalence for a given buffer

Given N individual $d \pm U_{95}(d)$ estimates for a particular NMI and assuming that they are independently drawn from a relatively normal distribution, a combined "measurement capability" DEq, $D \pm U_{95}(D)$, for that NMI can be estimated from the mean of the d, the standard deviation of the d, and the pooled $U_{95}(d)$

$$\begin{aligned} D &= \sum_{i=1}^{N} d_i \Big/ N; \quad u(D) = \sqrt{\bar{u}^2(d) + s^2(d)} \\ U_{95}(D) &= 2\,u(D) \\ \bar{u}^2(d) &= \sum_{j=1}^{N} \left(\frac{U_{95}(d_i)}{2}\right)^2 \Big/ N \\ s^2(d) &= \begin{cases} 0 & N = 1 \\ \sum_{i=1}^{N} (d_i - D)^2 \Big/ (N-1) & N < 1 \end{cases} \end{aligned} \tag{5}$$

where i indexes over the individual estimates. This $U_{95}(D)$ estimated in this manner can be considered as conservatively large since the among-temperature

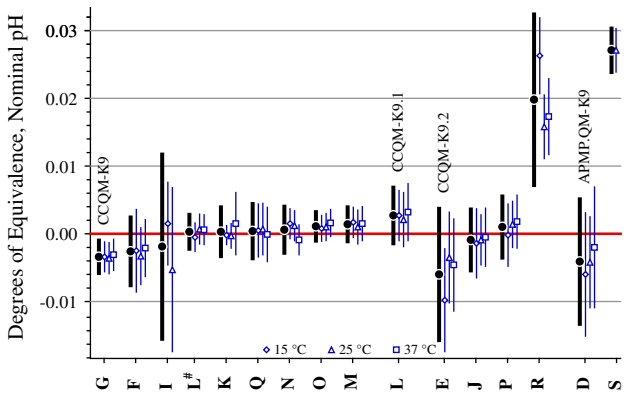

Fig. 1 Dot-and-bar plot of degrees of equivalence estimated by variance propagation for all participants in CCQM-K9, -K9.1, -K9.2, or APMP.QM-K9 who reported primary method pH results. The *vertical axis* displays degrees of equivalence, $D \pm U_{95}(D)$ and $d \pm U_{95}(d)$. The *horizontal axis* is used to separate the NMIs. The *filled circles* and *thick vertical lines* represent the combined $D \pm U_{95}(D)$ for each NMI as estimated from Eq. 5. The NMIs are sorted in order of increasing D within each KC; the KC is identified above the results for the participant with the lowest-valued D within that KC. The *open symbols* and *thin vertical lines* represent $d \pm U_{95}(d)$ for measurements made at 15 °C (*diamond*), 25 °C (*triangle*), and 37 °C (*square*) as specified in the KC Final Reports. The *thick horizontal line* represents zero bias; the *thin horizontal lines* are visual guides

variability, estimated from the standard deviation of the d_i, includes contributions from the within-temperature variability, estimated as the pooled $U_{95}(d_i)/2$. However, these $U_{95}(D)$ will always be at least as large as the expected within-temperature $U_{95}(d_i)$ and will closely approach $2{\cdot}s(d)$ as between-temperature differences become dominant. Note that $u(D)$ is not scaled by $\sqrt{N}$ since $D \pm U_{95}(D)$ is intended to be characteristic of individual measurement processes rather than any estimate of the central tendency of N processes. The variance propagation results for all five buffer systems are listed in Tables S1.b to S5.b of the ESM, along with the d and $u(d)$ recalculated from the reported results as listed in Tables S1.a to S5.a.

Of course, that the $d \pm U_{95}(d)$ can be mathematically combined does not address the question as to whether combining them is chemically reasonable. Figure 1 displays the $d \pm U_{95}(d)$ for all NMIs that reported primary pH results in the CCQM-K9, -K9.1, K9.2, and APMP.QM-K9 studies of the phosphate buffer system along with the combined $D \pm U_{95}(D)$. These $d \pm U_{95}(d)$ estimates are taken directly from the Final Reports or calculated using the data and formulae provided in those reports. The coherence of the $d \pm U_{95}(d)$ over the three temperatures for nearly all of the NMIs suggests that combining the individual estimates is reasonable. If the validity of the combination is accepted, then the $D \pm U_{95}(D)$ provides a snapshot of the NMI's phosphate buffer primary pH measurement capabilities from 15 °C to 37 °C.

Revisiting the estimation of degrees of equivalence

Since estimating $D \pm U_{95}(D)$ is outside the scope of the CIPM MRA's "measurement standard" paradigm, the question arises whether even more informative estimates could be achieved using data analysis approaches that do more than just propagate reported summary estimates.

Key comparison reference value, V_{KC}

While many location estimators have been proposed for evaluating a KCRV and recent guidance provided for choosing and calculating ones appropriate to particular circumstances [23], all of the KCs considered here have used either the median when there was significant between-result variance, s_b^2, or the Graybill–Deal weighted mean [21], x_{GD}, when s_b^2 was considered insignificant. The x_{GD} is defined as

$$x_{GD} = \sum_{i=1}^{n} \frac{x_i}{u^2(x_i)} \Bigg/ \sum_{i=1}^{n} \frac{1}{u^2(x_i)} \tag{6}$$

where i indexes over all the accepted results in a KC and n is the number of such results. Three of the root KCs (CCQM-K9, -K17, and -K20) used x_{GD} as their KCRV estimate for all temperatures studied.

