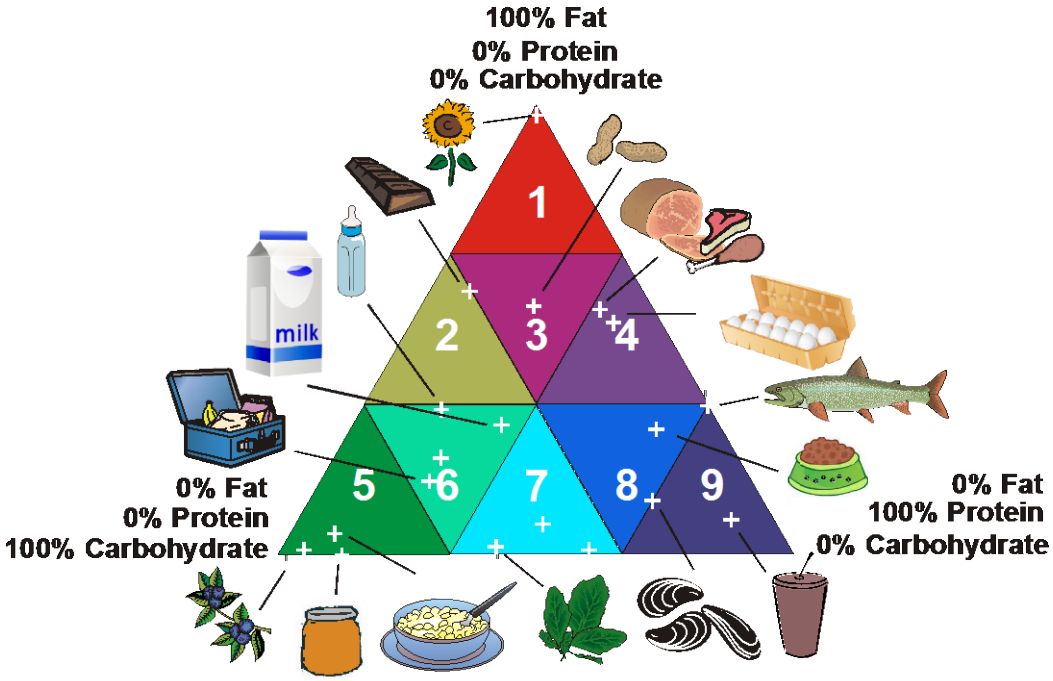
NIST Special Publication 260-181

The ABCs of Using Standard Reference Materials in the Analysis of Foods and Dietary Supplements: A Practical Guide

Katherine E. Sharpless
Katrice A. Lippa
David L. Duewer
Andrew L. Rukhin

http://dx.doi.org/10.6028/NIST.SP.260-181



NIST
National Institute of Standards and Technology
U.S. Department of Commerce

NIST Special Publication 260-181

The ABCs of Using Standard Reference Materials in the Analysis of Foods and Dietary Supplements: A Practical Guide

Katherine E. Sharpless
Katrice A. Lippa
David L. Duewer
Chemical Sciences Division
Materials Measurement Laboratory

Andrew L. Rukhin
Statistical Engineering Division
Information Technology Laboratory

http://dx.doi.org/10.6028/NIST.SP.260-181

June, 2014

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director

National Institute of Standards and Technology Special Publication 260-181.
Natl. Inst. Stand. Technol. Spec. Publ. 260-181, 41 pages (June 2014)
CODEN: NSPUE2

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.SP.260-181

Abstract

Although ‘nutraceuticals’ and ‘functional foods’ seem to be ill-defined terms, both are often used to indicate a food that contains compounds providing benefits beyond basic nutrition, often with the expectation that these compounds are protective against chronic disease. The National Institute of Standards and Technology (NIST) has been producing food-matrix Standard Reference Materials (SRMs) since the mid-1970s. Early materials were characterized solely for elements. Values were assigned for organic constituents of food-matrix SRMs beginning in 1996 and in dietary supplements beginning in 2006. Although none of the NIST food or dietary supplement SRMs were categorized as functional foods *per se*, many have values assigned for components that put the ‘functional’ in functional foods – e.g., antioxidants, phytonutrients, minerals, vitamins, etc. Recommendations for use of these and other natural-matrix SRMs as quality assurance tools are discussed in this paper: from selecting an appropriate material to validating analytical methods, characterizing in-house quality control materials, and establishing traceability.

Keywords

Food, Dietary Supplement
Certified Reference Material (CRM), Standard Reference Material (SRM)
Quality Control (QC), Metrological Traceability

Table of Contents

Tables

Figures

Introduction

Various labeling requirements are imposed on food and dietary supplement manufacturers [1,2], and regulatory agencies and consumers expect product labels to be accurate. In addition to providing accurate product labels, dietary supplement manufacturers are required to follow current good manufacturing practices (cGMPs) [3], including setting product specifications and ensuring that analytical methods are appropriate for assessing whether or not these specifications are met. If a particular material is studied in a clinical trial, investigators must demonstrate that constituents in the test material are the same among batches. Certified reference materials (CRMs) can be used to address these requirements.

