Experimental Design and Data Evaluation Considerations for Comparisons of Reference Materials

<Version 27-July-2012>

Published: Accred Qual Assur (2012) 17:567-588
http://DX.DOI.ORG/10.1007/s00769-012-0920-4

David L. Duewer[1*], Hugo Gasca-Aragon[1,2], Katrice A. Lippa[1], Blaza Toman[3]

1 Analytical Chemistry Division
National Institute of Standards and Technology (NIST)
Gaithersburg, MD USA 20899-8390

2 Department of Mathematics & Statistics
University of Massachusetts
Amherst, MA 01003-9305

3 Statistical Engineering Division
National Institute of Standards and Technology (NIST)
Gaithersburg, MD USA 20899-8980

* Corresponding Author: 100 Bureau Drive, MS 8390
301-975-3935 (Phone)
301-926-8671 (Fax)
david.duewer@nist.gov

Disclaimer

Certain commercial software is identified in this report to specify the experimental procedure as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the software is necessarily the best available for the purpose.

Abstract

The analysis of reference materials (RMs) can help assess the equivalence of chemical measurement processes. When two or more RMs are available for a given measurand, confidently establishing the equivalence of measurement processes requires the RMs to be capable of yielding equivalent results. Evaluating the degrees of equivalence among RMs that differ in analyte quantity and perhaps matrix composition requires an approach other than that used to assess results for samples of a single material. We have more than a decade of experience with an approach compares the assigned values of RMs to a simple linear model of the relationship between those values and measurement results ideally made under repeatability conditions. In addition to accessing the metrological equivalence of specific RMs, the equivalence of the value-assignment capabilities of the organizations that issue the RMs can also be accessed. This report summarizes our experience with the design of and analysis of studies using this approach and provides numeric and graphical tools for estimating degrees of equivalence. We divide the required tasks into four steps: 1) design, 2) measurement, 3) definition of a reference function, and 4) estimation of degrees of equivalence. We regard the experimental design and measurement tasks as most critical to the eventual utility of the comparison, since creative mathematics can not fully compensate for poor planning or erratic measurements.

Keywords

Consultative Committee for Amount of Substance – Metrology in Chemistry; degrees of equivalence; generalized distance regression; metrological comparability, metrological equivalence, reference function; reference material

Introduction

A reference material (RM) is formally defined as a "material, sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process [1,2]. The analysis of RMs that deliver specified values of a property can help assess the equivalence of measurement processes. When only one such RM exists for a given chemical measurand (a given analyte in a given matrix with measurement results expressed in specified units) then the results of all measurement processes for that measurand can be tied together through that one common reference. However, when two or more RMs are available then confidently establishing the equivalence of measurement processes requires that those materials have been shown capable of yielding equivalent results. This is particularly true for materials that have been value-assigned by different organizations or by the same organization but at different times, by different analysts, or using different analytical approaches.

Because different RMs for a given measurand may deliver different levels of the analyte in somewhat different matrices, the evaluation of degrees of equivalence among these materials requires approaches that differ from those used to assess results for subsamples of a single material. In support of the Joint Committee for Traceability in Laboratory Medicine's efforts to enhance the reliability of clinical laboratory measurements, [3] development of a general protocol for demonstrating the extent of equivalence among RMs that deliver the same measurand was initiated in 2001 [4]. Practical techniques for the conduct and evaluation of such comparisons were established in a series of studies conducted from 2003 through 2008 [5]. The essence of these techniques is the comparison of the assigned values to measurement results ideally made under repeatability conditions through a mathematical model developed using a method that respects the uncertainties in both the assigned and measured values.

Building on this experience, in 2009 the Organic Analysis Working Group (OAWG) of the Comité International des Poids et Mesures (CIPM) Consultative Committee for Amount of Substance - Metrology in Chemistry (CCQM) decided to directly compare the measurement services that its member organizations, mostly National Metrology Institutes (NMIs), deliver to their customers. The premise of these studies is that demonstrating the equivalence of the value assignment of particular materials also demonstrates the equivalence of the organizations' RM-related processes. In addition to calibration and the determination of the analyte concentration in a complex matrix, these processes include the assessment of material homogeneity and stability and may include packaging, storage, and shipping capabilities.

Measurements for two such comparisons have been completed and the data analysis methods developed to evaluate the degrees of equivalence for both the materials and their organizations' measurement services have been accepted in principle by the OAWG: CCQM-K79 "Ethanol in aqueous matrix" and CCQM-K80 "Creatinine in human serum." Final reports for both studies will become publically available after their formal acceptance [6]. We here describe the experimental design considerations and data evaluation tools developed from our experience with these studies and our review of results from several comparisons conducted by the CIPM CCQM Gas Analysis Working Group (GAWG) on purpose-certified gas mixtures.

We divide the tasks required by a RM evaluation into four steps: 1) design, 2) measurement, 3) definition of a reference model, and 4) estimation of degrees of equivalence. We regard the design (Step 1) and measurement (Step 2) tasks as most critical, since creative mathematics can not fully compensate for poor planning or erratic measurements. Each of these steps will be discussed in its own section after first presenting the mathematical model that motivates the design, measurement, and evaluation considerations.

Terminology

Formally, measurements made under "repeatability conditions" are made by one analyst using the same set of supplies and equipment over a relatively short period of time. Ideally, the measurements to which the assigned values are to be compared can all be made in a single measurement campaign, where "campaign" refers to a single "run" or "batch" of measurements accomplished in the requisite "relatively short period of time." Practically, however, measurements may have to be performed by a team of analysts in two or more campaigns and the campaigns may have to be separated in time by several days to weeks. When the measurements are made in more than one campaign but by the same team using the same equipment as quickly as resources allow, then the measurement conditions are more accurately described as "intermediate within-laboratory single-team same-equipment short-as-possible-term". However, for simplicity's sake, we refer to measurements made under these conditions to be "repeatability measurements".

We have previously used the term "comparability" when discussing the extent of agreement among different reference materials that deliver the same measurand [4,5]. However, "comparability" now has a formal metrological definition that is not compatible with this usage [1]. The metrological comparability of measurement results refers to their "metrological traceability to the same reference" - in approximate translation, the qualitative property of being expressed in the same units. For example, a creatinine result expressed in mg/g is not comparable with one expressed in mg/dL - although it can be made comparable if the density of the material has been determined.

The likewise formally defined term "metrological compatibility" of measurement results addresses the quantitative agreement of measurement results relative to their uncertainties [1]. While quite compatible with our prior use of comparability, as currently defined the term is specifically limited to the comparison of the results of different measurement processes for a specified measurand in a *single* material.

To our knowledge, there is as yet no internationally recognized term to describe the quantitative comparison of results from *different* materials. However, the CIPM Mutual Recognition Arrangement (CIPM MRA) [7] that provides the technical basis for comparing national metrology services defines the term "degree of equivalence" as expressing the compatibility of a given result and a reference value. Further, it uses the term "metrological equivalence" in the context of measurement capabilities rather than just particular results.

Given the authority that the CIPM MRA has among NMIs, we chose to use the term "equivalence" in the context of comparing materials and certification processes. Metrologically equivalent RMs are with high confidence expected to deliver their analyte quantities within their stated uncertainty intervals and metrologically equivalent certification processes are with high confidence expected to be capable of providing equivalent RMs. However, "metrological equivalence" does *not* imply "exactly the same." Given that there is seldom just one practical approach to a chemical characterization task, metrologically equivalent certification processes may use quite different analytical techniques. Further, RMs that are metrologically equivalent when evaluated with a given measurement process may differ in many ways that impact their practical utility, including but not limited to: analyte level, stated uncertainty, stability, quantity per unit, commutability among measurement procedures, and - not least - cost.

Model

The main data analysis challenge in comparing RMs is establishing a descriptive relationship between two sets of values, call them *V* (for assigned *Value*) and *R* (for measurement process *Response*), both characterizing the same group of similar but intrinsically somewhat different materials. For straight-line relationships, the general model is

$$R = \alpha + \beta V + E \tag{1}$$

where α is the intercept, β the slope, and *E* the residual random error. While relationships other than linear can in principle be considered, those more complex than this two-parameter straight line will in general be much less practical.

So-called "ordinary least squares (OLS)" techniques are appropriate for estimating the α and β parameters when the *V* are known exactly. When the *R* are known within some estimate of uncertainty, call it $u(R)$, then α and β can be estimated using techniques that minimize the uncertainty-weighted residual sum of squares *E*:

$$E = \sum_i^{n_m} \varepsilon_i^2; \quad \varepsilon_i^2 = \left(\frac{R_i - \hat{R}_i}{u(R_i)} \right)^2; \quad \hat{R}_i = \hat{\alpha} + \hat{\beta} V_i \tag{2}$$

where the *i* subscript indicates the values for a particular material, n_m is the number of materials studied, and the "^" (read as "hat") above a symbol indicates an estimated value. The simple OLS procedures available in spreadsheet software are adequate when the $u(R_i)$ are about the same for all materials. Weighted OLS procedures, widely available in most data evaluation software, are required when the $u(R_i)$ differ significantly among the materials.

OLS techniques are not appropriate when neither the *V* nor *R* is known exactly. Problems involving two or more sets of inexactly known values were addressed at least as early as 1879 [8] and the general problem was "first clearly stated" [9] in 1901 [10]. The challenge of evaluating bias between two chemical measurement systems was addressed in 1943 [11], although the proposed solution required the ratio between the measurement uncertainties, $u(R_i) / u(V_i)$, to be the same for all materials. Empirical non-linear optimization software that addressed bias evaluation for non-constant $u(R_i) / u(V_i)$ became available at least as early as 1984 [12]. A mathematically rigorous solution was published in 1987 [13] and was independently developed in 2001 as part of an international documentary standard for use with calibration gas mixtures [14]. This approach provides estimates of α and β that minimize the uncertainty-normalized differences between the observed values and their projection onto the line found by minimizing *E*:

$$E = \sum_i^{n_m} \varepsilon_i^2; \quad \varepsilon_i^2 = \left(\frac{R_i - \hat{R}_i}{u(R_i)} \right)^2 + \left(\frac{V_i - \hat{V}_i}{u(V_i)} \right)^2; \quad \hat{R}_i = \hat{\alpha} + \hat{\beta} \hat{V}_i \tag{3}$$

where $\hat{\alpha}$, $\hat{\beta}$, and the $\hat{R}_i$ and $\hat{V}_i$ are all estimated simultaneously.

Computation

The small increase in notational complexity of Equation 3 relative to Equation 2 disguises a very large increase in computational difficulty. Techniques continue to be developed to more efficiently use all of the available information relevant to the comparison and/or for non-linear relationships between the two sets of results [15-17]. However, several specialized software systems designed for two-parameter linear

relationships are currently available, including: the stand-alone commercial B_LEAST [14], the web-available Excel [18]-based freeware systems FREML [19] and XLGENLINE [20], and the Excel-based RegViz exploratory tool described in [5]. While differing in calculation methods and input/output details, these systems yield very similar parameter estimates when given the same data. We use the RegViz system for its graphical capabilities and our ability to quickly add functionality to it. We have confirmed the RegViz parameter estimates with the well-validated and intuitive FREML system developed by Ellison and the somewhat more general XLGENLINE developed at National Physical Laboratory (NPL) of the UK. The Supplementary Information presents exemplar analysis using each of these systems

Application of Equation 3 requires that all of the input data be stated as {value, uncertainty} pairs and that normalizing each residual (value – predicted value) to the uncertainty of the value efficiently gives all of the residuals the same scale. This not only assumes that each uncertainty is well-estimated, it assumes that all of the uncertainties are of the same type; i.e., that all of the {value, uncertainty} pairs have the same distributional form. The software systems mentioned above tacitly assume that these pairs define $N(\mu_i, \sigma_i^2)$ Gaussian (i.e., normal) kernel distributions, with each "value" estimating the kernel center, μ_i, and the squared "uncertainty" estimating its variance, σ_i^2. For these assumptions to be valid, the number of degrees of freedom, *v*, associated with each uncertainty estimate must be "large."

The very similar regression methodologies used by these and other data analysis systems have been (confusingly) given many different names [14,15,19,21]. We chose to use the term "generalized distance regression" (GDR) [22].

History of Use Within the CCQM

The GAWG's 2003 CCQM-P41 "Greenhouse gases. 2. Direct comparison of primary standard gas mixtures" [23] was the CCQM's first published study that compared assigned values with measurements made under repeatability conditions. Primary standard gas mixtures (PSMs) are primary standards of gas concentration prepared by NMIs and others for use as their untimate reference for gas measurements [24]. CCQM-P41 used a design that the GAWG pioneered in 2001 for comparing the value-assignments of single cylinders of purpose-prepared primary standard gas mixtures (PSMs) as made by each of the study participants using measurements made by a single laboratory over a short period of time. CCQM-P41 evaluated nine PSMs having targeted CO_2 amounts ranging from 350 μmol/mol to 380 μmol/mol and CH_4 amounts from 1.6 μmol/mol to 2 μmol/mol. Similar designs were used in three studies in 2006 and 2007: CCQM-P73 "Nitrogen monoxide gas standards (30-70) μmol/mol" [25], CCQM-K54 "Direct comparison of PSMs of hexane in methane" [26], and CCQM-K53 "Oxygen in nitrogen" [27]. CCQM-P73 evaluated two PSMs each from 12 NMIs or designated institutes (DIs), one with a targeted amount of NO in the range of 30 μmol/mol to 50 μmol/mol and the other in the range of 50 μmol/mol to 70 μmol/mol. CCQM-K54 evaluated eight PSMs with hexane amounts ranging from 120 μmol/mol to 200 μmol/mol. CCQM-K53 evaluated 12 PSMs with O_2 amounts from 99.0 μmol/mol to 101.2 μmol/mol.