It is now better appreciated that use of x_{GD} is justified only in the unusual case where s_b^2 is both truly zero and all of the $u(x)$ are credible. For situations where s_b^2 is appreciable but the x follow an approximately unimodal symmetric distribution and the $u(x)$ are at least plausible, the DerSimonian–Laird (DL) [24] weighted mean, x_{DL}, is more appropriate [25]. Commonly used in clinical meta-analysis, x_{DL}, is identical to x_{GD} when s_b^2 is zero but approaches the arithmetic mean, x_{mean}, as s_b^2 becomes large relative to the $u(x_i)$. The x_{DL} is defined as

$$x_{DL} = \sum_{i=1}^{n} \frac{x_i}{s_b^2 + u^2(x_i)} \Bigg/ \sum_{i=1}^{n} \frac{1}{s_b^2 + u^2(x_i)}$$

$$s_b^2 = \mathrm{MAX}\left[0, \left(\sum_{i=1}^{n} \frac{(x_i - x_{GD})^2}{u^2(x_i)} - n + 1\right) \Bigg/ \left(\sum_{i=1}^{n} \frac{1}{u^2(x_i)} - \sum_{i=1}^{n} \frac{1}{u^4(x_i)} \Bigg/ \sum_{i=1}^{n} \frac{1}{u^2(x_i)}\right)\right] \tag{7}$$

where "MAX" is the function "return the largest value of the arguments." Since x_{DL} asymptotically approaches x_{mean}, it is as sensitive as x_{mean} itself to the presence of discordant results and is only appropriately used after any and all such results have been identified, reviewed by the submitting NMI, and excluded if a cause for the discordance is identified.

Due to what was considered appreciable s_b^2, the CCQM-K18 and -K19 studies used the median of the accepted x,

x_{median}, to estimate the KCRV at each temperature studied. While appropriate for any distribution and robust to minority populations of discordant values, x_{median} is not a very efficient estimate of location (that is, it is more variable than x_{mean} when applied to normally distributed data) and does not make use of any information provided by the $u(x)$ even when they are quite informative [26].

Standard uncertainty of the key comparison reference value, $u(V_{KC})$

The three root KCs that used x_{GD} as their KCRV estimates regarded that estimator's usual standard uncertainty, $u(x_{GD})$,

$$u(x_{GD}) = \sqrt{1 \Big/ \sum_{i=1}^{n} \frac{1}{u^2(x_i)}}, \tag{8}$$

as too small for use as the $u(V_{KC})$. Instead, a weighted standard deviation estimated using the same inverse-variance weighting used to define x_{GD} was used to provide estimates that take non-zero s_b into account. While sometimes referred to as the "external consistency" uncertainty [3, 7, 27], this estimate is more simply termed the "Graybill–Deal weighted standard deviation" and is defined as

$$u_{GD}(x_{GD}) = \sqrt{\sum_{i=1}^{n} \frac{1}{u^2(x_i)} \frac{(x_i - x_{GD})^2}{n-1} \Big/ \sum_{i=1}^{n} \frac{1}{u^2(x_i)}} \tag{9}$$

While providing more chemically reasonable $u(V_{KC})$ for these studies than does $u(x_{GD})$, this approach does not address the x_{GD}'s bias towards x that have very small $u(x)$.

The two studies that estimated the KCRV values as the x_{median} used a scaled version of the robust median absolute deviation from the median (MAD) dispersion estimate to estimate $u(V_{KC})$:

$$\begin{aligned} u_{MAD}(V_{KC}) &= \mathrm{MAD}(x) \frac{1.858}{\sqrt{n-1}} \\ &\equiv \mathrm{MEDIAN}(|x - x_{median}|) \frac{1.858}{\sqrt{n-1}} \end{aligned} \tag{10}$$

where MEDIAN is the function "find the median value of the specified list of values" and the scaling factor of $1.858/\sqrt{(N-1)}$ adjusts the estimate to (1) have the approximately the same coverage as a standard deviation for normally distributed data, (2) compensate for the lower efficiency of x_{median} relative to x_{mean}, and (3) compensate for the relatively small N. While robust to the inclusion of discordant values, the MAD is inefficient compared to the standard deviation when applied to normally distributed data.

While various approaches for estimating uncertainties for weighted means have been proposed that provide more efficient coverage intervals [28, 29], the original estimate associated with x_{DL} is [24]

$$u(x_{DL}) = \sqrt{1 \Big/ \sum_{i=1}^{n} \frac{1}{s_b^2 + u^2(x_i)}}. \tag{11}$$

Linkages between studies

The CIPM MRA does not specify how results from successor KCs are to be linked to those of a root KC; however, it does mandate [1] that "The results of the RMO key comparisons are linked to key comparison reference values established by CIPM key comparisons by the common participation of some institutes in both CIPM and RMO comparisons. The uncertainty with which comparison data are propagated depends on the number of institutes taking part in both comparisons and on the quality of the results reported by these institutes." The CCQM has chosen to link successor and RMO KCs using the same general methods.

When a successor or RMO KC uses materials and methods that are sufficiently similar to those used in a root—as is the case for the primary pH studies considered here, the studies can be directly linked through results provided by one or more "anchor" NMIs who successfully participated in a prior KC. For example, results in the successor CCQM-K9.1 are linked to the KCRV of the root CCQM-K9 through results provided by one anchor who made full sets of measurements in both studies, CCQM-K9.2 is linked to CCQM-K9 through the results of two such anchors, and APMP.QM-K9 is linked through results of one anchor from CCQM-K9, one from CCQM-K9.1, and one from CCQM-K9.2. The linkages for all of the pH studies considered here are detailed in Tables S1.a to S5.a of the ESM.

To date, degrees of equivalence for participants in a successor pH KC have been estimated using a "National standard" paradigm assuming that DEq are unchanging over time and samples:

$$d = x - V_{KC} + V_R - V_S$$

$$u(d) = \sqrt{u^2(x) + u^2(V_{KC}) + u^2(V_R) + u^2(V_S) - 2\rho(V_R, V_{KC})u(V_R)u(V_{KC})} \tag{12}$$

$$U_{95}(d) = k_{95}u(d)$$

where V_R is a reference value estimated from the results of the anchor participants in previous studies, $u(V_R)$ is its estimated standard uncertainty, V_S is a reference value estimated from the results of the anchor participants in the successor KC, $u(V_S)$ is its estimated standard uncertainty, and $\rho(V_R, V_{KC})$ is the correlation between prior studies' reference values and the KCRV. Although V_R has (nearly) always been estimated from a subset of the participants in the root KC, none of the other quantities are estimated from the same data sets and so are not expected to be strongly correlated. As with the $d \pm U_{95}(d)$ estimated for the participants in the root KC, $\rho(V_R, V_{KC})$ has typically been ignored and k_{95} asserted to be 2.