An analytical chemist's primary concern in the laboratory is whether or not his/her results are consistently correct. One way to determine whether an analytical method is working properly and generating accurate results is through use of suitable CRMs. CRMs come in two flavors: those intended for use in calibrating measurement systems and those primarily intended to help validate the performance of measurement systems. Figure 1 shows how these two types of materials fit within the complete measurement process, connecting your measurements to the International System of Units (SI).

A calibration CRM is typically either a "neat" material of established high purity or a simple solution gravimetrically prepared from a material of established high purity. Such materials are used to establish the relationship between the quantity of a measurand and the results from a measurement process. (A "measurand" is the "quantity intended to be measured" [4]; for our purposes this refers to the quantity of a given analyte in a particular sample matrix.)

Validation/verification materials are typically preparations of naturally occurring materials of a comparable matrix as the unknown samples that a) contain desired levels of the measurand(s) of interest, b) have been processed to have uniform composition, and c) may have been modified for improved stability. When developing an analytical method, these materials are used to evaluate whether the calibration relationship established for a measurement process is valid when analyzing a particular sample matrix. When implementing a method in a given situation (particular equipment, reagents, analysts, etc.), these materials are used to verify that the developed measurement process indeed provides the expected results.

What your measuring system actually measures may differ from what you intend to measure. However, if the result of your measurement on a suitable natural-matrix CRM agrees with its certified value, it is likely that your measurement process has been suitably characterized and is capable of producing correct results for unknown samples and, ultimately, that the composition of a product can be specified and labeled accurately.

Standard Reference Materials® (SRMs®) are CRMs characterized by the U.S. National Institute of Standards and Technology (NIST). In this paper, we describe the steps that you would take to effectively use SRMs: selecting appropriate materials, establishing traceability, validating and verifying analytical methods, and characterizing in-house quality control materials. While our focus is on the use of our SRMs, the steps are similar for all CRMs regardless of the source [5,6]. Certificates of Analysis on NIST's website [7] (and the websites of other CRM

manufacturers and distributors) as well as other publications provide the information that you will need to evaluate and select an appropriate material [8,9,10].