The OAWG's 2009 CCQM-K80 creatinine comparison evaluated 17 materials from six organizations with creatinine amounts ranging from 3.1 mg/kg to 57 mg/kg. This was the first CCQM material comparison to use customer-deliverable materials for an analyte in a complex natural matrix, and to require that the procedure used to make the repeatability measurements have a linear response over more than a three-fold range in analyte level. The 2010 CCQM-K79 ethanol comparison evaluated 27 materials from nine organizations with ethanol amounts ranging from 0.1 g/kg to 334 g/kg. While evaluating a stable analyte in relatively simple matrices, this large comparison required the repeatability measurement procedure to provide a linear response over more than a thousand-fold range.

Step 1: Design the Study

RM comparisons involve two sets of data characterizing the same group of materials. One set is provided by the organizations that value-assign the RMs and thus reflects the capabilities of those organizations. We designate this set with the symbol *V*. The other set is provided by a single coordinating laboratory that may be, but is not necessarily, one of the participants in the study. This laboratory makes measurements on all of the materials under as close to repeatability conditions as practical, producing a set of measurement results that are intended to reflect the materials' relative measurand quantities. We designate this set with the symbol *R*. For notational simplicity, *V* and *R* are used for both the abstract data sets and for the observed values of those data sets.

Assigned Values (V)

A metrologically valid assigned value reported consists of an expected value (V_i) and its expanded uncertainty, $U(V_i)$. While the expected value can be assumed to reflect the organization's best efforts regardless of the nature of the material, the components included in the uncertainty estimate and how the uncertainty is expressed may differ depending on the nature of the comparison.

Purpose-assigned materials

For comparisons that involve only materials that are value-assigned especially for the study, the nature of the assigned uncertainties may be dictated by the study protocol. The GAWG's CCQM-P41, -P73, -K53, and -K54 study protocols focused on the gravimetry and purity verification uncertainty components that are related to the preparation of the PSMs.

The uncertainties for these GAWG comparisons were reported in expanded form at the 95 % level of confidence, $U_{95}(V_i)$. With one exception where the expansion factor was stated as 1.96, the factor used to expand the standard combined uncertainties for all of the PSMs was implicitly assumed or explicitly stated to be 2. This choice of expansion factor indicates that the *v* for each the uncertainty estimate is "large." For these studies,

$$u_\infty(V_i) = U_{95}(V_i)/2 \tag{4}$$

thus provides "large sample" uncertainty estimate for the kernels. To the extent that the V_i and $U_{95}(V_i)$ are correctly estimated, the $\mathrm{N}\left(V_i, u_\infty^2(V_i)\right)$ kernels describe the same information for all of the PSMs.

RMs

While different organizations may not use the same analytical techniques and statistical models to value-assign their RMs, the assigned uncertainties can be assumed to include "all" of the relevant components related to the determination of analyte quantity, material homogeneity, and material stability. These assigned uncertainties are reported as expanded uncertainties having defined coverage of the true value, typically with $V_i \pm U_{95}(V_i)$ providing coverage at a 95 % level of confidence.

The documents describing RMs seldom provide the computational details of how uncertainties are estimated nor are such details generally made publically accessible. Thus it is not generally possible to infer the assessed distribution for a given assigned value without access to privileged information. However, the 95 % coverage interval for a $\mathrm{N}\left(V_i, u_\infty^2(V_i)\right)$ normal kernel density provides the same coverage as is specified by $V_i \pm U_{95}(V_i)$ [28]. Expanded uncertainties specified with other than a 95 % level of confidence can also be converted into the $u_\infty(V_i)$ form; see [5] for details.

Thus to the extent that the V_i are correctly estimated and that the $U_{95}(V_i)$ do capture "all" of the relevant uncertainty components, the $\mathrm{N}\left(V_i, u_\infty^2(V_i)\right)$ provides a sufficient description of an RM's assigned value.

Repeatability Measurements (R)

Different measurement processes may be differentially influenced by material properties other than the quantity of the analyte (e.g., viscosity, particle size, or chemical interferences). The measurement process used to provide the repeatability results should be as free from such influences as possible. Demonstrating the equivalence of the nominal analyte content of materials using such a process does not guarantee the equivalence of the materials when assessed with other measurement systems. However, establishing this baseline equivalence should help other studies to better access the characteristics of other measurement processes.

For the GDR approach to be useful, the measurement results for a given material made under repeatability conditions must also be specified as a consistent set of expected values, R_i, and uncertainties, $u_\infty(R_i)$. There are two major considerations: the nature of the measurement process and the design followed in making the measurements.

Measurement Process

GDR analysis for the linear relationship of Equation 1 requires only that: 1) the relationship between the V_i and R_i be truly linear, 2) the measurement process be maintained in statistical control throughout the measurements, and 3) the imprecision of the measurement process be fit-for-purpose.

The measurement process must provide measurement results that fairly represent the analyte quantity in all of the study materials. The shorter the span of measurand quantities in the studied materials (e.g., the 1.02-fold range of CCQM-K53), the less important the linearity of the measurement process. The greater the span (e.g., the 3000-fold range of CCQM-K79), the more critical linearity becomes. Thus the nature of the materials to be evaluated will dictate the selection of the measurement process and, once selected, the capabilities of the process will dictate which materials are suitable for evaluation.

Maintaining a measurement process in true "repeatability condition" over anything but a truly short period of time may not be practical. However, use of internal and/or external controls can help provide measurements that are a precise as practical when measurements must be conducted over longer periods [29]. "As precise as practical" will depend on the measurement resources available for the comparison. As measurement imprecision increases, the analytical cost of the measurements is likely to decrease but it will become increasingly difficult to identify true differences among the RMs. We believe that the Goldilocks [30] "just right" precision will generally be about the median relative uncertainty of the assigned values, $u_\infty(V_i)/V_i$, of all materials included in the study. However, a method for more objectively determining what constitutes fit-for-purpose imprecision for a particular comparison should be developed [31].

Since both the V_i and R_i terms in Equation 3 are scaled by their respective uncertainties, it is immaterial whether or not the R_i, have the same scale as the V_i. That is, it does not matter whether or not the measurement process is externally calibrated. The nature of the analytical process used to make the repeatability measurements is not relevant as long as it has the appropriate stability, precision, and linearity characteristics. The process design should include suitable mechanisms for control of within- and between-campaign measurement drift, such as run-order randomization as well as the use of control materials.

Study design for single units of intrinsically homogenous materials

Figure 1a displays a measurement design appropriate for single units of materials that are of quite uniform composition, such as PSM cylinders. Obtaining two or more complete sets of measurements in independent measurement campaigns helps to ensure that any run-order effects are averaged over the materials. Making two or more replicate measurements within each campaign helps to ensure that technically invalid results can be recognized.

This simple one-factor with summary data design enables an adequate assessment of measurement variability when within-campaign imprecision is small compared to the between-campaign intermediate imprecision. A statistical model for this design is:

$$R_{ij} \sim \mathrm{N}\left(\mu_i, \sigma_{\mathrm{r},i}^2\right) \tag{5}$$

where i indexes the materials, j the campaigns, "~" indicates "is distributed as," μ_i is the population mean for the coordinating laboratory's measurements, and $\sigma_{\mathrm{r},i}$ is the true measurement repeatability imprecision for the material.

If the within-campaign imprecision is not small relative to the between-campaign imprecision, the design is readily extended (without requiring more measurements, assuming that at least two replicate measurements are made per campaign) using one-factor analysis of variance (ANOVA) to estimate the standard uncertainty of each mean value (see below).

The mean of the campaign means, R_i, is an appropriate estimate of μ_i. A standard uncertainty for this mean can be estimated from the standard deviation of the campaign means, $\hat{\sigma}\left(R_{ij}\right)$, and the number of campaigns, n_c:

$$u\left(R_i\right) = \hat{\sigma}\left(R_{ij}\right) \Big/ \sqrt{n_\mathrm{c}} \quad .$$

Assuming that all materials are measured in the same manner, the v for the $u(R_i)$ will be the same (n_c-1) for all materials.

In the four GAWG comparisons, the R_i were estimated as the mean of mean values from three or more measurement campaigns. The $u(R_i)$ were estimated either from the standard deviation of the mean of the campaign means, the standard deviation of the means pooled over all of the materials, or from the otherwise-determined repeatability/reproducibility characteristics of the measurement system.

Study design for batch-certified materials of uniform composition

Figure 1b displays a one-factor with replicated data design appropriate for batch-certified materials that are of quite uniform composition throughout each unit, such as presumptively homogenous solution RMs like the ethanol in aqueous solution materials of CCQM-K79. Since batch-certified units of any given material may differ slightly in composition, evaluating two or more units helps ensure representative assessment. While it is not productive to subsample a homogenous material, making independent replicate measurements can again help to average out measurement artifacts and to identify technically invalid results.

This design enables assessment of the uncertainty for the measurement of a material from between-unit and measurement repeatability sources of variability, enabling testing the measurement results for each material for significant between-campaign effects and appropriately pooling those results based upon the outcome of the tests. The least complex model for describing the measurements made using this design is:

$$R_{ijk} \sim \mathrm{N}\left(\mu_i + \gamma_{ij}, \sigma_{\mathrm{r},i}^2\right) \tag{6}$$

where j now indexes the units, k indexes the independent replicates per unit, $\sigma_{\mathrm{r},i}$ is assumed to be the same (or, equivalently, $\sigma_{\mathrm{r},i}/\mu_i$ is the same) across all units and γ_{ij} are between-unit differences. The γ_{ij} are assumed to be

$$\gamma_{ij} \sim \mathrm{N}\left(0, \sigma_{\mathrm{c},i}^2\right)$$

where $\sigma_{\mathrm{c},i}$ reflects the true between-campaign variability for the material. Depending on the nature of the measurement process, it may be desirable to make two or more measurements on each of the independently prepared replicates. As above, these replicate measurements can be averaged and their contribution to the total measurement variance disregarded when the within-replicate variability is small compared to the between-replicate or between-unit variances. Assuming the within-replicate variance is small, this design is readily evaluated using classical one-factor random effects analysis of variance (one-factor ANOVA); if it is not small relative to the other sources then a more complex ANOVA analysis may be useful (see below).

The mean of all measurements, R_i, is again an appropriate estimate of μ_i. A standard uncertainty of the R_i can be estimated as:

$$u\left(R_i\right) = \sqrt{\frac{n_\mathrm{r}\,\hat{\sigma}_{\mathrm{c},i}^2 + \hat{\sigma}_{\mathrm{r},i}^2}{n_\mathrm{r}\,n_\mathrm{c}}}$$

where n_r is the number of replicates and $\hat{\sigma}_{\mathrm{r},i}$ and $\hat{\sigma}_{\mathrm{c},i}$ are one-factor ANOVA estimates. If $\hat{\sigma}_{\mathrm{c},i}$ is not significantly greater than zero then this suggests the model of Equation 6 reduces to that of Equation 5 and the uncertainty estimate is associated with $\nu = n_\mathrm{c}n_\mathrm{r} - 1$ degrees of freedom; however, if $\hat{\sigma}_{\mathrm{c},i}$ is significantly greater than zero then $\nu = n_\mathrm{c}$-1. For example, assuming $n_\mathrm{c} = 2$ and $n_\mathrm{r} = 3$ then ν could range from 2·3-1 = 5 to 2-1 = 1. Thus, even though the different materials are measured in the same manner, the ν for the standard uncertainties may differ.

See the Supplementary Material for tests to estimate the statistical significance of variance components; for much more complete information on one-factor ANOVA, see, e.g., [32]. The Supplementary Material presents exemplar analyses using commercial software.

Study design for batch-certified materials of potentially non-uniform composition

Figure 1c displays a two-factor nested design appropriate for batch-certified materials that may *not* be of uniform composition throughout a given unit, such as granular solid or viscous liquid RMs like the creatinine in serum materials of CCQM-K80. For such materials, sub-sampling within each unit as well as evaluating two or more units can help ensure a representative assessment of each material. Making independent replicate measurements of each such aliquot of each material again helps to identify technically invalid results. And again it may be desirable to average two or more measurements on each independently prepared replicate and to disregard this imprecision component when the within-replicate imprecision is small compared to the between-replicate, between-aliquot, or between-unit imprecision.

This design enables assessment of the uncertainty associated with within- and between-material variability in addition to measurement repeatability, enabling testing the measurement results for each material for significant within- and between-campaign effects and appropriately pooling those results based upon the outcome of the tests. The least complex model for describing the measurements made using this design is:

$$R_{ijkl} \sim \mathrm{N}\left(\mu_i + \gamma_{ij} + \delta_{ijk}, \sigma_{\mathrm{r},i}^2\right) \tag{7}$$

where k now indexes the independent aliquots, l indexes independent replicates per aliquot, δ_{ijk} are between-aliquot differences, and $\sigma_{r,i}$ (or $\sigma_{r,i}/\mu_i$) is again assumed to be the same across all units of a given material. The δ_{ijk} are assumed to be

$$\delta_{ijk} \sim \mathrm{N}\left(0, \sigma_{a,i}^2\right)$$

where $\sigma_{a,i}$ reflects the true between-aliquot (or within-unit) material variability and is assumed to be the same (or $\sigma_{a,i}/\mu_i$ is the same) for all units. While potentially over-simplified, this model with these assumptions is likely fit-for-purpose given that RMs are designed to be homogenous and stable. Use of more complex models, e.g., allowing the $\sigma_{a,i}$ to be other than constant or proportional to analyte quantity, generally will require more data in order to provide reliable estimates.