In the successor studies involving two or more anchor participants, V_R and V_S have been estimated from x_{mean}; the standard deviation, $s(x)$; and pooled uncertainty of the anchor participants' results, $\bar{u}(x)$. The V_S and its standard uncertainty, $u(V_S)$, are readily estimated:

$$V_S = \sum_{j=1}^{n} x_{Sj} \Big/ n; \qquad u(V_S) = \sqrt{\frac{\bar{u}^2(x_S) + s^2(x_S)}{n}}$$

$$\bar{u}^2(x_S) = \sum_{j=1}^{N} u^2\left(x_{Sj}\right) \Big/ n \tag{13}$$

$$s^2(x_S) = \begin{cases} 0 & n = 1 \\ \sum_{i=1}^{N} \left(x_{Si} - V_S\right)^2 \Big/ (n-1) & n > 1 \end{cases}$$

where j indexes over the anchors, n is the number of anchors, x_S are the results for the anchors in the successor KC, and $u(x_S)$ are the standard uncertainties for the anchor values.

When all anchors successfully participated in the *same* prior KC, the estimation process for the prior reference value, V_R, is analogous to the above with the x_S replaced by x_R. However, when some of the anchors participated in different studies (as in APMP.QM-K9), the "national standard" paradigm re-centers all of the anchor values to have the value they "should have had":

$$x_{adj} = x_R + d_R; \quad u\left(x_{adj}\right) = u(d_R) \tag{14}$$

where x_{adj} designates a re-centered value, x_R is the value in the most recent KC that the anchor successfully participated in, d_R is the DEq in that KC, and $u(d_R)$ is its standard uncertainty. The uncertainty associated with x_R, $u(x_R)$, is not included the calculation of $u\left(x_{adj}\right)$ since it is already included in $u(d_R)$.

The measurement capability paradigm suggests a much simpler calculation. If a participant's result does not reflect the fixed bias of a national standard, successful participation in a prior KC implies only that all anchor participants are expected to routinely realize true values within their assessed uncertainties. The DEq for the non-anchor participants in the successor KC is thus independent of results in the root KC:

$$d = x - V_S$$
$$u(d) = \sqrt{u^2(x) + u^2(V_S)}; \quad U_{95}(d) = k_{95}u(d) \tag{15}$$

When there is only one anchor participant, the k_{95} expansion factor in Eqs. 12 and 15 must be assigned by expert judgment.

Reference value estimators

When there is more than one anchor participant in a successor KC, using Eq. 13, i.e., estimating V_S as x_{mean}, does not make efficient use of the information provided in the reported $u(x)$. As in the estimation of the KCRV, estimating V_S as x_{DL} (Eq. 7) and $u(V_S)$ as $u(x_{DL})$ (Eq. 11) makes more complete use of the available information. Further, use of the same estimators for the V_{KC} and Vs provides a philosophically consistent approach to the analysis of the successor KCs.

Leave-one-out reference values

Estimating a KCRV using all accepted results can be considered to provide the closest realization of an SI unit that can be estimated using a consensus process. However, using that KCRV to estimate the $d \pm U_{95}(d)$ for a $x \pm U_{95}(x)$ used in the determination of the KCRV may result in non-negligible values for the often-ignored $\rho(x,V_{KC})$ term in Eq. 2. This can be avoided by estimating each $d \pm U_{95}(d)$ relative to a reference value that is independent of the associated $x \pm U_{95}(x)$. At the cost of additional calculations and an $\sqrt{n/(n-1)}$ increase in the estimated uncertainty, the same estimator used for V_{KC} can provide individual reference values for the $d \pm U_{95}(d)$ for each $x \pm U_{95}(x)$ using all of the accepted results except itself. This leave-one-out (LOO) approach is a routine tool for assessing the predictive utility of regression models [30]. LOO is a particularly useful tool for identifying the influence of particular values on consensus summaries and the consequences of such inclusion on the other values [31].

When the measurement capability linkage of Eq. 15 is used, the LOO-estimated DEq for participants in a root KC does not impact the DEq estimated for participants in successor studies since these are linked only to the KCRV of the root and the measurements made by the anchor participants in the successor KC itself. In any case,

eliminating the potential distortion from ignoring non-zero $\rho(x,V_{KC})$ places the $U_{95}(d)$ estimates for root and successor KC participants on more equal footing.

Use of corrected and imperfect results

It can happen that an NMI recognizes computational oversights only after the results of a KC have been revealed. While the DEq for such an NMI must be estimated from the originally reported results, when the error results from miscalculation then the WG may choose to use a transparently corrected result in determining the KCRV. In CCQM-K9, the NMI who reported the errant result had to demonstrate its capability in a successor KC. In these circumstances, an issue arises when the NMI is an anchor in a later successor: which result should be used as the link? The approach used by the EAWG has been to link to the result from the successor KC. However, since the KCRV of the root KC is based in part on that NMI's corrected result, linkage through the corrected result shortens the linkage chain for later participants without further compromise. As this shortening does not benefit the anchor participant but impacts only those NMIs that are linked through that anchor, "measurement capability" DEq should be based on the most direct valid linkage.

Occasionally, too, results are reported that are valid in their own right but that are excluded from formal inclusion in the KC and so cannot be used to estimate a national standard DEq. Such exclusions include but are not limited to measurements made at not quite the KC's design conditions and values submitted without an accompanying uncertainty budget. Given that the proposed process for combining results is already well outside the scope of the CIPM MRA's paradigm, it seems reasonable to try to make use of such data after conservative adjustment. For example, (1) measurements made at an off-target temperature could be interpolated to the target if the approximate temperature dependence of the measurements can be estimated or (2) missing uncertainties could be estimated as the "worst case" of previously supplied complete data, assuming that sufficient such data were available. While it would be inappropriate to base critical decisions primarily on resurrected data, ignoring available information is inefficient.

Parametric Bootstrap Monte Carlo analysis

The DEq uncertainty estimates detailed above generally follow the conventional propagation rules, with the exception that degrees of freedom and known correlation issues are routinely ignored. Given the relatively small number of data available for estimating a V_{KC} or V_S, the assumption that $k_{95} = 2$ provides about a 95 % level of confidence coverage interval about the true value is difficult to justify. And, while the correlation between a given location estimate and a datum used in its estimation can be determined, the functional relationship can be fairly complex.