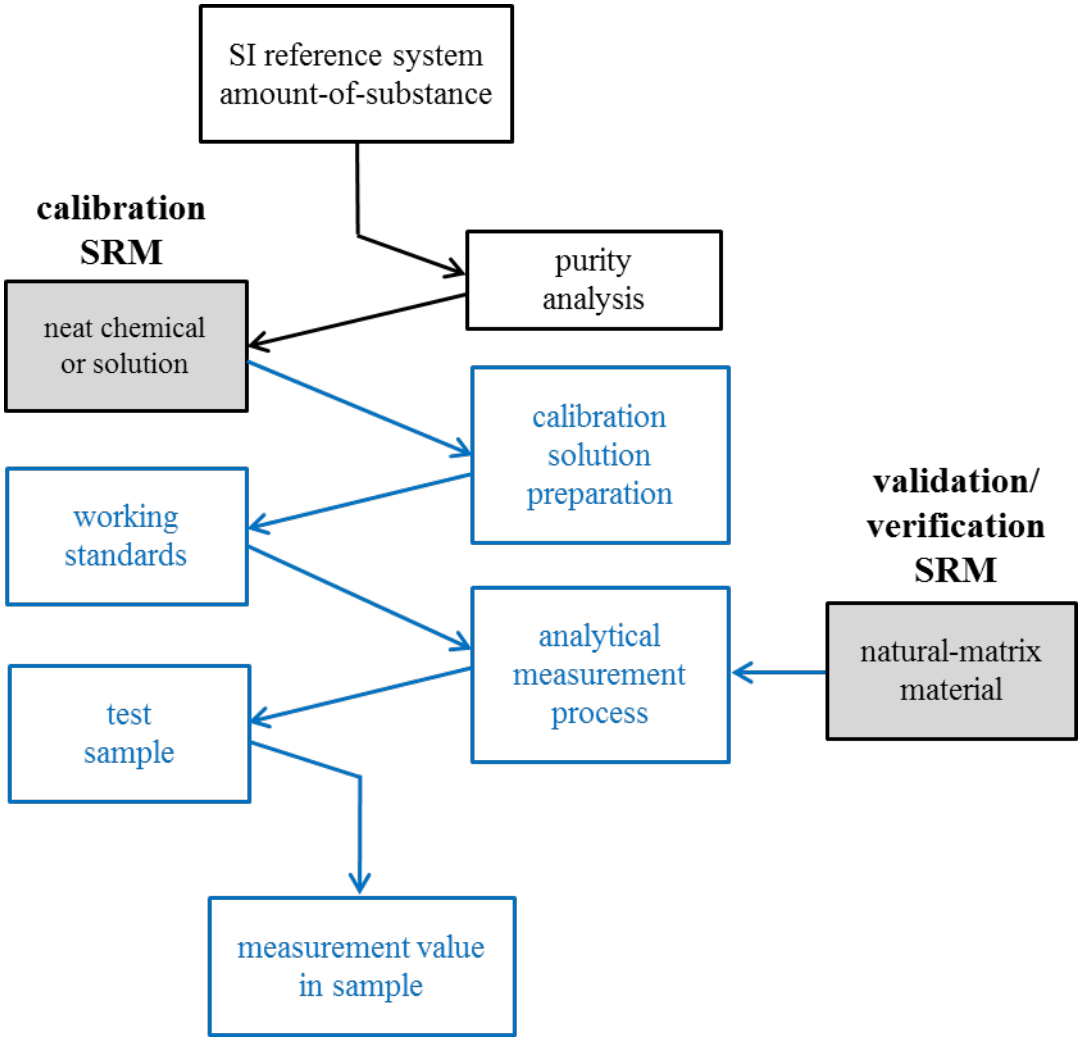


Figure 1. Uses of an SRM

The boxes to the left and right in Figure 1 represent materials, those in the center represent activities, and the arrows indicate the flow of the complete measurement process from the SI reference system to the desired measurement value. The boxes and lines in black denote the materials and activities provided by NIST and other CRM providers. The boxes and lines in blue denote the materials and activities that you must provide.

Step A. Selection of SRMs

Choose an SRM with a matrix that is similar to that of your test samples: any matrix effects or sample preparation issues encountered in one sample should be encountered in the other. For example, if your unknown samples contain a high level of fat, a high-fat SRM would likely be more appropriate for quality assurance than would a high-protein SRM. However, along with a matrix match, measurand quantities should also be matched. Interferences and blank issues may arise if the quantities in the different materials are quite different. Also, results for test samples and the SRM should lie within the same general range on the calibration curve. Often the Official Methods of Analysis of AOAC International (and standard test methods developed by other standards organizations) will specify an analyte range when defining the scope of the method, and this range is also an important consideration.

Next, check that the uncertainty of the SRM's assigned value(s) is fit for your intended purpose [11]. The assigned uncertainty should be small relative to the total uncertainty targeted for your test samples. As a rule of thumb, the SRM's assigned uncertainties should be about one-third or less of the target uncertainty to ensure that uncertainty in the certified value will have negligible influence on the results of your measurements.

To determine the appropriate SRM for analysis with food samples, the AOAC International triangle can be used (Figure A.1). This triangle was developed by AOAC in an effort to promote analytical methods to address nutrition labeling requirements. If a method provides accurate results for one or two foods within a sector, the method is expected to provide accurate results for other foods in that sector. NIST extended this model, expecting that one or two SRMs in a sector should be useful as quality assurance tools when testing other foods within that sector. If the fat, protein, and carbohydrate content of the test sample are not known, approximations can be obtained from the U.S. Department of Agriculture (USDA) nutrient database [12] and used to estimate the test sample's position in the triangle. Once an appropriate natural-matrix SRM is identified, the expected measurand content should be compared to its certified value.

For dietary supplements, where there is not a fat, protein, and carbohydrate composition to consider, the analyst might look at food-matrix SRMs and also at botanical and agricultural materials, environmental matrices (e.g., for measurement of contaminants), and geological materials (e.g., for measurement of elements in solid oral dosage forms) to see whether one of those matrices might have values assigned for the measurand of interest. Again, the optimal SRM will be of a "similar" matrix with a "similar" measurand quantity, but the user must decide what is similar enough based on available knowledge about the materials and on his/her own analytical methods.

As an example, we will consider the analysis of an imaginary product, a 50 gram dark chocolate protein bar. This protein bar contains mass fractions of 15 % protein, 15 % fat, and 50 % carbohydrate, which puts it in sector 6 of the AOAC triangle. An additional consideration is that our product is fortified with fat- and water-soluble vitamins and several minerals. Fortification often results in a single form of the vitamin that largely swamps out any naturally occurring forms. Therefore, it would also be best to have the same forms in the SRM as in our material. Again, the compositional information needed to evaluate the suitability of an SRM is available

on the material's Certificate of Analysis on the NIST website [7]; other CRM providers make available similar information in their own websites and catalogs.