As above, it may be desirable to make two or more measurements on each of the independently prepared replicates. These replicates can be averaged and their contribution to the total variance disregarded when the within-replicate variability is small compared to the between-replicate, -aliquot, or -unit variances. However, if the intrinsic measurement imprecision is not negligible then expanding the model to a three-factor nested design does not require any additional measurements and only modestly increases the complexity of the data analysis.

As in the one-factor designs, the mean of all measurements, R_i, is an appropriate estimate of μ_i. A standard uncertainty of this mean is typically estimated as

$$u\left(R_i\right) = \sqrt{\frac{n_r\, n_a\, \hat{\sigma}_{u,i}^2 + n_r\, \hat{\sigma}_{a,i}^2 + \hat{\sigma}_{r,i}^2}{n_r\, n_a\, n_u}}$$

where n_u is the number of units and n_a is the number of aliquots per unit. However, computing appropriate values for $\hat{\sigma}_{u,i}$, $\hat{\sigma}_{a,i}$, and $\hat{\sigma}_{r,i}$ is best done by an experienced data analyst using an appropriate method [33]. Software designed for the analysis of mixed linear models can be used such as the MIXED procedure in SAS [34] or the "lmer" function from the lme4 package for R [35]. The Supplementary Material presents exemplar analyses.

As with the one-factor with replicated data design, the absence of significant within- and between-unit variance components reduces the model of Equation 7 to that of Equations 5 or 6 and to standard uncertainty estimates for different materials having different v. The v can range from as few as n_u-1 when the units are significantly different to as many as $n_r n_a n_u$-1 when there are no appreciable differences in the measurements among the units and the aliquots. The v for different scenarios and tests for estimating the statistical significance of variance components are presented in the Supplementary Information.

Uncertainty Estimates for the Gaussian Kernels

As noted above, the one-factor with summary data design provides $u(R_i)$ estimates that all have the same v. Regardless of whether or not these standard uncertainties are converted into "large sample" form, the conversion will be the same for all materials and will thus provide a consistent description for all materials.

Also as noted above, designs that explicitly consider replicated data can provide $u(R_i)$ estimates that have very different v. Defining the results as $\mathrm{N}\left(R_i, u^2\left(R_i\right)\right)$ kernels may thus not provide a consistent description for all of the materials. There are a number of ways to address this issue, including modifying the GDR technology to explicitly utilize the v information [16]. However, the currently available GDR software can be used by first expanding the standard uncertainties into $U_{95}(R_i)$ form and then applying the

logic of Equation 4 to give "large sample standard uncertainties" $u_\infty(R_i)$. These estimates will provide a consistent description for all of the materials.

Unfortunately, if ν is small for any of the materials there may be no completely satisfactory way to expand the standard uncertainties.

Student's *t* expansions

For a given $u(R_i)$ with a given number of degrees of freedom, ν_i, the usual long-term frequency (frequentist) method for estimating a 95 % level of confidence interval is expansion with the appropriate two-tailed Student's *t* factor, $t_s(0.05,\nu_i)$:

$$U_{95}(R_i) = t_s(0.05,\nu_i)u(R_i) \quad .$$

The issue now becomes one of determining ν_i.

When results for all units (and aliquots) are indistinguishable, then the ν_i will be 1 less than the total number of replicates. Unfortunately, when there are relatively large differences between the measurement results for the different units, ν_i will be n_u-1 regardless of the number of replicates and aliquots. While this issue is relatively unimportant should n_u be larger than about 4 (giving a minimum ν of 3), it causes a headache for n_u of 2 or 3: $t_s(0.05,1) \approx 12.7$, $t_s(0.05,2) \approx 4.30$, $t_s(0.05,3) \approx 3.18$, $t_s(0.05,4) \approx 2.78$, $t_s(0.05,54) \approx 2.00$, and $t_s(0.05,\infty) \approx 1.96$. The huge change from ν of 1 to ν of 2 merely reflects that variance estimates using a small number of measurements are not reliable.

Pooling across materials: Pooling the various variance component estimates across materials can provide estimates for some to all of the components that will have larger and more reliably estimated combined ν. For instance, the CCQM-K54 report states "Prior to regression, the standard uncertainty associated with the response $u(y)$ has been pooled. The differences in these uncertainties … are mainly due to random effects and the limited number of degrees of freedom: all $\nu_i = 2$. Consequently, there is no point in assuming that one response would be more accurate than another." A pooled variance expressed in standard deviation form is calculated as:

$$\sigma = \sqrt{\sum_i^n \nu_i \sigma_i^2 \Big/ \sum_i^n \nu_i}$$

where the σ_i^2 is the individual variance, n is the number of individual variances, and ν_i is the number of degrees of freedom for each estimate. The ν for the pooled variance is the sum of the individual ν_i. Thus, even when the individual variances have $\nu = 1$, pooled estimates can have large ν and thus be quite reliable.

However, pooling can only be justified when the variance estimates actually represent the same variance components. To the extent that repeatability conditions are maintained over all campaigns, this is likely to be true for between-replicate estimates. When the sample preparation process is complex and applied uniformly to all materials, it may be appropriate for between-aliquots. And, most unfortunately, it is unlikely to be true for between-unit estimates.

Bayesian Markov Chain Monte Carlo (MCMC) Estimation

Bayesian analysis is based on a somewhat different definition of probability than the usual frequentist interpretation that underpins classical statistical inference. Under the Bayesian paradigm, parameters such as the measurand value and variance components have probability distributions that quantify our knowledge about them. The estimation process starts with quantification of prior knowledge about the

parameters followed by specification of the statistical model that relates the parameters to the data. The statistical model for the three designs considered here is the same as described in Equations 5, 6, and 7.

The components of these models are combined via Bayes Theorem to obtain posterior distributions for the parameters. These distributions update our knowledge about the parameters based on the evidence provided by the data. This analysis can produce a probability distribution for each of the μ_i (the true value of analyte quantity estimated by the measurement mean, R_i) which encompasses all of the information and variability present in the data but is confined by bounds based on prior knowledge. The process yields a probability interval which is interpretable as an uncertainty interval [36]. Even though the probability distribution for the μ_i may not be available in closed form, Markov Chain Monte Carlo (MCMC) empirical Bayesian methods enable computation of coverage intervals. Software systems suitable for computing the intervals, such as WinBUGS and OpenBUGS [37], are freely available and (relatively) easy to use.

Ideally, Bayesian analysis can proceed using very conservative, minimally-informative priors (e.g., very broad Gaussian distributions for the μ_i) and let the data mostly determine the posterior distribution of the measurand. Unfortunately, unconstrained Bayesian MCMC component of variance models with small v may yield unreasonably large variance estimates [38]. In general, somewhat informative priors are required for n_u less than three. However, when these priors are carefully defined the analysis can validly produce probability distributions for the μ_i which encompass the available information on the materials and all of the variability present in the data. The Supplementary Material presents exemplar WinBUGS analyses for one-factor with replicated data and two-factor nested designs.

Confounding of Campaigns and Units

The multi-unit designs pictured in Figure 1 assume that all analyses can be conducted in one continuous measurement campaign. This approach would best approximate true "repeatability conditions" while enabling averaging-out run-order and other potential measurement process artifacts. However, as previously described, making all measurements in a single campaign may not be practical for studies requiring a very time-consuming measurement process or involving a large number of materials. In such cases, the study must be designed to control the influence of possible between-campaign changes in the measurement process.

"Cross classification" experimental designs could best enable identification of the true sources of variation. Such designs would require measuring the different replicates (and aliquots) from each of the units in separate campaigns (see [32] for further information). However, these replicates (and aliquots) would perforce need to be processed at different times with the opened unit stored between campaigns and/or the processed replicates and aliquots themselves stored between campaigns. In either case, the materials could degrade or become contaminated during storage. Storing an opened unit might also violate material usage requirements specified by the organization that value-assigned the material.

The one-factor with replicated data and two-factor nested designs in Figure 1 are appropriate as presented if all replicate measurements for a given unit are made during a single campaign. However, this confounds any true between-unit variability with potential systematic between-campaign changes in the measurement process. That is, it will not be possible to differentiate between $\sigma_{c,i}$, and $\sigma_{u,i}$ as any observed variability could be attributed to either source. For the remainder of this document, we use $\sigma_{c,i}$ to refer to this confounded variability.

Ensuring the integrity of the study materials by measuring each unit in a single campaign seems the better choice, at least for material preparations that could change composition during storage. The resulting confounding of between-campaign and between-unit variance may also help to avoid unwarrantedly identifying particular materials as "significantly variable" on the basis of too few data.

Number of Materials

The number of materials included in an RM comparison is constrained to be between the number required for a reliable analysis of the results and the number of suitable materials that are available, the number of measurements that can be accomplished in a single campaign, or the resources available to the coordinating laboratory. Using the linear model of Equation 1, it is possible to detect the non-equivalence of one material in a set of just three materials. However, it is not possible to identify which material is non-equivalent with only three materials. The practical lower limit on the number of materials is thus four.

While the cost of a comparison increases with the number of measurements, the scope and reliability of the conclusions that can be reached increases with the number of materials compared. For many measurands, the upper limit on the number of RMs that can be compared is likely to be the number of available RMs. Should there be more organizations that wish to participate in a comparison than can be accommodated by the coordinating laboratory, the study should be redesigned or more resources allocated so that all qualified organizations can participate. Should the number of participating organizations be tractable but the total number of available RMs be too large, then selection criteria need to be established.

While it is desirable that all participants be represented by about the same number and diversity of materials, the scope of a study is improved by including as diverse a collection of materials as is possible. While what constitutes "diversity" depends upon the study's nature, at least two participants should provide materials that represent each aspect of that diversity to ensure that limitations or artifacts of the measurement process can be recognized. Such aspects include known differences in sample matrix (e.g., lyophilized vs. frozen or with vs. without added preservative) as well as the bottom and top ends of the analyte quantity range. Different studies may require different trade-offs between balanced representation and breadth of scope, but in all cases it is essential that the selection process be transparent and fair to all participants.

Number of Units Per Material

While comparisons of individually value-assigned materials are relatively simple to summarize, such comparisons do not address all of the potential sources of variation in batch value-assigned RMs. In particular, comparisons of single units do not address the between-unit heterogeneity that is often a significant component of batch-assigned uncertainty. Making measurements on two or more units of batch-certified materials helps avoid attributing differences to bias in the value-assignment rather than to variability in material composition or analytical mistakes. We thus recommend that the minimum number of units per material for batch-certified materials is two. But what is the "best" number? There are several considerations, including: the purpose of the study, the representativeness of the measurements, and the cost of the materials and measurements.

ISO Guide 35:2006 "Reference materials – General and statistical principles for certification" suggests that 10 units is a minimum number for evaluating homogeneity [39] in a newly produced RM. However, comparisons between previously value-assigned materials are not intended to discover heterogeneity and thus can use fewer than 10 units of each material. While summary statistics will become more

representative with increasing number of units, the improvement is not linear. The 95 % confidence t_s factors discussed above suggest that an expanded uncertainty based on the analysis of five units should be fairly representative of the material. While Bayesian and related techniques can be used to estimate credible uncertainty intervals directly from distributional assumptions and the "raw" measurements, the v for each factor should be at least three unless outside constraints can be justified. These considerations suggest that evaluating three to five units of each material would simplify the data analysis and be adequately representative for metrological equivalency.

The cost to the value-assigning organizations increases nearly linearly with the number of units per material requested, but for most materials the cost per material is relatively small and in any case is a shared burden. The cost to the coordinating laboratory also increases with increasing number of units. While this cost depends on the nature of the measurement process, it is in any case borne by just that one laboratory. As this cost can be quite substantial, the decision as to the exact design should be largely based on the analytical resources available to, and affordable by, the coordinator.

Prepare for the Unexpected

The minimum number of units per material requested from participating institutions should be at least one more than the number to be evaluated. The extra unit(s) provide an "insurance policy" in case of technical failure or in case a participant should challenge the measurement results for one or more of their materials. In the latter case, a third party should be asked to evaluate the challenged result(s).

Step 2: Make the Measurements

The coordinating laboratory has the responsibility of providing the highest practical quality measurements for each of the study materials. While quite demanding, the challenges involved are routine for organizations that produce certified reference materials.

While the diversity of measurement processes that may be appropriate for RM comparisons precludes specific discussion, the following general principles apply. 1) To ensure that the results of the comparison can be accepted by all of the participating institutions and any sponsoring body such as the CCQM, all of the processes employed - from receiving the materials to summarizing the results - must be transparent and open to inspection. 2) As noted previously, the term "practical" is definable only within the context of the particular measurand, the study materials, and the available analytical resources. However, for a study to meaningfully speak to the equivalence of RMs or the processes used in their value assignment, the relative imprecision of the results from the measurement process should be as small or smaller than the median relative assigned uncertainty of the study materials. 3) For studies involving lengthy measurement processes or multiple measurement campaigns, periodic analysis of one or more control materials may be necessary for achieving the desired small measurement imprecision. For perfectly stable materials measured under perfectly stable conditions, control measurements can provide unambiguous evidence of that stability. For materials prone to degradation or measurement processes prone to within- and/or between-campaign drift, the control measurements can be used to at least partially compensate for systematic changes [29]. The coordinating laboratory' analysts must use their knowledge of their measurement processes to decide on the number and nature of such controls and how often they are measured.