Parametric Bootstrap Monte Carlo (PBMC) analysis is one approach that provides a relatively simple and convenient method for estimating coverage intervals directly from just the reported data. Assuming that all of the reported $x \pm U_{95}(x)$ credibly specify $N(x,(U_{95}(x)/2)^2)$ normal kernel distributions, then empirical posterior distributions for all d values estimated from Eqs. 1 or 15 can be estimated by (1) repetitively sampling all of the input values within their distributions, providing one PBMC sample per reported result for each set, (2) estimating V_{KC} and V_S for each of the PBMC sets, and (3) estimating and storing the d (call them d_{MC}) for all of the resampled results in each set. This methodology is closely related to the methods described in [32] and to empirical Bayesian analysis [33].

While not particularly efficient in terms of computer resources, PBMC can be readily implemented in any computational environment that supports user definition of programs for the evaluation of specialized functions (e.g., x_{DL}) and for the storage of intermediate results. Since spreadsheets can provide a familiar working environment that simplifies the definition and maintenance of the linkages between root and successor KCs, PBMC analysis within a spreadsheet environment can be quite efficient in terms of analysts' resources when appropriate care is taken in their design.

Assuming that a suitably large number of PBMC samplings, N_{MC}, are available, d can be estimated from the empirical 50 percentile of the stored PBMC results:

$$d = \text{PTILE}(50, d_{MC}) \equiv \text{MEDIAN}(d_{MC}) \tag{16}$$

where "PTILE" is the function "return the p percentile of the specified values" and for $p = 50$ is identical to the median. Credible uncertainty intervals about d can be estimated in the same manner, with the 95 % level of confidence interval estimated from the 2.5 and 97.5 percentiles: PTILE(2.5,d_{MC}) and PTILE(97.5,d_{MC}). If the ratio $(d - \text{PTILE}(2.5,d_{MC}))/(\text{PTILE}(97.5,d_{MC}) - d)$ is about 1, then the usual symmetric 95 % confidence interval on d can be estimated as

$$U_{95}(d) = (\text{PTILE}(97.5, d_{MC}) - \text{PTILE}(2.5, d_{MC}))/2. \tag{17}$$

However, if the ratio is far from 1 then the interval can either be reported as asymmetric,

$$\begin{aligned} {}^{-}U_{95}(d) &= d - \text{PTILE}(2.5, d_{MC}), \\ {}^{+}U_{95}(d) &= \text{PTILE}(97.5, d_{MC}) - d, \end{aligned} \tag{18}$$

or as the larger of the two half-intervals,

$$U_{95}(d) = \mathrm{MAX}({}^{-}U_{95}(d), {}^{+}U_{95}(d)) \quad (19)$$

Asymmetric intervals are the narrowest intervals that provide the stated coverage; however, the familiar symmetric form may be more convenient for use in further calculations. While the symmetric estimates of Eq. 19 are conservative, they increasingly over-estimate the length of the interval as asymmetry increases.

Using the same sets of PBMC d_{MC} used to estimate the d in Eq. 16, the "measurement capability" DEq that combines results for all temperatures in a given buffer, the D for a given NMI of Eq. 5 can be estimated as

$$D = \mathrm{PTILE}\left(50, \bigcup_{t=1}^{T} d_{\mathrm{MC}t}\right) \quad (20)$$

where t indexes over the temperatures, T is the number of temperatures, $d_{\mathrm{MC}t}$ are the PBMC values for the given temperature, $\bigcup d_{\mathrm{MC}t}$ is the union of all the PBMC d_{MC} for a given NMI, and the number of d_{MC} is the same for all temperatures. The $U_{95}(D)$ can be estimated using the same approaches and decision criteria detailed in Eqs. 17–19:

$$U_{95}(D) = \left(\mathrm{PTILE}\left(97.5, \bigcup_{t=1}^{T} d_{\mathrm{MC}t}\right) - \mathrm{PTILE}\left(2.5, \bigcup_{t=1}^{T} d_{\mathrm{MC}t}\right)\right)/2$$

or

$${}^{-}U_{95}(D) = D - \mathrm{PTILE}\left(2.5, \bigcup_{t=1}^{T} d_{\mathrm{MC}t}\right);$$

$${}^{+}U_{95}(D) = \mathrm{PTILE}\left(97.5, \bigcup_{t=1}^{T} d_{\mathrm{MC}t}\right) - D$$

or

$$U_{95}(D) = \mathrm{MAX}({}^{-}U_{95}(D), {}^{+}U_{95}(D)) \quad (21)$$

Figure 2 displays the PBMC-estimated $d \pm U_{95}(d)$ and $D \pm U_{95}(D)$ for all NMIs that provided results for primary pH measurements in phosphate buffer. All of the expanded uncertainties are estimated conservatively as the maximum of the two half-intervals. At graphical resolution, the differences between the national standard estimates of Fig. 1 and the measurement capability estimates of Fig. 2 are quite small.

Figure 3 provides a high-resolution comparison between the DEq as reported in the Final Reports and those estimated using PBMC and the several estimation and linkage modifications proposed above. All of the pH differences are small with none larger than 0.003 and most less than 0.001, but the pattern of changes attributable to specific modifications may be of interest. Figure 3a visualizes the differences in d, $U_{95}(d)$, D, and $U_{95}(D)$ attributable to the PBMC estimation method itself. The d are essentially unaffected; the D are mostly unaffected except for those NMIs where the distribution of the combined d_{MC} is not

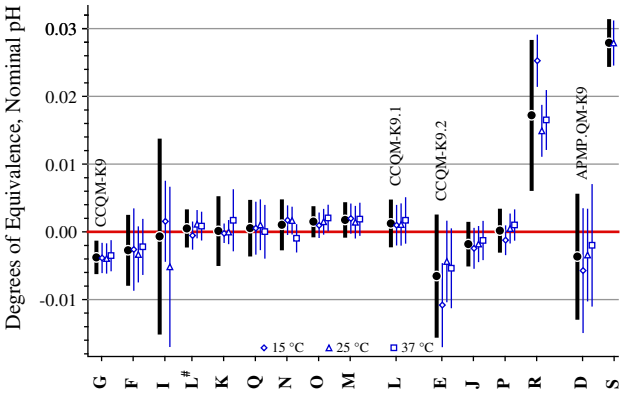

Fig. 2 Dot-and-bar plot of PBMC estimated degrees of equivalence for the CCQM-K9, -K9.1, -K9.2, and APMP.QM-K9 participants. The graphical format is identical to that of Fig. 1

well described as symmetric unimodal. For these NMIs, the PBMC-estimated median d_{MC} is somewhat closer to the ideal zero D than the arithmetic average. The PBMC-estimated $U_{95}(d)$ for the CCQM-K9 participants are somewhat smaller than the reported values. The PBMC-estimated $U_{95}(d)$ for the participants in successor studies are either unchanged or somewhat larger, depending on which KC is considered. The $U_{95}(D)$ are essentially unaffected, again except for the NMIs where the combined d_{MC} distribution is significantly asymmetric.