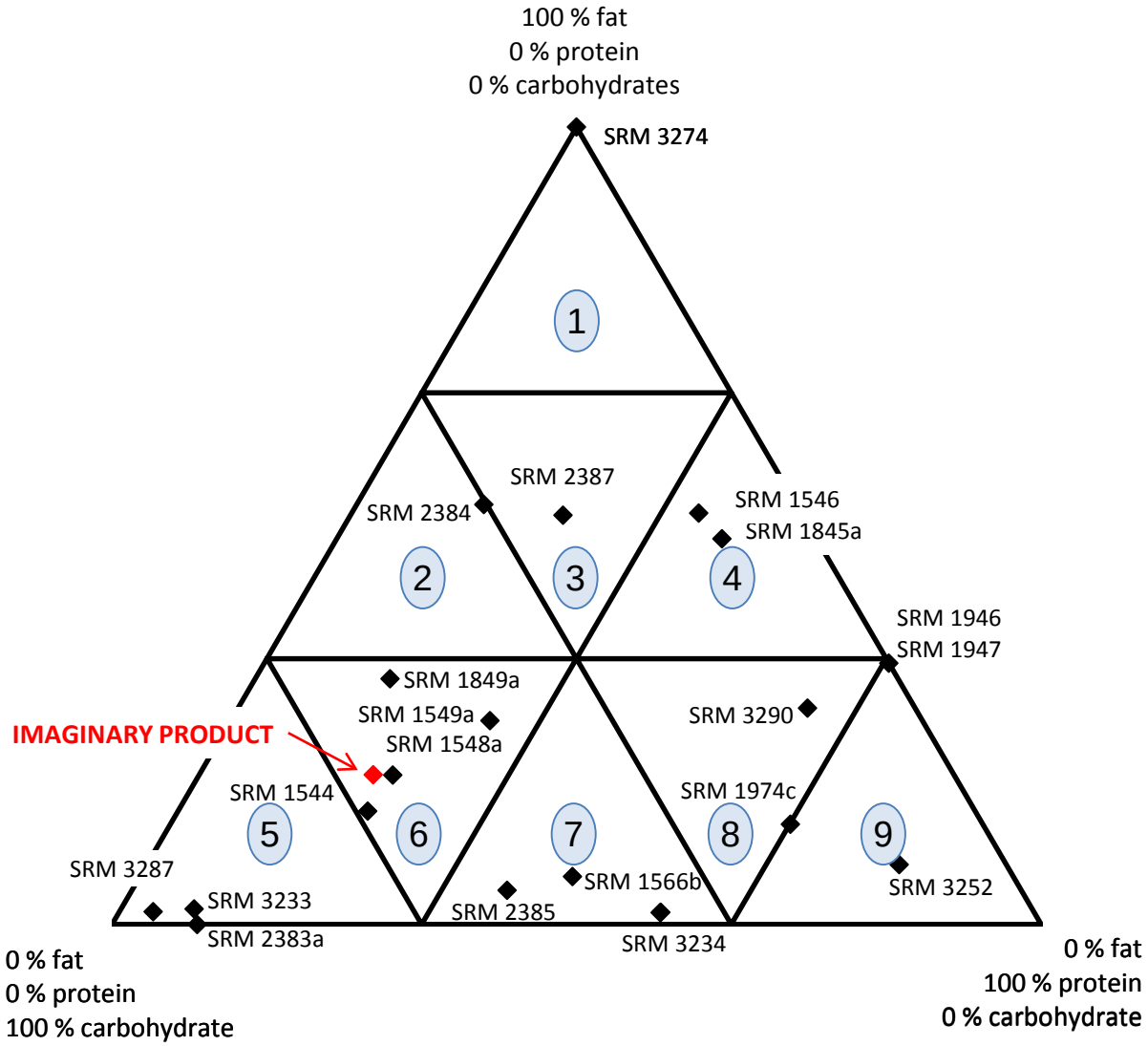


Figure A.1. Location of SRMs in AOAC Food Triangle

The measurands currently required on nutrition labels for processed foods sold in the U.S. include the organic constituents cholesterol and vitamins A (retinol) and C (ascorbic acid), as well as fat, protein, carbohydrate, dietary fiber, and sugar [13]. Since measurement processes for some organics are relatively sensitive to the sample matrix, we look first at the SRMs that are in sector 6 (Table A.1). SRM 1849a Infant/Adult Nutritional Formula appears to be a good choice for use as a quality control material for vitamins and proximates because it 1) has assigned values for the organic measurands of greatest interest and (2) is fortified with the same forms of vitamins as those found in our imaginary protein bar.

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.SP.260-181

Table A.1. Organic measurands of interest in SRMs in sector 6 of the AOAC food triangle.

Material	Cholesterol, mg/kg	Retinol, mg/kg	Ascorbic Acid, mg/kg	Fat, %	Protein, %	Carbohydrates, %	Sugar, %
Anticipated mass fraction in the imaginary 50 g protein bar	**low**	≈3	≈300	**15**	**15**	**50**	**?**
SRM 1544 Fatty Acids and Cholesterol in a Frozen Diet Composite	148.3 ± 9.4	not assigned	not assigned	3.7 ± 0.6	5.3 ± 0.3	16.9 ± 1.5	not assigned
SRM 1548a Typical Diet	not assigned	not assigned	not assigned	19.41 ± 1.45	18.08 ± 0.42	58.36 ± 1.53	not assigned
SRM 1549a Whole Milk Powder	981 ± 71	not assigned	41.9 ±2.5	26.98 ± 0.66	25.64 ± 0.31	38.43 ± 0.95	not assigned
SRM 1849a Infant/Adult Nutritional Formula	137.4 ± 2.9	7.68 ± 0.32	784 ±65	30.4 ± 0.95	13.72 ± 0.92	51.6 ± 1.3	47.6 ± 5.5