Step 3: Establish the Consensus Model

We term a consensus-defined relationship between assigned, $\{V_i, u(V_i)\}$, and repeatability measurement, $\{R_i, u(R_i)\}$, values to be a Reference Function (RF). The CO_2 data of GAWG's CCQM-P41 are used in this and the next Sections to illustrate many of the data analysis issues that must be addressed in establishing an RF. These data used are as reported in Table 7 of [23]. These data have a number of properties that make them nearly ideal exemplars for this discussion: there are only nine materials so that the influence of individual materials can easily be envisioned, the range in analyte quantity is sufficiently narrow so that all of the values and their uncertainties can be visualized, the degrees of equivalence among the materials are good but imperfect, and the data have been published in the open literature.

The GDR Reference Function, RF

Assumptions

Equation 1 is our model for straight-line relationships. While measurement systems exist that are non-linearly related to the true amount of substance [15], we confine our discussion to two-parameter (intercept and slope) linear relationships.

We define the uncertainties for both the V_i and R_i as one-half of their expanded uncertainties at the 95 % level of confidence. As in Equation 4, we describe these estimates as "large sample standard uncertainties" and use the notation $u_\infty(V_i)$ and $u_\infty(R_i)$. We regard the $\{V_i, u_\infty(V_i)\}$ and $\{R_i, u_\infty(R_i)\}$ values as defining mutually independent Gaussian $\mathrm{N}\left(V_i, u_\infty^2(V_i)\right)$ and $\mathrm{N}\left(R_i, u_\infty^2(R_i)\right)$ kernel distributions.

Estimating the RF

Replacing the generic estimate of sample variability or uncertainty of Equation 3 with the corresponding large-sample standard uncertainties, the GDR estimate of a RF minimizes the sum of the squared uncertainty-normalized differences between the observed values and their projection onto the consensus line:

$$E = \sum_i^{N_m} \varepsilon_i^2; \quad \varepsilon_i^2 = \left(\frac{R_i - \hat{R}_i}{u_\infty(R_i)}\right)^2 + \left(\frac{V_i - \hat{V}_i}{u_\infty(V_i)}\right)^2; \quad \hat{R}_i = \hat{\alpha} + \hat{\beta}\hat{V}_i \quad . \tag{8}$$

The RF is defined by the estimated intercept and slope parameters $\hat{\alpha}$ and $\hat{\beta}$. The $\{\hat{V}_i, \hat{R}_i\}$ pairs estimate the uncertainty-weighted projection of the observed data onto the RF.

Envisioning the GDR estimates

Figure 2a is the primary RegViz graphical display for the analysis of the CCQM-P41 CO_2 data, using the assumption that all of the CCQM-P41 PSMs value-assignments are equally valid. This display of the data confirms the basic linearity of the repeatability measurement system. It also identifies a potential anomaly with one or both of the low CO_2 amount PSMs. However, the display does not provide sufficient graphical resolution to visualize the very small uncertainties for many of the materials even over the narrow 1.1-fold range of CO_2 amount. Figure 2b provides a high-resolution display as a series of "thumbnail" scatterplots, one per PSM.

Figure 3 illustrates key attributes of the ellipse shown in each of thumbnails. Since "large sample" standard uncertainties are used in Equation 8, the ε_i have units of "large sample standard uncertainty." Since the ε_i^2 are defined from two independent variables, they are χ^2 distributed with two degrees of freedom. The value that corresponds to 95th percentile for such distributions is about 5.99, thus the 95 %

critical value for the ε_i is $\sqrt{5.99} = 2.45$ rather than the $\sqrt{3.84} = 1.96 \approx 2$ appropriate for univariate systems that are χ^2 distributed with one degree of freedom. Therefore, each ellipse bounds all points that are 2.45 ε_i distant from $\{V_i, R_i\}$. An uncertainty-weighted distance of greater than 2.45 units suggests that the material is not equivalent to the ensemble of all study materials. Thus the CCQM-P41 PSMs A, B, C, G, and perhaps F may be not equivalent. However, the true level of confidence provided by this test is less than 95 % since it does not consider the uncertainty in the RF itself.

The GDR Reference Function Uncertainty, U_{95}(RF)

The uncertainty interval on the RF may have a different width at every locus along the RF line. It is relatively straightforward to estimate Working-Hotelling confidence intervals for variance-weighted OLS regressions [9]. In principle, similar confidence intervals for GDR can be calculated from *E*; the $\hat{\alpha}$, $\hat{\beta}$, $\hat{V}_i$,and $\hat{R}_i$ parameter estimates; and the covariances among these parameters. In practice, GDR confidence intervals are more readily estimated by simulation.

Parametric bootstrap Monte Carlo (PBMC)

The parametric bootstrap Monte Carlo (PBMC) is a powerful tool for estimating the distribution of outcomes from a complex analysis [36,40,41]. In PBMC GDR analyses, the entire ensemble of $\{V_i, R_i\}$ pairs are replaced a fairly large number of times, n_{MC}, with pseudo-values randomly drawn from the $\mathrm{N}\left(V_i, u_\infty^2(V_i)\right)$ and $\mathrm{N}\left(R_i, u_\infty^2(R_i)\right)$ kernels and a new GDR analysis performed on the ensemble of pseudo-values. All of the "interesting" results estimated from each PBMC analysis are stored and, after a suitably large number of analyses, the distribution of each result empirically summarized. For GDR analysis, these results include but are not be limited to the RF parameters and the $\{\hat{V}_i, \hat{R}_i\}$.

PBMC is typically used to estimate uncertainty intervals from the empirical percentiles of the distributions. For example, the two-tailed 95 % level of confidence interval for some result *Q* of the GDR analysis of the original data can be estimated from the 2.5 % and 97.5 % percentiles of the PBMC estimates of *Q*, call them *q*:

$$PTILE(2.5, q) \le Q \le PTILE(97.5, q)$$

where "PTILE(*p*,*q*)" is the function "return the p^{th} percentile of the specified list of *q* values." If the ratio $(Q - \mathrm{PTILE}(2.5, q))/(\mathrm{PTILE}(97.5, q) - Q)$ is about 1, then the usual 95 % confidence interval on *X* can be estimated as:

$$U_{95}(Q) = (\mathrm{PTILE}(97.5, q) - \mathrm{PTILE}(2.5, q))/2 \quad .$$

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

$$^{-}U_{95}(Q) = Q - \mathrm{PTILE}(2.5, q) \qquad ^{+}U_{95}(Q) = \mathrm{PTILE}(97.5, q) - Q$$

or as the larger of the two half-intervals

$$\pm U_{95}(Q) = \mathrm{MAX}\left(^{-}U_{95}(Q), ^{+}U_{95}(Q)\right)$$

where "MAX(.)" is the function "return the largest value of the arguments." Asymmetric intervals are the narrowest intervals that provide the stated coverage; however, symmetric over-estimates may be more convenient for use in further calculations.

When only the central 95 % of the distributions are of interest, n_{MC} typically needs to be only a few hundred to several thousand for $U_{95}(q)$ estimates to be stable to two significant digits. The only tools

needed to perform a PBMC analysis (other than the data analysis system used to analyze the original data) are mechanisms for: 1) generating the pseudo-values, 2) re-doing the analysis using the pseudo-values, 3) storing the results, and 4) summarizing the resulting lists of values. This can be done "by hand" but is most conveniently done with specialized software. Uniquely among the available GDR freeware, the RegViz system estimates the variability for all estimated quantities using PBMC techniques.

Strongly influential materials and leave-one-out (LOO) analysis

The RF can be strongly influenced by materials having small $u_\infty(V_i)$ and/or $u_\infty(R_i)$. Moreover, the magnitude and nature of this influence is strongly affected by the amount of the analyte in the material relative to that of the other materials. Materials with the lowest and highest quantity amounts have greater influence on the RF than those in the middle of the ensemble.

Leave-one-out cross-validation (LOO) is a routine tool for improving the predictive utility of a model [40]. For RM comparisons, LOO can be accomplished by excluding individual materials from their own evaluation. Materials that are equivalent to the ensemble when excluded from the analysis are reliably identified as truly equivalent. Materials that appear to be equivalent when included but are "non-equivalent" when excluded are identified as being at best marginally equivalent. The upper and lower panels of Figure 4 compare the "leave all in" (LAI)-PBMC and LOO-PBMC 95 % confidence intervals for the CCQM-P41 CO_2 data. The LOO-PBMC U_{95}(RF) interval is considerably wider, indicating the presence of one or more materials that strongly influence the RF.

A disadvantage of the LOO approach is that it requires the estimation of as many "individual-RF"s as there are materials. However, only the product $n_{MC}n_m$ needs to be large for reliable 95 % level-of-confidence intervals on the RF parameters. For stable summarization of results for specific materials, n_{MC} itself needs to be large.

Given the RF uncertainty, an inexact but reasonable graphical test for non-equivalency is for the entire 95 % level of confidence interval on the RF to be exterior to the ellipse. By this test and using the LOO-PBMC intervals, CCQM-P41 PSMs A, C, and G are not equivalent to the ensemble with at least a 95 % level of confidence.

Identifying influential materials

Figure 5 illustrates an exploratory graphical tool for identifying highly influential materials, comparing the ε_i estimated with LAI-PBMC with those estimated with LOO-PBMC. Circles with crosses that are not substantially on the diagonal line indicate materials that strongly influence the RF. Circles outside the ±2.45 standard uncertainty square are potentially anomalous, most probably as a result of underestimated $u_\infty(V_i)$ and/or $u_\infty(R_i)$. While PSMs B, F, G, C, and A appear potentially anomalous, only PSM A appears strongly influential.

Identifying consequential materials

While the presence of a material in the GDR model may or may not strongly influence the consensus solution, it presence may have consequence for the other materials in the study. Figure 6 visualizes the consequences of including each of the CCQM-P41 materials in the GDR model. Inclusion of a material has a negative consequence when its presence causes the ε_i of one or more other materials in a given PBMC iteration to change from being less than the critical distance of 2.45 to being greater than that value. Likewise, inclusion of the material has a positive consequence when its presence causes the ε_i of one or more other materials to change from being greater than 2.45 to being less than that value. Defining "strongly consequential" as causing at least one material's ε_i to shift across the critical distance at least half the time, PSMs C, G, and A are strongly consequential.

Effect of excluding influential and/or consequential materials

Strongly influential and consequential materials must be identified and decisions made on whether or not they should be included in the RF estimation. Figure 4c displays the results for the CCQM-P41 data after excluding PSM A from the RF evaluation. The other eight PSMs appear to be metrologically equivalent, albeit the $u_\infty(V_i)$ for C, G, and I appear to be underestimated.

Step 4: Determine the Degrees of Equivalence

Definitions for "Degree of Equivalence" for Individual Materials

For comparisons of results for nominally identical samples of one material, the degree of equivalence for a given result is defined in the CIPM MRA as: [7]

$$d_i = x_i - x_{\text{ref}} \tag{9}$$

where x_i is the reported result and x_{ref} is the reference value. The best possible d_i is zero, when the result is identical to x_{ref}. However, this definition is inadequate for studies where two equivalently important sets of measurement results are compared.

Over a narrow analyte amount range
The signed orthogonal distance provides a consistent definition of the degree of equivalence for multi-material studies where the uncertainties are approximately independent of the analyte quantity amount:

$$d_i = \text{SGN}\left(V_i - \hat{V}\right)\sqrt{\left(V_i - \hat{V}\right)^2 + \left(\left(R_i - \hat{R}_i\right)/\hat{\beta}\right)^2} \tag{10}$$

where the function SGN(.) returns the signum (sign, ±1) of its argument, $\text{SGN}\left(V_i - \hat{V}\right)$ defines whether the observed $\{V_i, R_i\}$ pair is "above" or "below" the RF line, and division by $\hat{\beta}$ ensures that the measurement-related terms have the same scale as the assigned values.

Figure 7 contrasts the definitions of Equations 9 and 10 for three materials with measured V and R results that lie on a line parallel to the RF but have very different measurement uncertainties: one with $u(R_1) >> u(V_1)$, the second with $u(R_2) = u(V_2)$, and the third with $u(R_3) << u(V_3)$. Using Equation 9, material$_1$ (relatively heterogenous with an underestimated assigned uncertainty) would be considered more equivalent than material$_3$ (relatively homogenous with an overestimated assigned uncertainty). Using Equation 10, these materials are equivalently equivalent.

Over a wide analyte amount range
The definition of Equation 10 is not suited for studies involving materials having $u(R_i)$ and $u(V_i)$ that are proportional to the quantity amount ($u(R_i) \propto \mu$ and $u(V_i) \propto \mu$). For such materials, it is more appropriate to consider the percent *relative* degree of equivalence:

$$\%d_i = 100\frac{d_i}{\left(V_i + \left(R_i - \hat{\alpha}\right)/\hat{\beta}\right)/2}$$

where the R_i are transformed through the RF to have the same origin and scale as the V_i.

Definition of the Uncertainty of the Degree of Equivalence for Individual Materials

For comparisons where Equation 9 is appropriate, the first-order uncertainty estimate for d_i is:

$$U_p(d_i) = k_p\, u(d_i),$$
$$u(d_i) = \sqrt{u(x_i)^2 + u(x_{\text{ref}})^2 - 2\,\rho(x_i, x_{\text{ref}})u(x_i)u(x_{\text{ref}})}$$

where k_p is the coverage factor intended to yield an expanded uncertainty such that the interval $d_i \pm U_p(d_i)$ includes the true degree of equivalence with a p-level of confidence, $u(x_i)$ is the standard uncertainty of

the reported value, $u(x_{ref})$ is the standard uncertainty of the reference value, and $\rho(x_i, x_{ref})$ is the correlation between the reported value and the reference value. The magnitude of the correlation depends whether the x_i was used to help define the x_{ref} and, if so, how the x_{ref} was estimated. Note that when x_{ref} is a consensus estimate, ignoring the correlation term leads to larger $u(d_i)$ and thus to optimistic equivalence claims. When the interval $d_i \pm U_{95}(d_i)$ contains zero then the uncertainties in the two terms are considered to adequately account for the observed bias.