Figure 3b depicts the changes attributable to the use of x_{DL} for the reference values in CCQM-K9, -K9.2, and APMP.QM-K9. None of the d and D are changed by more than about ± 0.0005. The $U_{95}(d)$ and $U_{95}(D)$ are on average very slightly smaller than the values provided in the reports or estimated from them. Figure 3c depicts the change resulting from linking CCQM-K9.2 to the KCRV using the corrected value reported in CCQM-K9 by one of the anchor NMIs rather than that NMI's official DEq estimated in CCQM-K9.1. The change only affects the APMP.QM-K9 participants. Figure 3d depicts the change resulting from using LOO evaluation for the CCQM-K9 participants, where the d and D become on average about 0.0002 farther from zero and the $U_{95}(d)$ and $U_{95}(D)$ become uniformly about 0.0003 larger. These small changes have virtually no effect on the DEq estimated for the participants in the successor KCs.

Figure 3e depicts the change in linkage from the "national standard" paradigm of Eq. 12 to the "measurement capability" paradigm of Eq. 15. The d and D for the participants in the successor KCs are changed by up to ± 0.002, reflecting the elimination of the V_{R} bias-correction resulting in a small majority of the DEq becoming closer to the ideal zero. The $U_{95}(d)$ and $U_{95}(D)$ for these NMIs rather uniformly become about 0.0005 shorter, reflecting the elimination of the $u(V_{\mathrm{R}})$ uncertainty component.

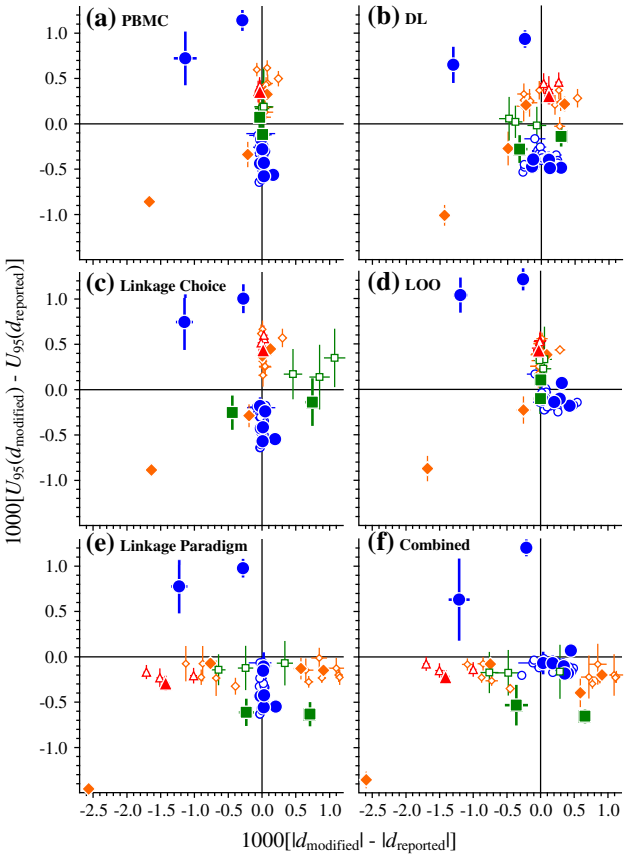

Fig. 3 Differences between the degrees of equivalence and their expanded uncertainties for the CCQM-K9, -K9.1, -K9.2, or APMP.QM-K9 participants as reported and as estimated using the proposed modified approaches. The *panels* display differences due to the use of **a** the PBMC estimation process, **b** DerSimonian–Laird weighted mean to estimate all reference values, **c** linking to a corrected value in CCQM-K9 rather than to its replacement in CCQM-K9.1, **d** leave-one-out evaluation of CCQM-K9 results, **e** measurement capability paradigm linkage, and **f** the combination of all the proposed modifications. For all *panels*, the *horizontal axis* displays differences in absolute d and D; the *vertical axis* displays differences in $U_{95}(d)$ and $U_{95}(D)$. Negative values along either axis indicate that the reported values are further from the ideal zero than those estimated using the proposed modification. *Small open symbols* represent temperature-specific differences in d and $U_{95}(d)$; *large solid symbols* differences in the estimated D and $U_{95}(D)$; *circles* estimates from CCQM-K9, *triangles* CCQM-K9.1, *diamonds* CCQM-K9.2, and *squares* APMP.QM-K9. The *bars* on all *symbols* represent 95 % level of confidence intervals on the PBMC estimates, based on 9 sets of 1000 random draws

Figure 3f depicts the whole of the proposed modifications. The great majority of the observed changes are attributable to use of (1) the measurement capability paradigm, (2) PBMC analysis, (3) x_{DL} as the estimator for the reference values in both the root and successor KCs, and (4) LOO analysis of the DEq for participants in the root KC. Note that each of these modifications can have very different effects on the participants in the root and in the successor KCs, and the magnitude of the changes observed with the CCQM-K9, -K9.1, K-9.2, and APMP.QM-K9 studies may not predict their relative impact on other measurement systems.

The PBMC results for all five buffer systems are listed in Tables S1.c to S5.c of the ESM.