U.S. nutrition regulations also require labeling the contents of sodium, calcium, and iron [13]. You expect the protein bar to contain sodium, calcium, and iron mass fractions of no more than 2500 mg/kg, 300 mg/kg, and 15 mg/kg, respectively. You also want to ensure that your product is not contaminated with lead. While we first look at the SRMs in sector 6, measurement processes for the total quantity of elements can be somewhat less sensitive to the sample matrix than those for organic measurands, so we also look at materials in neighboring sectors (Table A.2).

While it's best to have a close match between analyte levels in the validation SRMs and the target levels in the protein bar, SRMs with levels that are within an order-of-magnitude (between 10-fold smaller to 10-fold larger) of the targets are likely be acceptable as long as they are within the scope of your analytical method. Levels that are fairly close to the levels that you anticipate in your protein bar are highlighted in bold red; levels that are within an order of magnitude are identified by footnote "a". The material you identified as the organic control material, SRM 1849a, is a reasonable match for the desired sodium content, although SRM 1549a Whole Milk Powder and SRM 1566b Oyster Tissue are better. However, SRM 1849a has calcium and iron contents that are much greater than the targets and does not have an assigned value for lead. While you can dilute your digests of SRM 1849a to measure calcium and iron, this would also dilute sodium and other matrix elements that might affect the determination of calcium and iron. You decide instead that your quality assurance program will need to use several different materials.

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.SP.260-181

None of the sector 6 materials is a good match for calcium. The best matches are SRM 2383a Baby Food Composite, SRM 3287 Blueberry (Fruit), and possibly SRM 2385 Slurried Spinach. While the sector 6 SRM 1548a Typical Diet would be a reasonable choice for iron (and maybe also sodium, depending on the scope of the method and the range of your calibration curve), both SRM 3287 and SRM 2385 are better matches. To control costs, you wish to minimize the number of SRMs needed. Exploring other information available through the website, you find that SRM 2383a requires refrigeration while SRM 3287 delivers both calcium and iron, can be stored at room temperature (five packets of ≈ 5 g each). While SRMs 2385 is an acceptable choice for calcium and is perhaps the best choice for iron, (four jars of ≈ 70 g each) and requires refrigeration. On balance, you decide that SRM 3287 is the material of choice for calcium and iron, but would not be appropriate for sodium or lead.

Table A.2. Elements of interest in SRMs in sectors 2, 5, 6, and 7 of the AOAC food triangle.

Sector	Material	Sodium, mg/kg	Calcium, mg/kg	Iron, mg/kg	Lead, mg/kg
6	**Anticipated mass fraction in the imaginary 50 g protein bar**	**<2500**	**300**	**15**	**none**
6	SRM 1544 Fatty Acids and Cholesterol in a Frozen Diet Composite	not assigned	not assigned	not assigned	not assigned
6	SRM 1548a Typical Diet	8132[a] ± 942	1967[a] ± 113	**35.3 ± 3.77**	**0.044 ± 0.009**
6	SRM 1549a Whole Milk Powder	**3176 ± 58**	8810 ± 240	1.8[a] ± 0.7	not assigned
6	SRM 1849a Infant/Adult Nutritional Formula	**4265 ± 83**	5253 ± 51	175.6 ± 2.9	not assigned
2	SRM 2384 Baking Chocolate	40 ±2	840[a] ±74	132[a] ±11	not assigned
5	SRM 2383a Baby Food Composite	195 ±29	**324.6 ±5.0**	4.42[a] ±0.51	not assigned
5	SRM 3233 Fortified Breakfast Cereal	6830[a] ±120	36910 ±920	766 ±36	not assigned
5	SRM 3287 Blueberry (Fruit)	16.39 ±0.74	**323 ±16**	**12.20 ±0.74**	not assigned
7	SRM 1566b Oyster Tissue	**3297 ±53**	838[a] ±20	205.8 ±6.8	**0.308 ±0.009**
7	SRM 2385 Slurried Spinach	47 ±1	**624 ±40**	**17.1 ±1.9**	not assigned
7	SRM 3234 Soy Flour	2.52 ±0.45	3191 ±56	80.3[a] ±2.7	not assigned

[a] Values within an order of magnitude of the target analyte levels.

SRM 1548a is the best available match to the low lead content you require for your product, and may be good enough for sodium. SRM 1566b has a higher lead level than you'd want to find in your protein bar, but you could dilute it validate your calibration at several different levels including the very low level you anticipate. Given that the oyster tissue has a reasonable amount of sodium in it, you decide to buy SRM 1566b for your sodium and lead control.