For multi-material studies over either a narrow or wide range of analyte amount

Given the correlations among the $\hat{\alpha}$ and $\hat{\beta}$ parameters of the RF and the $\{V_i, u_\infty(V_i), R, u_\infty(R_i)\}$ from which they were estimated, estimating $u(d_i)$ or $u(\%d_i)$ and $U_{95}(d_i)$ or $U_{95}(\%d_i)$ with the usual Taylor's series approximation is a complex and a somewhat daunting task. However, PBMC again provides a simple and transparent method for estimating the relevant uncertainties: evaluate the d_i or $\%d_i$ during each of the n_{MC} PBMC analyses and then empirically summarize these results as described above. This includes all of the correlations in the estimates without requiring their explicit estimation.

The effect of the $\rho(x_i, \hat{\mu})$ correlation can be eliminated for each of the materials by using LOO-PBMC results to estimate the uncertainties. This results in wider uncertainty intervals for most materials, with the magnitude of enlargement depending on the number of materials in the data set, the magnitudes of the observed uncertainties, and the location of the observed values within the data set.

Reporting RF Degrees of Equivalence for Materials

When symmetric 95 % level of confidence intervals are used, the d_i or $\%d_i$ PBMC-estimated $d_i \pm U_{95}(d_i)$ or $\%d_i \pm U_{95}(\%d_i)$ can be reported using the same tabular and graphical formats as used with single-material studies. Should asymmetric intervals be deemed the more appropriate summaries, the $-U_{95}$ and $+U_{95}$ estimates should be treated separately. Figure 8a illustrates a dot-and-bar plot of $d_i \pm U_{95}(d_i)$ intervals for the entire ensemble of CCQM-P41 materials as estimated using LOO-PBMC. This plot differs from the usual format used for single-material comparisons only in that the horizontal axis displays the certified values rather than an equally-spaced and often arbitrarily ordered label for each result.

The Effect of Excluding Consequential Materials

While highly influential and consequential values can strongly bias an reference value when estimated using methods that are inappropriate to the actual distribution of values [42], the resulting bias affects all of the values in the study in the same manner and is therefore likely to be closely scrutinized. This is not true with reference functions. Figure 8b displays the $d_i \pm U_{95}(d_i)$ for the CCQM-P41 materials with Material "A" excluded from the RF calculations. Note in particular the changes for materials "B", "D", and "I": the effect of excluding a material is dependent on the relative locations of the materials. *It is thus critical that any and all concerns about influential materials be identified and resolved quickly and transparently.*

Definitions for "Degree of Equivalence" for Organizations

For RM comparisons that involve more than one material from some participating organization, the results for all of the materials from each participating organization need to be combined in some manner to estimate the degrees of equivalence of that participant's certification capabilities. While the

established methods for combining such results have pitfalls [43], assuming Gaussian distributions and averaging the $d_i \pm U_{95}(d_i)$ for every material has the advantage of simplicity. A degree of equivalence for each participant, $D \pm U_{95}(D)$, can be estimated as:

$$D = \sum_{i=1}^{n} d_i \Big/ n; \quad U_{95}(D) = 2\,u(D)$$

$$u(D) = \begin{cases} U_{95}(d_1)/2 & n = 1 \\ \sqrt{\sum_{i=1}^{n} \left(\frac{U_{95}(d_i)}{2} \right)^2 \Big/ n + \left(\sum_{i=1}^{n} (d_i - D)^2 \Big/ n - 1 \right)^2} & n > 1 \end{cases}$$

where i indexes over the n materials submitted by that participant.

While relatively straightforward, defining $u(D)$ from the between- and within-material uncertainties does not address the possible influence of non-Gaussian distributions. However, the same PBMC analysis used to estimate the $d_i \pm U_{95}(d_i)$ can be used to estimate the $D \pm U_{95}(D)$. Assuming that the same sufficiently large number of pseudo-values are available for all materials:

$$D_{\text{abs}} = \text{PTILE}\left(50, \bigcup_{i=1}^{n} \left(d_{\text{abs}\,i} \right) \right)$$

$$U_{95}(D_{\text{abs}}) = \left(\text{PTILE}\left(97.5, \bigcup_{i=1}^{n} \left(d_{\text{abs}\,i} \right) \right) - \text{PTILE}\left(2.5, \bigcup_{i=1}^{n} \left(d_{\text{abs}\,i} \right) \right) \right) / 2$$

where $\bigcup_{i=1}^{n} \left(d_i \right)$ is the union of all the PBMC pseudo-values for the materials submitted by the participant. This estimate for $U_{95}(D)$ assumes that the distribution of the combined d_i is approximately symmetric. If this assumption is not viable, asymmetric intervals can be estimated as discussed above. Relative degrees of equivalence, $\%D \pm U_{95}(\%D)$, can be estimated using the same formula applied to the analogous $\%d_i$ estimates.

Reporting Degrees of Equivalence for Certification Capabilities

The $D \pm U_{95}(D)$ or $\%D \pm U_{95}(\%D)$ can be reported in the same formats used with the d_i or $\%d_i$ for individual materials. However, with the focus on participants rather than materials, the horizontal axis of the dot-and-bar graphs should use the ordering formats used with single-material comparisons. In addition, it is informative to plot the component $d_i \pm U_{95}(d_i)$ dot-and-bars to one side of their summary $D \pm U_{95}(D)$ values.

Figure 9a displays $d_i \pm U_{95}(d_i)$ for the CCQM-P73 study, where each participant prepared two NO in N_2 PSMs of somewhat differing NO amount [25]. The dot-and-bars are labeled with a code character for each participant followed by a digit reflecting the relative NO amount. While a number of the materials are not equivalent, none of the materials has high influence.

Figure 9b displays the $D \pm U_{95}(D)$ for the CCQM-P73 data, with the estimates for the participants displayed in alphabetical order. The $d_i \pm U_{95}(d_i)$ for the participant's materials are displayed to the immediate right of each participant's $D \pm U_{95}(D)$ dot-and-bar. Displaying the $d_i \pm U_{95}(d_i)$ along side the $D \pm U_{95}(D)$ enables the viewer to identify the relative consistency of the evidence for each participant.

Acknowledgements

We thank Johanna E. Camara (NIST, Gaithersburg) and Rosemarie Phillips (BAM, DE) for their insights and experience in the practicalities of CRM comparison measurements; Wolfram Bremser (BAM, DE) and Stephen L.R. Ellison (LGC Limited, UK) for helpful discussions on approaches to analyzing CRM comparison results, Jolene D. Splett (NIST, Boulder) for her expertise with SAS and the interpretation of its results, Chih-Ming Wang (NIST, Boulder) for his careful review (and correction) of statistical concepts and notation, Katherine E. Sharpless (NIST, Gaithersburg) for her efforts towards making this document more accessible, and the anonymous reviewers who carefully and thoughtfully critiqued the original draft of this report.

References

1 JCGM 200 (2008) International vocabulary of metrology - Basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology. Sèvres, France. http://www.bipm.org/en/publications/guides/vim.html

2 Emons H. The 'RM family' – Identification of all of its members. Acced Qual Assur 2006;10:690-691.

3 Armbruster D, Miller RR. The Joint Committee for Traceability in Laboratory Medicine (JCTLM): A Global Approach to Promote the Standardisation of Clinical Laboratory Test Results. Clin Biochem Rev 2007;28(3):105–114.

4 JCTLM (2006) Joint Committee for Traceability in Laboratory Medicine Quality System Procedure JCTLM WG1-P-04A Process for Comparing Certified values of the Same Measurand in Multiple References materials (CRMs). http://www.bipm.org/utils/en/pdf/WG1-P-04A.pdf.

5 Duewer DL, Lippa K, Long SE, *et al*. Demonstrating the comparability of certified reference materials. Anal Bioanal Chem, 2009;395(1):155-169.

6 Status accessible through: http://kcdb.bipm.org/appendixB/KCDB_ApB_search.asp

7 CIPM. 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. http://www.bipm.org/en/cipm-mra/documents/

8 Kummell CH. Reduction of observation equations which contain more than one observed quantity. The Analyst (Annals of Mathematics) 1879;6(4):97-105.

9 Feigelson ED, Babu GJ. Linear Regression in Astronomy. II. Astrophysical J 1992:397:55-67.

10 Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 1901;6(2):559-572.

11 Deming WE. Statistical adjustment of data. John Wiley & Sons, NY (1943).

12 Christian SD, Tucker EE. LINGEN - A general linear least squares program. J Chem Educ 1984;61(9):788.

13 Ripley BD, Thompson M. Regression Techniques for the Detection of Analytical Bias, Analyst 1987;112(4):337-383.

14 ISO. ISO 6143:2001 Gas analysis - Comparison methods for determining and checking the composition of calibration gas mixtures, International Organization for Standardization. Geneva, 2001.

15 Milton MJT, Harris PM, Smith IM, Brown AS, Goody BA. Implementation of a generalized least-squares method for determining calibration curves from data with general uncertainty structures. Metrologia 2006;43(4):S291-S298

16 Guenther FR, Possolo A. Calibration and uncertainty assessment for certified reference gas mixtures. Anal Bioanal Chem 2011;399:489-500.

17 Toman B, Duewer DL, Gasca Aragon H, Guenther FR, Rhoderick GC. A Bayesian approach to the evaluation of comparisons of individually value-assigned reference materials. Anal Bioanl Chem, 2012;403:537-548.

18 Microsoft Corporation, Redman, WA USA. http://office.microsoft.com/en-us/excel/

19 Analytical Methods Committee (AMC). Linear Functional Relationship Estimation by Maximum Likelihood. http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/FREML.asp

20 National Physics Laboratory. XLGENLINE. http://www.eurometros.org

21 Van Huffel. Total Least Squares and Errors-in-variables Modeling. Comp Stat Data Anal 2007;52:1076-1079.

22 Bartholomew-Biggs M, Butler BP, Forbes AB. Optimization algorithms for generalized distance regression in metrology, in Ciarlini P, Forbes AB, Pavese F, Richter D (eds), Advanced Mathematical and Computational Tools in Metrology IV, Series on Advances in Mathematics for Applied Sciences, 2000;53:21-31.

23 van der Veen AMH, Brinkmann FNC, Arnautovic M, *et al*. International comparison CCQM-P41 Greenhouse gases. 2. Direct comparison of primary standard gas mixtures. Metrologia 2007;44:08003

24 NPL. What is a Primary Standard Gas Mixture (PSM)? (FAQ - Gas Standards). http://www.npl.co.uk/science-technology/chemical-metrology/faqs/what-is-a-primary-standard-gas-mixture-(psm)-(faq-gas-standards) (2007) accessed 25-May-2012

25 Wielgosz RI, Esler M, Viallon J, *et al*. International Comparison CCQM-P73: nitrogen monoxide gas standards (30-70) µmol/mol. Metrologia 2008;45:08002

26 van der Veen AMH, Chander H, Ziel PR, *et al*. International comparison CCQM-K54: primary standard gas mixtures of hexane in methane. Metrologia 2010;47:08019

27 Lee J, Lee JB, Moon DM, *et al*. Final report on international key comparison CCQM-K53: oxygen in nitrogen, Metrologia 2010;47:08005

28 Ciarlini P, Cox MG, Pavese F, Regoliosi G. The use of a mixture of probability distributions in temperature interlaboratory comparisons. Metrologia, 2004, 41:116-121.

29 Salit ML, Turk GC. A Drift Correction Procedure. Anal Chem 1998;70:3184-3190.

30 http://en.wikipedia.org/wiki/The_Story_of_the_Three_Bears, accessed 25-May-012

31 Fearn T, Fisher SA, Thompson M, Ellison SLR. A decision theory approach to fitness for purpose in analytical measurement. Analyst 2002;126(6):818-24.

32 Ellison SLR, Barwick VJ, Farrant TJD. Practical Statistics for the Analytical Scientist: A Bench Guide, 2nd Ed. RSC Publishing, Cambridge, UK. 2009.

33 Searle SR, Casella G, McCulloch CE. Variance Components. Wiley-Interscience, Hoboken, NJ, USA. 1992.

34 SAS/STAT 9.2 User's Guide. SAS Institute Inc. Cary, NC USA. 2008.

35 Bates D, Maechler M. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-32. 2009. http://cran.us.r-project.org/web/packages/lme4/

36 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.

37 Lunn DJ, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions (with discussion), Statistics in Medicine 2009;28:3049–3082. See also http://www.mrc-bsu.cam.ac.uk/bugs/ and http://www.openbugs.info/w/

38 Gelman A, Carlin JB, Stern HA, Rubin DB. Bayesian Data Analysis, 2nd Ed. Chapman and Hall/CRC, Boca Raton, FL USA. 2004.

39 ISO. ISO GUIDE 35:2006 Reference materials – General and statistical principles for certification, International Organization for Standardization. Geneva, 2006.

40 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed. Springer, NY, NY USA. 2009. http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print5.pdf

41 Duewer DL, Kowalski BR, Fasching JL. Improving the reliability of factor analysis of chemical data by utilizing the measured analytical uncertainty. Anal Chem 1976;48:2002-2010.

42 Duewer DL. A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers. Accred Qual Assur 2008;13:193–216.

43 Lawn RE, Thompson M, Walker RF. Proficiency Testing in Analytical Chemistry. Royal Society of Chemistry, Cambridge, UK. 1997.