"Measurement capability" degrees of equivalence for all buffers

While each buffer system has its unique attributes, the $d \pm U_{95}(d)$ estimates for most NMIs in other buffers where measurements were reported at 15 °C, 25 °C, and 37 °C are about as self-consistent as they are in the phosphate buffer discussed above. Given that all $D \pm U_{95}(D)$ within-buffer estimates appear to "make chemical sense", it remains to explore how results can be combined across the buffers—and whether such combinations are chemically informative.

To meaningfully combine across the buffer systems, the magnitude and distributions of the quantities combined must be similar. Figure 4 displays the standard deviation, $s(x)$, the DerSimonian-Laird between-NMI component of variation, s_b, and the pooled (see Eq. 7) measurement uncertainties, $\bar{u}(x)$, estimated from the accepted results in the five root KCs. The $\bar{u}(x)$ are strikingly similar for all five buffers, indicating that the participating NMIs regarded the measurement processes as being of similar complexity. However, the reported measurement uncertainties do not fully account for the observed between-NMI variability in any of the buffer systems. The magnitude of the unexplained between-NMI variability is about the same and rather small in four of the buffers. Only in the carbonate system investigated in CCQM-K18 and -K18.1 the unexplained variability is significant—and can be entirely attributed to a reproducible offset in the measurement results reported by two NMIs. While not yet completely understood, this offset is believed to be related to the procedures used to account for slow loss of CO_2 from the buffer into the hydrogen flow in the Harned cell.

The carbonate buffer KCs are also unique in that, owing to the time required for measurement at each temperature, the KC protocol only involved measurements at 25 °C. It is plausible that primary pH measurements in this system may not be comparable to those in the other four buffers. However, the variability of the DEq in the carbonate system is not so much greater than that in the others to preclude attempting to combine them with those for the other buffers and evaluating the resulting combined values for chemical plausibility.

The number of temperatures evaluated in the EAWG's pH KCs does differ; further, KC participants do not always report results for all of the temperatures included in the KC

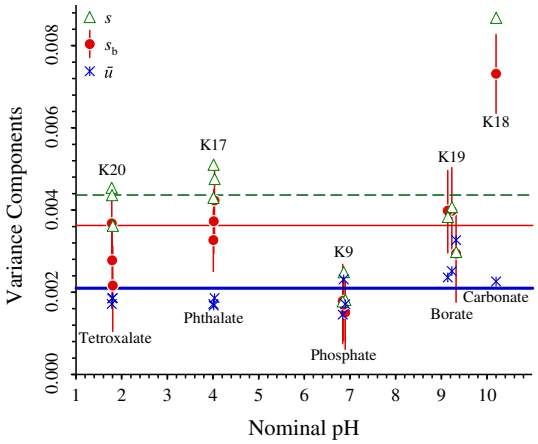

Fig. 4 Uncertainty components for the pH-related measurement results reported in the CCQM-K9, -K17, K-18, -K19, and -K20 key comparisons. The *horizontal axis* displays the KCRVs as estimated from the 15, 25, and 37 °C results accepted for use in estimating the KCRV. The *vertical axis* displays estimates of variability for these results. The *open triangles* represent the standard deviations, *s*, for the reported *x* in each of these root KCs; the *dashed horizontal line* the pooled value of the *s*. The *asterisk* represents the pooled uncertainty, $\bar{u}$, of the reported $u(x)$; the *thick horizontal line* their pooled value. The *solid circles* represent the DerSimonian–Laird estimate of between-NMI variability, s_b; the *thin horizontal line* their pooled value. The *horizontal* and *vertical lines* represent PBMC-estimated 95 % coverage intervals, based on 9 sets of 1000 random draws

design. To provide an "all buffer" DEq summary, $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$, for each NMI, this potential imbalance in the number of temperature-specific $d \pm U_{95}(d)$ available in different buffers requires modification of the single-buffer approaches for combining DEq. This is trivial for the propagation approach, requiring only that the $d \pm U_{95}(d)$ in Eq. 5 be replaced by the summary $D \pm U_{95}(D)$:

$$\boldsymbol{D} = \sum_{i=1}^{N} D_i \Big/ N; \quad u(\boldsymbol{D}) = \sqrt{\bar{u}^2(D) + s^2(D)}; \quad U_{95}(\boldsymbol{D}) = 2u(\boldsymbol{D})$$

$$\bar{u}^2(D) = \sum_{j=1}^{N} \left(\frac{U_{95}(D_i)}{2} \right)^2 \Big/ N; \quad s^2(D) = \begin{cases} 0 & N = 1 \\ \sum_{i=1}^{N} (D_i - \boldsymbol{D})^2 \Big/ (N-1) & N > 1 \end{cases} \tag{22}$$

where *i* now indexes over the buffers and *N* is number of buffer systems for which the NMI provided results.

Estimating $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$ is only a bit more complicated for the PBMC approach of Eq. 16. Using the same sets of PBMC d_{MC} used to estimate d, $U_{95}(d)$, D and $U_{95}(D)$:

$$\boldsymbol{D} = \text{PTILE}\left(50, \bigcup_{t=1}^{T} d_{MCt}\right) \tag{23}$$

where *t* now indexes over all temperatures in all buffers.. The $U_{95}(\boldsymbol{D})$ can be estimated using the analogous modifications to Eq. 21, again using the decision criteria discussed for Eqs. 17–19.

To ensure that each of the five buffer systems has equal influence on the all-buffer $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$ estimates, the total number of d_{MC} should be the same for all buffers, e.g., for each 1000 PBMC d_{MC} values generated for each of the three results reported in the phosphate buffer system there should be 3000 d_{MC} for the carbonate buffer's single result. While just a bookkeeping detail, having balanced numbers of d_{MC} is necessary for the PBMC process to yield equal-weighted estimates.

Figure 5 displays the variance propagation and PBMC-generated $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$ estimates for all NMIs reporting any primary pH result in any of the pH KCs listed in Table 1, with the $U_{95}(D)$ and $U_{95}(\boldsymbol{D})$ conservatively estimated as the maximum half-interval. Figure 5 uses the same dot-and-bar format used in Fig. 1, but with the thin lines representing the buffer-specific $D \pm U_{95}(D)$ rather than the within-buffer temperature-specific $d \pm U_{95}(d)$. At graphical resolution, the two methods provide very similar estimates; numeric values of the estimates are listed in Table S6 of the ESM. Figure S6 displays the PBMC results using symmetric and asymmetric $U_{95}(D)$ and $U_{95}(\boldsymbol{D})$ intervals.