And finally, you want to measure catechins present in the chocolate in your product. You've developed a new method and want to make sure that it works well. You search the NIST SRM website by the keyword "catechin" and discover five SRMs: a baking chocolate; a three-component set of catechin calibration solutions; and green tea as leaves, extract, and solid oral dosage forms (Table A.3). Because you want to make sure that your method extracts the catechins adequately from a matrix, you eliminate the calibration solution from consideration as a control material; but you do decide to select SRM 3257 to calibrate your method. (Note that if you did decide to use SRM 3257 to make sure your method was working properly, you could not also calibrate with it.) Since you're interested specifically in catechins from the chocolate in your protein bar, the baking chocolate SRM is the logical choice, but it doesn't hurt to look at the green tea materials as well. The levels in the green tea are much higher than those that you would expect in your product, where chocolate is merely an ingredient and the catechins in it will be diluted by other ingredients. So you choose SRM 2384 Baking Chocolate for method validation.

Table A.3. SRMs with values assigned for catechins.

Material	Catechin, mg/kg	Epicatechin mg/kg	Catechin monomers mg/kg
SRM 2384 Baking Chocolate	245 ±51	1220 ±240	1490 ±220
SRM 3254 *Camellia sinensis* (Green Tea) Leaves	1010 ±410	9000 ±1600	10000±1700[a]
SRM 3255 *Camellia sinensis* (Green Tea) Extract	9170 ±930	47300 ±6700	56500±6800[a]
SRM 3256 Green Tea-Containing Solid Oral Dosage Form	2630 ±180	12000 ±2600	14600±2600[a]
SRM 3257 Catechin Calibration Solutions	23.54 ±0.53	93.9 ±2.1	117.4±2.2[a]

[a] Calculated from the certified values

Step B. Use of SRMs for Establishing Metrological Traceability

Metrological traceability is the "property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty" [4]. Metrological traceability is different from chain-of-custody traceability: metrological traceability follows a series of measurement comparisons rather than a material passing through a series of possessions or a supply chain. In the case of your protein bar measurements, the traceability chain extends from the SI through the SRMs used to calibrate your measurement systems to the measurements you've made on the protein bar. Since every link in the measurement chain contributes uncertainty, your confidence in your protein bar measurement results will be maximized (that is, the measurement uncertainty will be minimized) by keeping the length of the chain as short as possible.

The most direct method of establishing metrological traceability to the SI is through the calibration of your measurement system with a certified calibration material, such as SRM 3128 Lead (Pb) Standard Solution or SRM 3257 Catechin Calibration Solutions, followed by validation of your measurement process with the appropriate natural-matrix SRM. The calibration solutions have been prepared from materials of established purity using appropriately calibrated balances and instruments under controlled environmental conditions. Neat materials of certified purity, such as SRM 911c Cholesterol, may be available in addition to or instead of pre-made solutions. In either case, assuming that you a) use suitably calibrated equipment under suitably controlled conditions to prepare whatever working standards are needed to calibrate your measurement systems, b) suitably document your use of those procedures, and c) correctly calculate the uncertainties, the measurements made with those systems can be made traceable to the certified value of the calibrant and through it to the SI units of mass (kilogram) or mole and, if required, volume (cubic meter).

When neither certified calibration solutions nor neat materials are available, traceability can be established by calibrating with a natural-matrix material. However, by definition, the traceability chain for natural-matrix materials has at least one more link than that of a calibration material and so will not provide the smallest possible measurement uncertainty. In the case of catechin, the relative uncertainty in the certified value in the baking chocolate (≈ 21 %) is almost an order of magnitude greater than that in the calibration solution (≈ 2.3 %). Further, calibrating with a natural-matrix material may bias your results unless your measurement process is very similar to the processes used to certify the SRM. Should you calibrate to SRM 2384 Baking Chocolate you are assuming that extracting catechins from the protein bar and from the baking chocolate have similar efficiencies and that the chromatography of the extracts has similar selectivity.

You can establish direct traceability to SI units by evaluating the purity of a non-certified neat material yourself. This can be an expensive and time-consuming process [14], but it is routinely accomplished by sophisticated analytical laboratories. Traceability would then be linked to the SI through the procedures you used to establish purity in addition to those used to prepare the working solutions.

For additional discussion of traceability, see: [15,16].

Step C. Use of SRMs for Precision Assessment

Depending on how you design your measurement program, repeated measurements of an analyte (call it measurand *X*) can be used to characterize different sources of variability (imprecision) in your analytical method for *X*. These measurements do not have to be made on SRMs but rather on any homogenous, stable material available in sufficient quantity having an appropriate matrix and analyte content.

Multiple analyses of a single preparation of a sample reflect the imprecision of the post-sample preparation aspects of your method. The post-preparation standard deviation, s_{post}, is readily estimated from the generic formula

$$s(x) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

where $s(x)$ is the estimated standard deviation of the measurements of *X*, each x_i is a measured value, *n* is the number of such values, and $\bar{x}$ is the mean (also known as the arithmetic average) of the values. Note: Appendix I lists and defines all of the symbols used in this document.