Acronyms and Symbols

Acronyms

ANOVA	analysis of variance
CCQM	Consultative Committee for Amount of Substance - Metrology in Chemistry
CIPM MRA	Comité International des Poids et Mesures Mutual Recognition Arrangement
DI	designated institute
GAWG	Gas Analysis Working Group
GDR	generalized distance regression
GUM	Guide to the Expression of Uncertainty in Measurement
LAI	leave all in
LOO	leave one out
MCMC	Markov chain Monte Carlo
NMI	national metrology institute
OAWG	Organic Analysis Working Group
OLS	ordinary least squares
PBMC	parametric bootstrap Monte Carlo
PSM	primary standard gas mixture
RF	reference function
RM	reference material

Functions

MAX(.)	maximum value in specified set of values
$N(\mu,\sigma^2)$	Gaussian distribution of specified mean and variance
PTILE(p,q)	given percentile of a specified set of values
SGN(.)	signum or "sign" (±1) of a value
$t_s(p,\nu)$	Student's t expansion factor for a specified coverage probability and number of degrees of freedom
$\bigcup$	union of specified sets of values

Meta-symbols

~	when to the left of a mathematical expression, indicates "is distributed as"
ˆ	when above a symbol, indicates an estimated quantity
{}	a defined set of values

Symbols

α	intercept
β	slope
ε_i	residual difference for the i^{th} material
E	residual random error
γ_i	between-unit differences of the i^{th} material
δ_{ij}	within-unit differences of the j^{th} aliquot of the i^{th} material
μ	mean

ρ	correlation
σ	standard deviation
$\sigma_{a,i}$	between-aliquot imprecision for the i^{th} material
$\sigma_{c,i}$	between-campaign imprecision for the i^{th} material, also the confounded between-campaign and between-unit imprecision when one unit is evaluated in each campaign
$\sigma_{r,i}$	repeatability imprecision for the i^{th} material
$\sigma_{u,i}$	between-unit imprecision for the i^{th} material
ν	degrees of freedom
d_i	degree of equivalence for the i^{th} material
D	degree of equivalence for a particular NMI/DI in a KC
$\%d_i$	degree of equivalence for the i^{th} material as a percentage of the value
$\%D$	degree of equivalence for a particular NMI/DI in a KC of the value
i	index over materials
j	index over campaigns and/or units
k	index over replicates or aliquots
l	index over replicates
k_p	coverage factor to provide coverage at a p-percent level of confidence
n	number
n_a	number of aliquots of each unit
n_c	number of measurement campaigns
n_m	number of materials
n_r	number of replicates of each aliquot or unit
n_u	number of units of each material
p	probability
q	PBMC estimates of the quantity Q
Q	generic symbol for a quantity
R	generic representation of “instrument response”
R_i	instrument response for the i^{th} material
R_{ij}	instrument response for the i^{th} material in the j^{th} measurement campaign
u	standard uncertainty
u_∞	“large sample” standard uncertainty
U_p	half-width of the p-% level of confidence expanded uncertainty
$^{-}U_p$	half-width of the 95 % level of confidence expanded uncertainty
$^{+}U_p$	half-width of the 95 % level of confidence expanded uncertainty
V	generic representation of “assigned value”
V_i	assigned value for the i^{th} material
x_i	participant-reported measurement result for a given study material

Figure Captions

Figure 1: Experimental designs for comparisons of RMs

a) One-factor with summary data design for single units of uniform composition; "*i*" indexes the unique materials, "n_c" is the number of measurement campaigns, and the dotted line represents "repeat"
b) One-factor with replicated data design for batch-certified materials of uniform composition; format as above with "Unit" denoting a unique container of the i^{th} RM, "n_u" the number of such units, "R" an independent replicate measurement and "n_r" the number of such measurements.
c) Two-factor nested design for batch-certified materials of potentially non-uniform composition; format as above with "A" denoting independent aliquots of a unique container of the i^{th} RM and "n_a" the number of such aliquots.

Figure 2: RegViz graphical displays of CCQM-P41 CO_2 GDR analysis.

a) Low-resolution overview scatterplot; the assigned values are plotted along the horizontal axis and the repeatability measurement responses along the vertical axis. The diagonal red line represents the RF. The small black crosses represent the $\{V_i \pm U_{95}(V_i), R_i \pm U_{95}(R_i)\}$ for the nine PSMs. The blue symbols, "A" to "I", identify the PSMs as sorted from smallest to largest amount of CO_2.
b) High-resolution thumbnail scatterplot; each thumbnail displays the extent of agreement between the RF and $\{V_i, R_i\}$ for one PSM. All of the thumbnails have the same scale, set by the largest $U_{95}(V_i)$ or $U_{95}(R_i)/\hat{\beta}$ of any material. The red lines represent the RF in relation to each material. The black ellipses enclose about 95 % of the density for the bivariate normal distribution $N(V_i, u_\infty(V_i), R_i, u_\infty(R_i)$. The dots centered in the ellipses mark $\{V_i, R_i\}$.

Figure 3: Key attributes of the thumbnail ellipses.
The diagonal red line represents the RF, the closed blue circle marks $\{V_i, R_i\}$, the open blue circle marks $\{\hat{V}_i, \hat{R}_i\}$, and the diagonal blue line between them represents the uncertainty-weighted distance, ε_i, between the observed values for the material and the RF. The thin blue lines and associated curly braces designate lengths along the *V* and *R* axes. While the figure uses an RF having unit slope to simplify the discussion, the image can be generalized by dividing the "*R*"-related terms by the estimated slope, $\hat{\beta}$.

Figure 4: Comparison of GDR RF models for CCQM-P41 CO_2.

a) RF as in Figure 2b with U_{95}(RF) estimated with "Leave-All-In (LAI)-PBMC; the green lines bounding the RF line represent the empirical 95 % level of confidence interval on the RF.
b) RF as in Figure 2b with U_{95}(RF) estimated with Leave-One-Out (LOO)-PBMC.
c) RF estimated excluding PSM "A" with U_{95}(RF) estimated with LOO-PBMC. The red label and circle in the first thumbnail indicate that the material is excluded from the RF estimation.

Figure 5: Identification of Influential Materials CCQM-P41 CO_2.
Each labeled circle represents the uncertainty-scaled distance for one PSM as estimated for the entire ensemble, $\varepsilon_{i,\mathrm{LAI}}$, (horizontal axis) with that estimated for all PSMs except itself, $\varepsilon_{i,\mathrm{LOO}}$. The bars represent PBMC-estimated 95 % level of confidence intervals on each of the two estimates. The black diagonal line represents equivalence between the two estimates. The red square bounds the -2.45 to +2.45 standard deviations region that is expected to include about 95 % of normally distributed values.

Figure 6: Identification of Consequential Materials for CCQM-P41 CO_2.
Each labeled circle represents the percentage of 1000 PBMC analyses where the presence of the material causes the $\varepsilon_{i,\mathrm{LOO}}$, for at least one other material to become larger than the 2.45 critical distance (a negative consequence, plotted along the horizontal axis) or become less than 2.45 (a positive consequence, plotted along the vertical axis). The bars represent estimated 95 % level of confidence intervals on the two percentages. The red lines mark the 50 % consequential thresholds. Materials whose presence in the data set does not have a very strong negative or positive consequence for any other material are located to the bottom left.

Figure 7: Contrast of two definitions for "degrees of equivalence".
Consider three materials that are uniformly offset from the RF but that have very different measurement uncertainties: $u(R_1) >> u(V_1)$, $u(R_2) = u(V_2)$, and $u(R_3) << u(V_3)$. The assigned values are plotted along the abstract horizontal axis and the repeatability measurement values are plotted along the abstract vertical axis. The red line represents the RF, the blue dots represents the observed measurements, the red dots represent the GDR estimates, the length of each black line connecting the observed and estimated values represents the uncertainty-weighted distance, and the ellipses represent the 95 % level of confidence area about the observed values. The dashed blue line through the blue dots and parallel to the RF is provided to emphasize that the three materials are the same unweighted distance from the RF.

Figure 8: Dot-and-Bar plots of the degrees of equivalence for CCQM-P41 CO_2.
a) LOO Degrees of equivalence estimated for all nine PSMs. The vertical axis displays the degrees of equivalence, d_i; the horizontal axis displays the certified values, V_i, Each open square represents the d for one PSM and its associated vertical bar represent the $U_{95}(d)$ confidence interval. The horizontal red line represents zero bias between the observed and predicted values. The horizontal grey lines are visual guide lines.
b) Degrees of equivalence estimated with PSM A excluded from the GDR model. The excluded material is marked with an open circle. Note that the d for PSM A does not change: a material is always excluded from its own evaluation in a LOO evaluation.

Figure 9: Dot-and-Bar plots of the degrees of equivalence for CCQM-P73 NO.
a) Degrees of equivalence of individual PSMs; format as in Figure 8. All PSMs shown as excluded from the GDR model were not included in the original analysis as being technically suspect or used as controls [25].
b) Degrees of equivalence of the value-assignment capabilities of the participating organizations. The format is similar to that of Figure 8, but the horizontal axis is used to segregate the organizations. Each solid circle represents the D_{abs} for one participant and its associated vertical bar represent the $U_{95}(D_{\mathrm{abs}})$ confidence interval. The open symbols and their vertical bars represent the $d_{\mathrm{abs}i} \pm U_{95}(d_{\mathrm{abs}i})$ for the PSMs value-assigned by that participant.

FIGURE 1

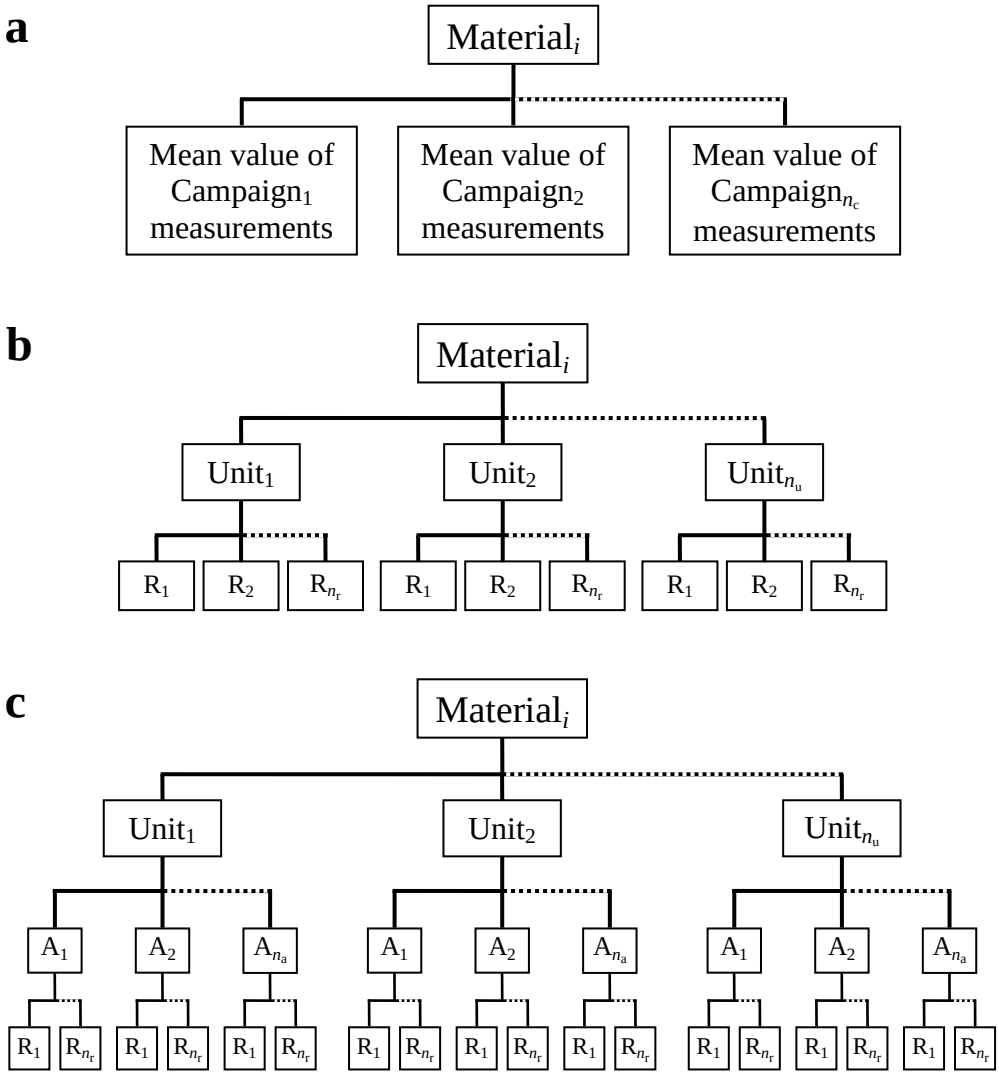

a
Material$_i$
Mean value of Campaign$_1$ measurements
Mean value of Campaign$_2$ measurements
Mean value of Campaign$_{n_c}$ measurements
b
Material$_i$
Unit$_1$
Unit$_2$
Unit$_{n_u}$
R$_1$
R$_2$
R$_{n_r}$
c
Material$_i$
Unit$_1$
Unit$_2$
Unit$_{n_u}$
A$_1$
A$_2$
A$_{n_a}$
R$_1$
R$_{n_r}$