The $D \pm U_{95}(D)$ for the carbonate buffer do not appear to be systematically different from those of the other buffer systems. For the large majority of NMIs, the DEq in different buffers are quite coherent. The reproducible and relatively large offset for the NMI coded as "T" has been previously noted and identified as the result of using a somewhat different electrochemical cell design than that used by most other NMIs.

Conclusion

The very similar values of the temperature-specific $d \pm U_{95}(d)$ for the primary pH measurement results reported by most KC participants in each of the five buffer systems suggest that combining them into buffer-specific $D \pm U_{95}(D)$ summaries provides chemically useful information—at least for the measurements made over the range of temperatures evaluated in that buffer. Likewise, the very similar values for the buffer-specific $D \pm U_{95}(D)$ for most NMIs suggest that combining them into the buffer-independent $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$ summaries may usefully summarize the primary pH measurement capabilities of the KC participants—at least for the five buffer systems and 15 °C–37 °C temperature range considered in this study.

While not essential to reaching the above conclusions, we propose a number of modifications to the methods

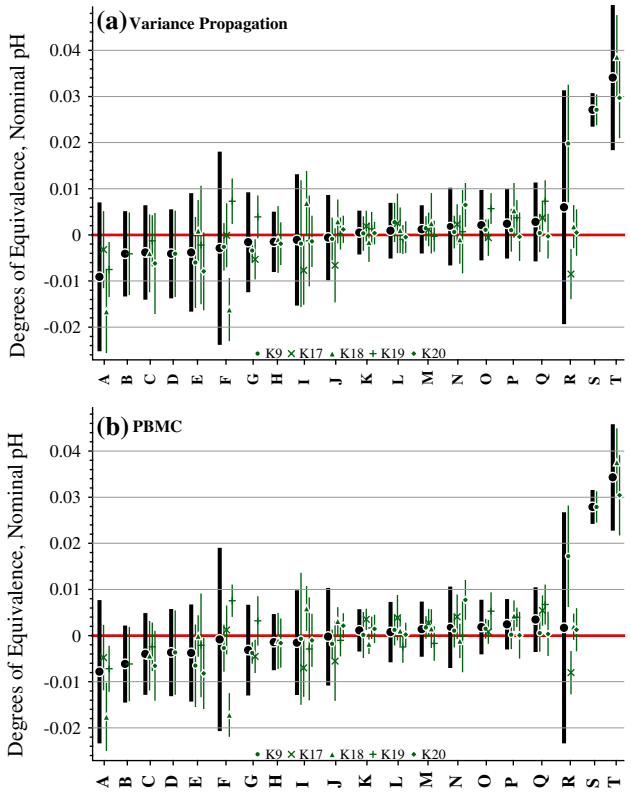

Fig. 5 Dot-and-bar plots of degrees of equivalence for all NMIs that reported primary method pH results in any of the 11 KCs listed in Table 1. **a** Variance propagation estimates, **b** PBMC estimates. The graphical format is similar to that of Fig. 1 with the exception that the *large solid circles* and *thick bars* represent the all-temperature all-buffer $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$ summaries, and the *smaller symbols* and *thin lines* represent the all-temperature $D \pm U_{95}(D)$ buffer-specific summaries. The *smaller solid circles* represent results for phosphate buffer (CCQM-K9, -K9.1, -K9.2, APMP.QM-K9), *times* phthalate buffer (CCQM-K17, EUROMET-K17), *solid triangles* carbonate buffer (CCQM-K18, -K18.1), *plus* borate buffer (CCQM-K19, -K19.1), and *solid diamonds* tetroxalate buffer (CCQM-K20)

usually used for CIPM MRA degrees of equivalence that may contribute to providing more representative estimates. The most significant of these are use of (1) "measurement capability" linkages between root and successor KCs, (2) Monte Carlo (PBMC and others) methods for evaluating the consequences of different distributional assumptions on the estimation of credible coverage intervals, (3) comparison of leave-one-out (LOO) degrees of equivalence estimates with those using the traditional approach to evaluate the influence of correlation, and (4) a modified dot-and-bar graphic for displaying summary estimates such as $D \pm U_{95}(D)$ and $\boldsymbol{D} \pm U_{95}(\boldsymbol{D})$.

The primary pH measurement results provided by the NMI participants in these pH-related KCs were chosen for study for a number of reasons, but chief among them is the remarkable agreement among the participant results over all of the solutions and evaluation temperatures thus far studied by the EAWG. If the degrees of equivalence for these measurements could not have been meaningfully combined, it would be highly unlikely that the results for less well understood and controlled measurement systems could be meaningfully combined. That the primary pH results can be combined using relatively simple analysis and display methods thus does not ensure that similarly meaningful summaries can be devised for other measurement systems, but it provides the incentive to attempt to do so.

Acknowledgments We thank all participants in CCQM-K9, -K9.1, K9.2, -K17, -K18, -K18.1, -K19, K19.1, -K20, APMP.QM–K9, and EUROMET.QM-K17 for their thoughtful contributions to the design of the studies and evaluation of the results and for their meticulous measurements. DLD thanks Katherine Sharpless and Katrice Lippa for their assistance and advice in preparing this report and the anonymous reviewers for their careful corrections and insightful suggestions.