When the x_i are merely repeated measurements of a single preparation of the sample made by one analyst over a short period of time, $s_{post} = s(x)$. Since the precision uncertainty components are combined in quadrature (i.e., the square root of the sum of the squares), then as long as s_{post} is less than about 0.3 times the total target imprecision (i.e., standard deviation), your post-preparation measurement process should contribute less than about 10 % of your targeted uncertainty. If the factor is more than about 0.3, you should try to improve the instrumental analysis aspects of your method – including confirming that the method's detection limit is adequate for your purpose.

If you analyze independently prepared subsamples of a material over a short period of time, you can estimate your method's repeatability imprecision. The repeatability standard deviation, s_r, can be estimated as above when the x_i represent single measurements of *n* independently prepared subsamples. As above, if s_r is less than about 0.5 times the targeted imprecision, the combined sample preparation and post-preparation processes should contribute less than about 30 % of your targeted uncertainty. If the factor is more than about 0.5, you should try to improve the sample preparation aspects of your method.

If you and several other analysts analyze independently prepared subsamples or you analyze your independently prepared subsamples over a sufficiently long period of time, you can estimate some form of what is termed "intermediate" imprecision, s_I. If the variability introduced by combining data from multiple analysts and/or collecting data over a long period of time is not within the target uncertainty, the method must be made more robust to small changes, the contributing influence factors that give rise to the variability must be identified and controlled, and/or the analysts must be provided with additional training. However, it's best to estimate s_I using one-factor analysis of variance (ANOVA) – described in most statistics texts and guides to laboratory statistics: see, for example, [17,18,19,20]. While the calculations are

fairly simple, they are much more efficiently (and reliably) performed using appropriate software.

Table C.1 lists (dry-lab) "results" for total catechin monomers in SRM 2384 of measurements performed by five analysts, along with basic summary statistics for each analyst. Let's say that all five groups of data were obtained using nominally the same analytical process, equipment, and supplies over roughly the same time period and that each value is the result for an independently prepared subsample. Table C.2 shows typical summary results from a one-factor ANOVA of these data. For our current purpose, only the number of measurements per group, n_j, and the within- and between-groups mean-squares, MS_{wth} and MS_{btw}, are of interest.

Table C.1. Catechin monomer validation "data."
Mass fractions in mg/g.

	Analyst				
Repeat	KS	DD	BN	KL	AR
1	1.362	1.494	1.389	1.460	1.357
2	1.388	1.504	1.415	1.530	1.453
3	1.392	1.647	1.424	1.560	1.510
4	1.412	1.650	1.441	1.580	1.573
5	1.415	1.692	1.455		
6	1.419	1.698	1.472		
7	1.426	1.741	1.482		
8	1.458	1.770	1.483		
9	1.467		1.519		
10	1.511		1.588		
n	10	8	10	4	4
Mean	1.425	1.649	1.467	1.533	1.473
s	0.044	0.102	0.057	0.053	0.092

Table C.2. One-way ANOVA for catechin monomer "data."
Mass fractions in mg/g

					One-Factor ANOVA					
Group	n_j	Sum	Mean	Variance	*Component*	*SumSq*	*df*	*MS*	*F*	*P-value*
KS	10	14.252	1.425	0.001896	Between	0.2522	4	0.0631	12.85	2.8E-06
DD	8	13.195	1.649	0.010329	Within	0.1522	31	0.0049		
BN	10	14.668	1.467	0.003258	Total	0.4044	35			
KL	4	6.130	1.533	0.002758						
AR	4	5.893	1.473	0.008408						

The repeatability standard deviation is just the square-root of MS_{wth}

$$s_r = \sqrt{MS_{\text{wth}}} = \sqrt{0.0049} = 0.070 \text{ mg/g.}$$

The between-analyst and/or between-chunk-of-chocolate intermediate imprecision is

$$s_{\mathrm{I}} = \sqrt{\frac{MS_{\mathrm{btw}} - MS_{\mathrm{wth}}}{n'}}$$

$$n' = \frac{N^2 - \sum_j^m n_j^2}{(m-1)N}$$

$$N = \sum_j^m n_j$$

where n' is the effective number of independent measurements per group, m is the number of groups, and N is the total number of measurements. If all analysts had made the same number of measurements (a "balanced design") then n' would just equal that number. In this example (an "unbalanced design" more typical of routinely collected data)

$$n' = \frac{36^2 - 296}{(5-1)36} \cong 6.94$$

and so

$$s_{\mathrm{I}} \cong \sqrt{\frac{0.0631 - 0.0049}{6.94}} = 0.092 \text{ mg/g.}$$

With the data at hand, you can't tell whether differences among the analysts or heterogeneity in the SRM unit is the major source of s_{I} imprecision, but this could be addressed by having analysts KS and DD analyze subsamples of the same chunk of the sample material. (You know that KS and DD are the analysts who should make additional measurements because of the five analysts the mean of KS's measurements is the smallest while the mean of DD's measurements is the largest.) Regardless, the expected combined standard deviation for the measurement process under the stated conditions of measurement combines the repeatability and intermediate components

$$s(x) = \sqrt{s_r^2 + s_I^2} = \sqrt{0.070^2 + 0.092^2} \cong 0.12 \text{ mg/g.}$$

If the within-group mean-square is greater than the between-group, $MS_{\mathrm{wth}} > MS_{\mathrm{btw}}$, then there is no significant intermediate imprecision and $s(x)$ can be taken to be s_{r}.