FIGURE 2

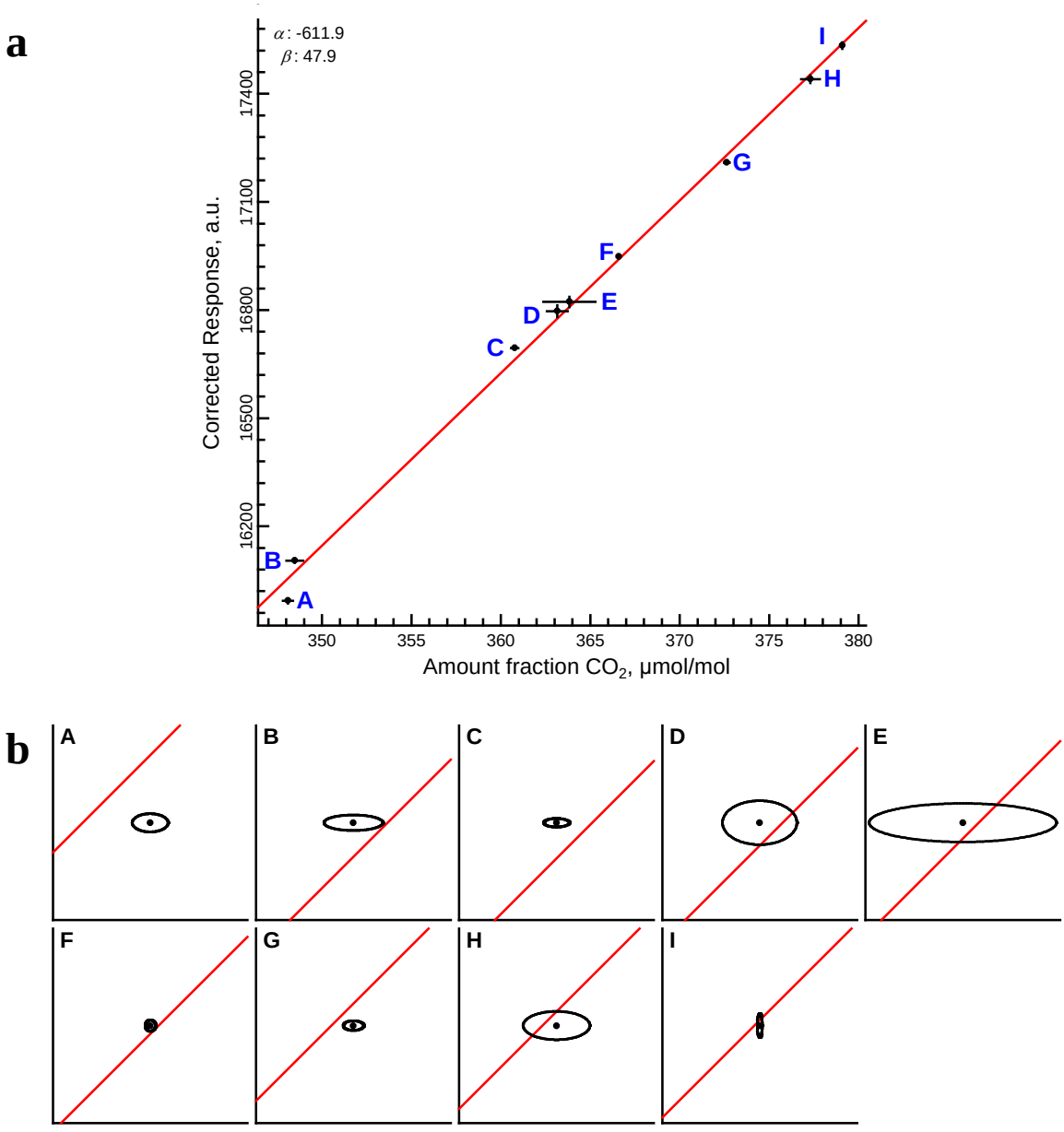

a
α: -611.9
β: 47.9
Corrected Response, a.u.
16200
16500
16800
17100
17400
Amount fraction CO2, µmol/mol
350
355
360
365
370
375
380
A
B
C
D
E
F
G
H
I
b
A
B
C
D
E
F
G
H
I

FIGURE 3

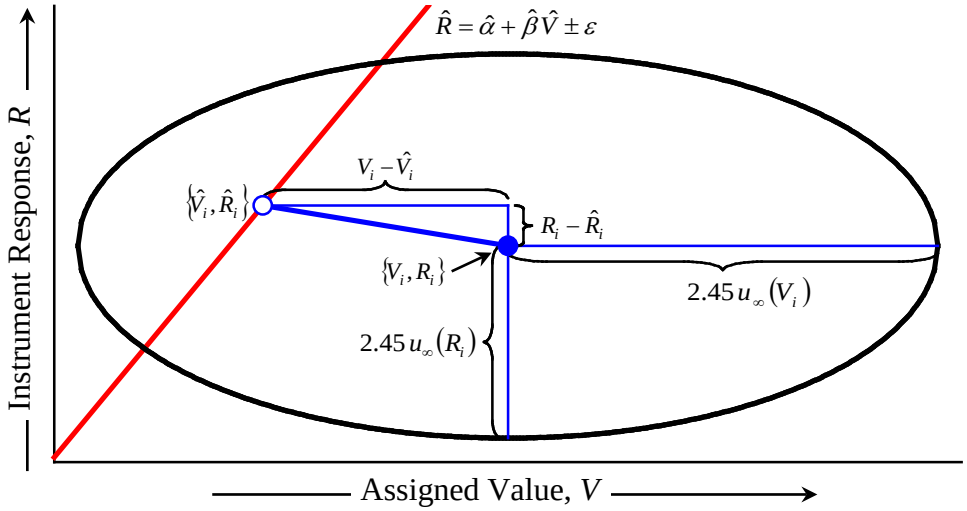

$\hat{R} = \hat{\alpha} + \hat{\beta}\hat{V} \pm \varepsilon$
$V_i - \hat{V}_i$
$\{\hat{V}_i, \hat{R}_i\}$
$R_i - \hat{R}_i$
$\{V_i, R_i\}$
$2.45\,u_\infty(V_i)$
$2.45\,u_\infty(R_i)$
Instrument Response, R
Assigned Value, V

FIGURE 4

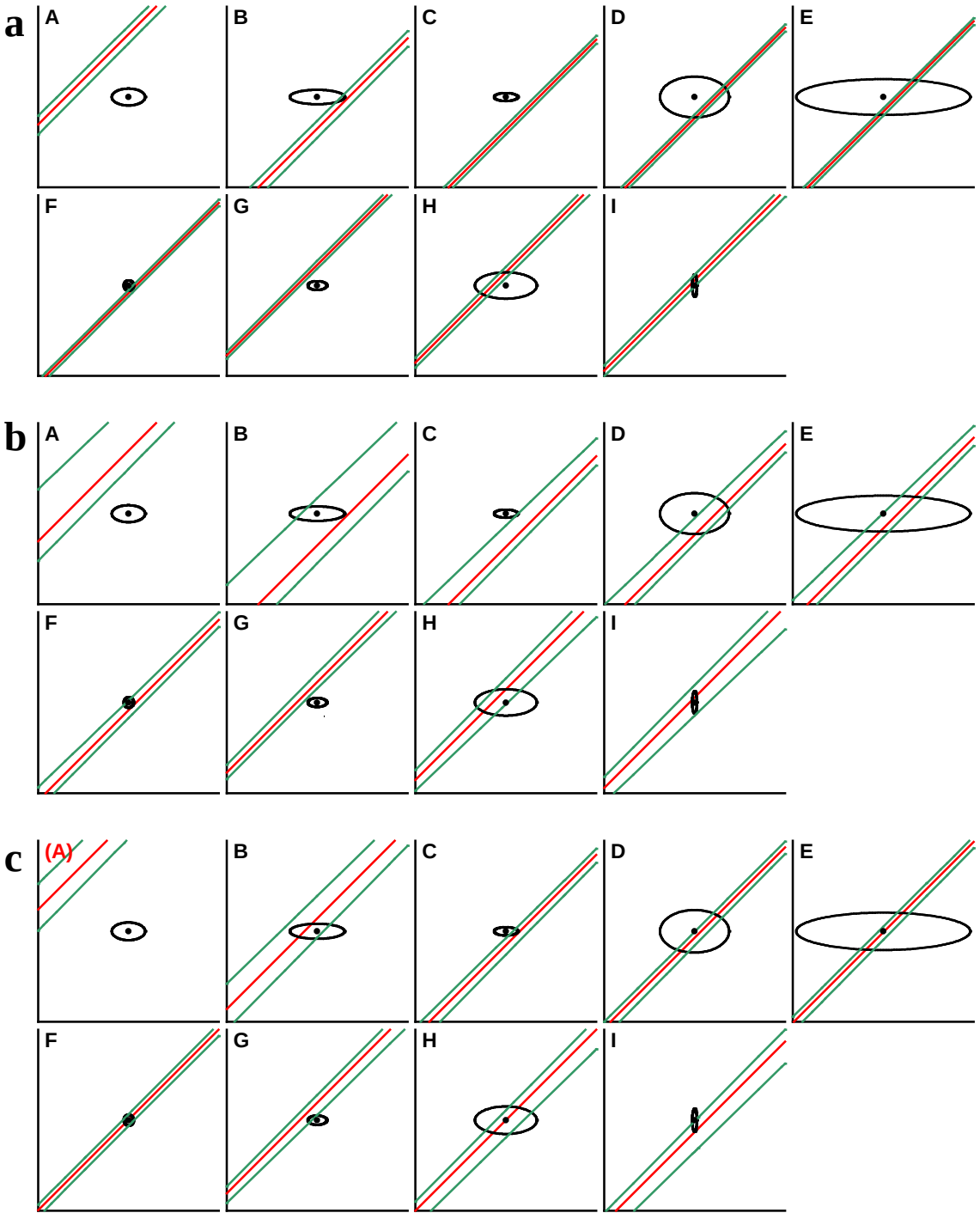


FIGURE 5

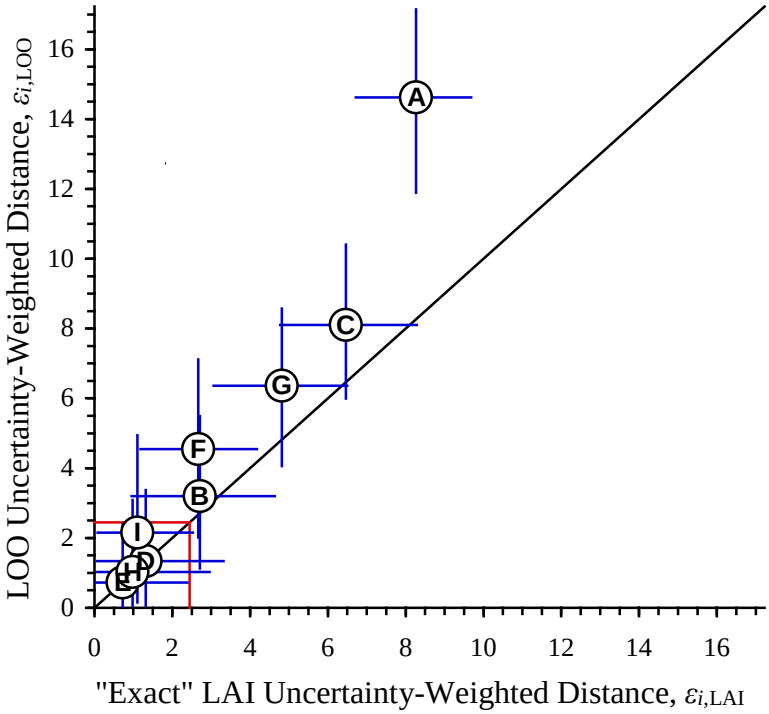

LOO Uncertainty-Weighted Distance, $\varepsilon_{i,LOO}$
"Exact" LAI Uncertainty-Weighted Distance, $\varepsilon_{i,LAI}$
0
2
4
6
8
10
12
14
16
A
B
C
D
E
F
G
H
I

FIGURE 6

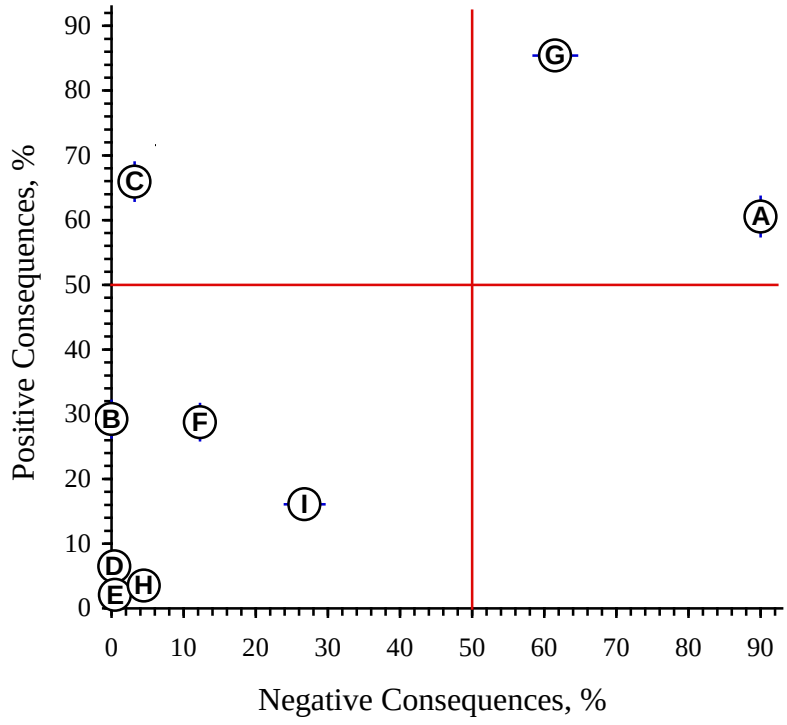

Positive Consequences, %
Negative Consequences, %
0
10
20
30
40
50
60
70
80
90
A
B
C
D
E
F
G
H
I