References

1. CIPM (2003) Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes. Comité international des poids et mesures. Paris, 14 October 1999 with Technical Supplement revisions of October 2003. http://www.bipm.org/utils/en/pdf/mra_2003.pdf
2. Viallon J. Protocol for the key comparison BIPM.QM-K1, Ozone at ambient level. Bureau International des Poids et Mesures, Pavillon de Breteuil, F-92312 Sèvres Cedex France 2007. http://www.bipm.org/utils/en/pdf/BIPM.QM-K1_protocol.pdf
3. Spitzer P (2001) Final report for CCQM-K9: pH determination on two phosphate buffers by Harned cell measurements. http://kcdb.bipm.org/AppendixB/appbresults/ccqm-k9/ccqm-k9_final_report.pdf
4. Spitzer P (2001) pH determination on two phosphate buffers by Harned cell measurements. Follow-up bilateral comparison SMU-PTB. http://kcdb.bipm.org/appendixb/appbresults/ccqm-k9/ccqm-k9_s_smu_final_report.pdf
5. Spitzer P et al (2008) Final report for CCQM-K9.2: subsequent key comparison on pH determination of phosphate buffer by Harned cell measurements. Metrologia 45(Tech Supp): 08006. doi:10.1088/0026-1394/45/1A/08006
6. Hioki A et al (2011) Final report of the key comparison APMP.QM-K9: APMP comparison on pH measurement of phosphate buffer. Metrologia 48(Tech Supp): 08012. doi:10.1088/0026-1394/48/1A/08012
7. Spitzer P et al (2003) Final report for CCQM-K17: pH determination on a phthalate buffer by Harned cell measurements. Metrologia 40(Tech Suppl): 08006. doi:10.1088/0026-1394/40/1A/08006
8. Spitzer P et al (2005) Final report of EUROMET Project 696: pH determination of a phthalate buffer. Metrologia 42(Tech Suppl): 08001. doi:10.1088/0026-1394/42/1A/08001

9. Vyskočil L et al (2007) Report of key comparison CCQM-K18: pH of carbonate buffer. Metrologia 44(Tech Suppl): 08011. doi:10.1088/0026-1394/44/1A/08011
10. Vyskočil L et al (2008) Final report of Subsequent key comparison CCQM-K18.1: pH of carbonate buffer. Metrologia 45(Tech Suppl): 08014. doi:10.1088/0026-1394/45/1A/08014
11. Spitzer P (2006) Key comparison CCQM-K19 on pH: final report. Metrologia 43(Tech Suppl): 08015. doi:10.1088/0026-1394/43/1A/08015
12. Spitzer P et al (2011) Final report on CCQM-K19.1: pH of borate buffer. Metrologia 48(Tech Suppl): 08010. doi:10.1088/0-1394/48/1A026/08010
13. Pratt KW (2009) Final report on key comparison CCQM-K20: pH of tetroxalate buffer. Metrologia 46(Tech Suppl): 08022. doi:10.1088/0026-1394/46/1A/08022
14. Spitzer P et al (2013) Final report on CCQM-K91: Key comparison on pH of an unknown phthalate buffer. Metrologia 50(Tech Suppl): 08016. doi:10.1088/0026-1394/50/1A/08016
15. Buck RP, Rondinini S, Covington AK, Baucke FGK, Brett CMA, Camões MF, Milton MJT, Mussini T, Naumann R, Pratt KW, Spitzer P, Wilson GS (2002) Measurement of pH. Defintion, standards and procedures (IUPAC Recommendations 2002). Pure Appl Chem 74(11):2169–2200. http://pac.iupac.org/publications/pac/pdf/2002/pdf/7411x2169.pdf
16. Máriássy M, Pratt KW, Spitzer P (2006) Major applications of electrochemical techniques at national metrology institutes. Metrologia 46:199–213. doi:10.1088/0026-1394/46/3/007
17. Spitzer P, Pratt KW (2011) The history and development of a rigorous metrological basis for pH measurements. J Solid State Electrochem 15:69–76
18. Bates RG, Guggenheim EA (1960) Report on the standardization of pH and related terminology. Pure Appl Chem 1(1):163–168. http://pac.iupac.org/publications/pac/pdf/2002/pdf/7411x2169.pdf
19. De Bièvre P, Dybkaer R, Fajgelj A, Hibbert DB (2011) Metrological traceability of measurement results in chemistry: concepts and implementation (IUPAC Technical Report). Pure Appl Chem 83(10):1873–1935. http://iupac.org/publications/pac/83/10/1873/pdf/
20. JCGM 100:2008. Evaluation of measurement data—Guide to the expression of uncertainty in measurement. BIPM, Sèvres, France. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf
21. Graybill FA, Deal RB (1959) Combining unbiased estimators. Biometrics 15:543–50. http://www.jstor.org/stable/2527652
22. Cox MG (2002) The evaluation of key comparison data. Metrologia 39:589–95. doi:10.1088/0026-1394/39/6/10
23. CCQM (2013) CCQM Guidance note: estimation of a consensus KCRV and associated degrees of equivalence, Version 10. http://www.bipm.org/cc/CCQM/Allowed/19/CCQM13-22_Consensus_KCRV_v10.pdf
24. DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Controlled Clinical Trials 7:177–88. doi:10.1016/0197-2456(86)90046-2
25. Rukhin AL (2009) Weighted means statistics in interlaboratory studies. Metrologia 46:323–31. doi:10.1088/0026-1394/46/3/021
26. Duewer DL (2008) A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers. Accred Qual Assur 13:193–216. doi:10.1007/s00769-008-0360-3
27. Birge RT (1932) The calculation of errors by the method of least squares. Phys Rev 40:207–227. doi:10.1103/PhysRev.40.207
28. Horn SA, Horn RA, Duncan DB (1975) Estimating heteroscedastic variances in linear models. J Am Stat Assoc 70(350):380–385. http://www.jstor.org/stable/2285827
29. Rukhin AL (2011) Maximum likelihood and restricted likelihood solutions in multiple-method studies. J Res Natl Inst Stand Technol 116(1):539–556. doi:10.6028/jres.116.004
30. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, Berlin. http://statweb.stanford.edu/~tibs/ElemStatLearn/
31. Duewer DL, Gasca-Aragon H, Lippa KA, Toman B (2012) Experimental design and data evaluation considerations for comparisons of reference materials. Accred Qual Assur 17:567–588. doi:10.1007/s00769-012-0920-4
32. JCGM 101:2008. Evaluation of measurement data—Supplement 1 to the "Guide to the expression of uncertainty in measurement"—Propagation of distributions using a Monte Carlo method. BIPM, Sèvres, France. http://www.bipm.org/utils/common/documents/jcgm/JCGM_101_2008_E.pdf
33. Elster C, Toman B (2009) Bayesian uncertainty analysis under prior ignorance of the measurand versus analysis using the supplement 1 to the Guide: a comparison. Metrologia 46(3):261–266. doi:10.1088/0026-1394/46/3/013