FIGURE 7

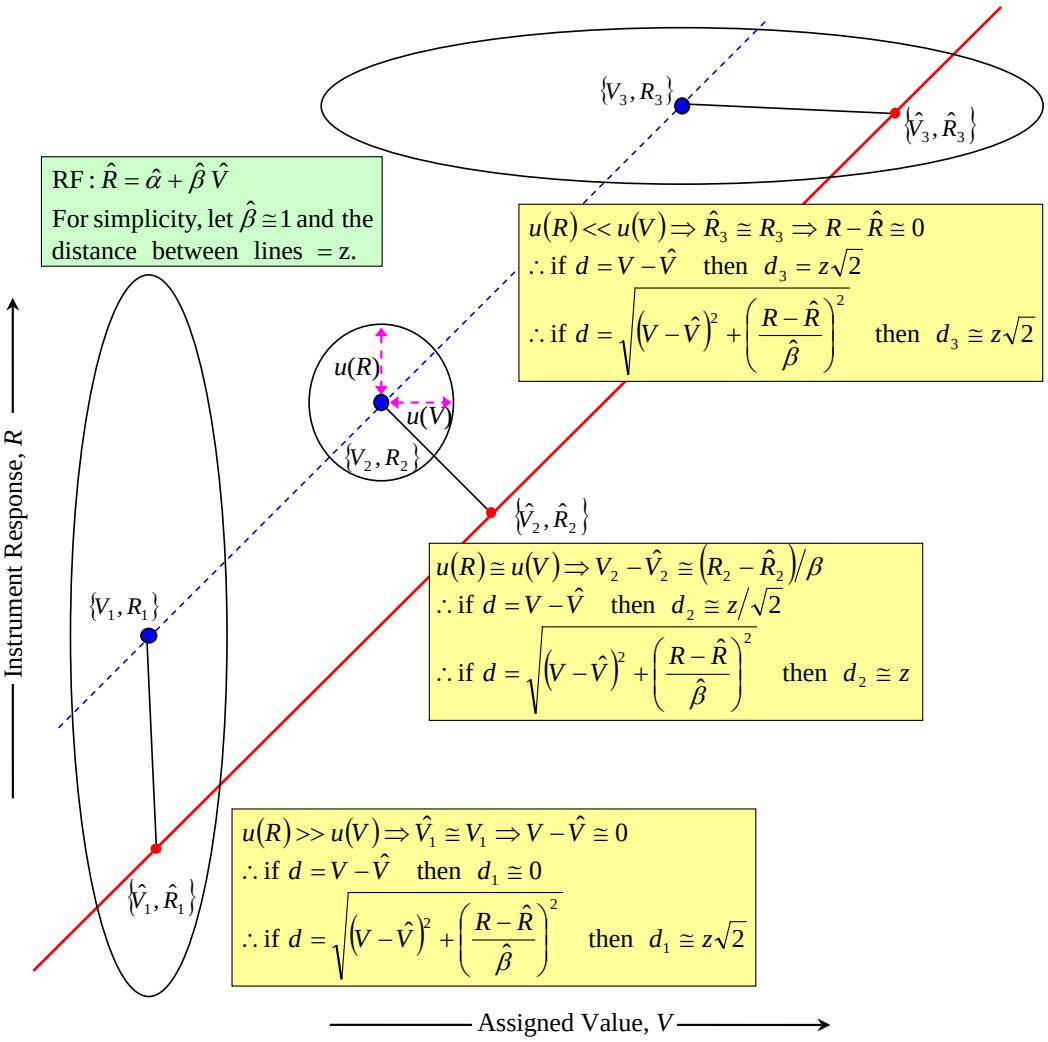

$\{V_3, R_3\}$
$\{\hat{V}_3, \hat{R}_3\}$
RF : $\hat{R} = \hat{\alpha} + \hat{\beta}\,\hat{V}$
For simplicity, let $\hat{\beta} \cong 1$ and the distance between lines = z.
$u(R) << u(V) \Rightarrow \hat{R}_3 \cong R_3 \Rightarrow R - \hat{R} \cong 0$
$\therefore$ if $d = V - \hat{V}$ then $d_3 = z\sqrt{2}$
$\therefore$ if $d = \sqrt{(V-\hat{V})^2 + \left(\frac{R-\hat{R}}{\hat{\beta}}\right)^2}$ then $d_3 \cong z\sqrt{2}$
$u(R)$
$u(V)$
$\{V_2, R_2\}$
$\{\hat{V}_2, \hat{R}_2\}$
$u(R) \cong u(V) \Rightarrow V_2 - \hat{V}_2 \cong (R_2 - \hat{R}_2)/\beta$
$\therefore$ if $d = V - \hat{V}$ then $d_2 \cong z/\sqrt{2}$
$\therefore$ if $d = \sqrt{(V-\hat{V})^2 + \left(\frac{R-\hat{R}}{\hat{\beta}}\right)^2}$ then $d_2 \cong z$
$\{V_1, R_1\}$
$\{\hat{V}_1, \hat{R}_1\}$
$u(R) >> u(V) \Rightarrow \hat{V}_1 \cong V_1 \Rightarrow V - \hat{V} \cong 0$
$\therefore$ if $d = V - \hat{V}$ then $d_1 \cong 0$
$\therefore$ if $d = \sqrt{(V-\hat{V})^2 + \left(\frac{R-\hat{R}}{\hat{\beta}}\right)^2}$ then $d_1 \cong z\sqrt{2}$
Instrument Response, R
Assigned Value, V

FIGURE 8

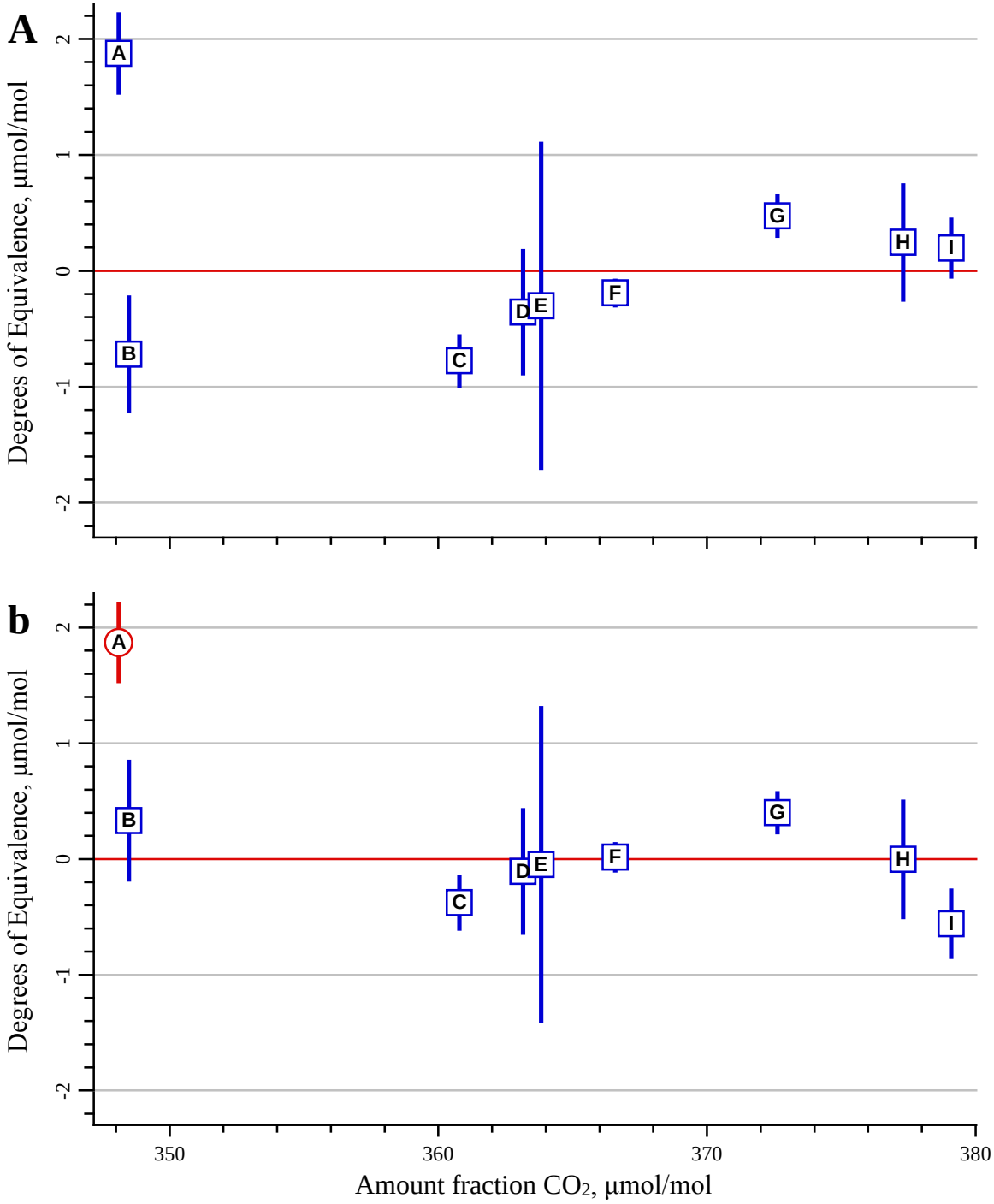
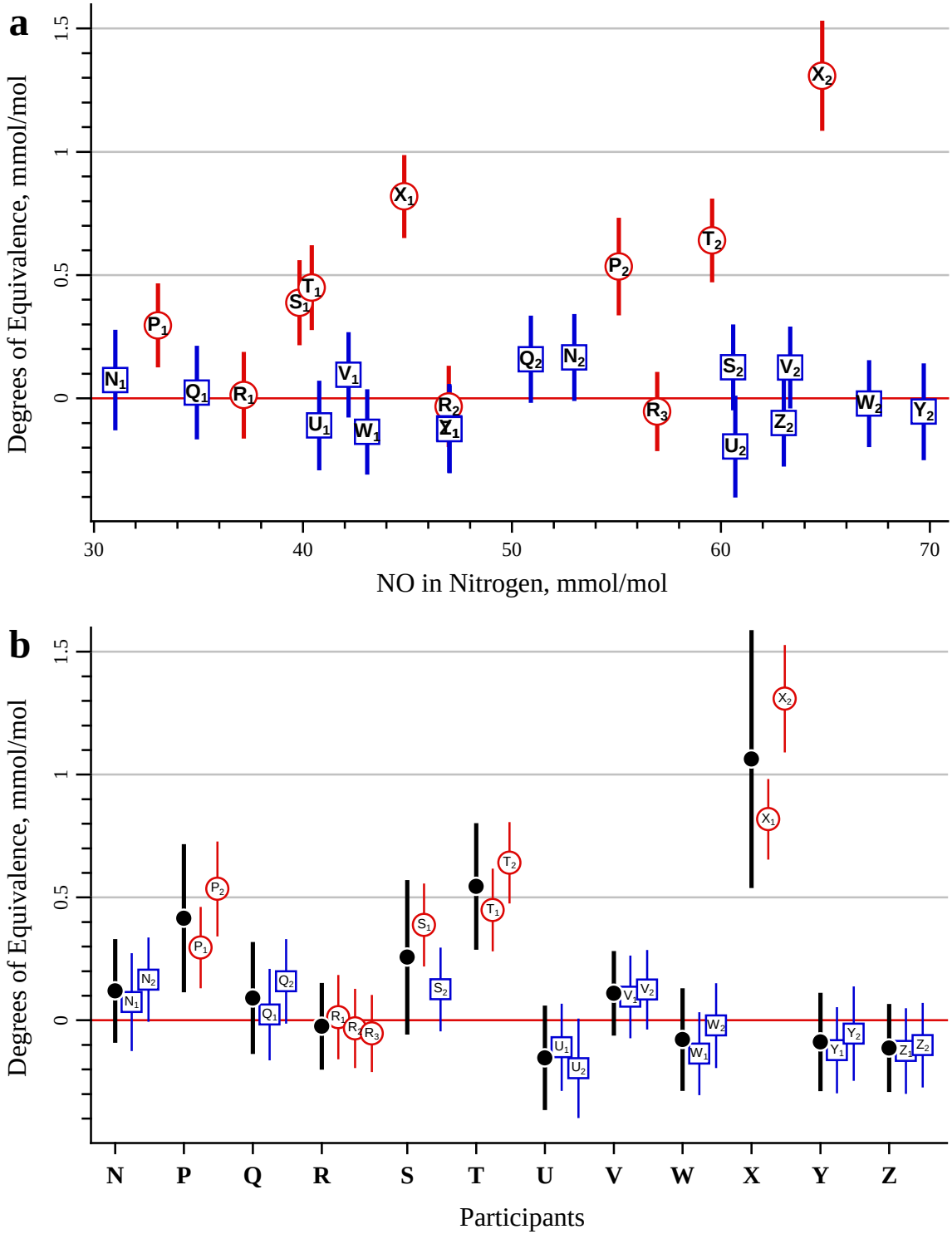

A
Degrees of Equivalence, µmol/mol
2
1
0
-1
-2
A
B
C
D
E
F
G
H
I
b
Degrees of Equivalence, µmol/mol
2
1
0
-1
-2
A
B
C
D
E
F
G
H
I
350
360
370
380
Amount fraction CO2, µmol/mol

FIGURE 9


a
Degrees of Equivalence, mmol/mol
1.5
1
0.5
0
30
40
50
60
70
NO in Nitrogen, mmol/mol
b
Degrees of Equivalence, mmol/mol
1.5
1
0.5
0
N
P
Q
R
S
T
U
V
W
X
Y
Z
Participants