Note: you can characterize more than one source of additional variability in a single data set given enough (good) data (e.g., several analysts each reporting data for the same control materials using a number of different extraction protocols and/or instruments multiple times a week for many weeks) and appropriate data analysis tools. However, the design and evaluation of such studies is well beyond the scope of this document. Consult with your favorite statistical consultant for further information.

Step D. Use of SRMs to Establish Trueness of Results

This section contains a lot of detailed explanation on interpreting your results relative to the certified value because these are the types of questions that we are most frequently asked when a customer's results don't seem to "agree" with the assigned value.

Different CRM suppliers express certified values and their uncertainties somewhat differently. For NIST natural-matrix SRMs characterized for chemical composition, the certified value of measurand *X* in the sample matrix is typically stated as $x_{\mathrm{NIST}} \pm U_{95}(x_{\mathrm{NIST}})$, where $U_{95}(x_{\mathrm{NIST}})$ is an expanded uncertainty such that the interval from $x_{\mathrm{NIST}} - U_{95}(x_{\mathrm{NIST}})$ to $x_{\mathrm{NIST}} + U_{95}(x_{\mathrm{NIST}})$ is expected to contain the true value of *X* in all units of the SRM with about a 95 % level of confidence. Note that this statement does not guarantee that the true value of *X* in *your* unit of the SRM is x_{NIST}, only that its true value is expected with high confidence to be somewhere within the certified interval. Conversely, there is a small probability (about 5 %) that the true value for your unit is somewhere outside the range.

While NIST's natural-matrix SRMs are intended to be homogeneous and relatively stable, the vast majority of these SRMs are batch-certified, and some degree of within-unit and between-unit heterogeneity is inevitable. The Certificate of Analysis for these SRMs will indicate the minimum sample mass that you should use for your analyses. This information is generally under the "Instructions for Use" section of the Certificate and is stated either as: a) the minimum sample mass for the certified value to be valid or b) the sample size at which homogeneity was assessed. The certified intervals take into account measurand heterogeneity at this sample size as well as the imprecision and between-method bias (systematic differences or "lack of trueness") of the measurement processes used to assign the certified value.

If you could make very many (say, 30 or more) measurements of independently prepared subsamples of a particular unit of a given SRM over a long period of time using an unbiased analytical method, then the mean of all your measurements, $\bar{x}_{\mathrm{lab}}$, should be within the $x_{\mathrm{NIST}} \pm U_{95}(x_{\mathrm{NIST}})$ interval. However, it is not practical for any SRM user to make such a large number of measurements: laboratory resources and the amount of material in each unit of the SRM are both limited. You are more likely to make a few measurements (say, three to five) and base your assessment of your method's bias on those. But there is a tension between the two possible conclusions: you don't want to conclude that a method is unbiased when it is in truth biased nor conclude that it is biased when in truth it is unbiased. Unfortunately, for any given number of measurements there is no way to minimize both risks at the same time; it is up to you to decide the relative importance of the two risks [19].

And always keep in mind that natural-matrix CRMs generally provide only "yellow light" (cautionary) validation of the fitness-for-purpose of your measurement procedure for real samples. While failure to agree with the certified value is a "red light" signal that your measurement procedure may not be fit for *any* similar sample, agreement is not a "green light" that all is well for *every* sample that your laboratory may need to analyze but only that it is fit for *samples having analyte content and matrix similar to the SRM*. For such materials, proceed, but with due caution.

D.1 Comparing your results to an SRM.

The comparison of your results to a certified value is generally known as “compliance assessment” but is more formally the determination of the “metrological compatibility of measurement results” [4]. The various aspects of assessing the compliance of chemical measurement procedures as well as practical assessment tools are discussed in the Eurachem/CITAC Guide "Use of uncertainty information in compliance assessment" [21] and the fundamental statistical limitations of these approaches have recently been discussed [22]. Appendix II outlines the philosophy for formal statistical assessment of compliance and provides detailed instructions for using sophisticated free computer programs for evaluating the examples, and Appendix III describes how to install and use the freeware software required to run the programs.

However, it is often possible to assess whether a measurement procedure is "true enough" using very simple data analysis methods. In our protein bar example, say you’re measuring total catechin monomers (the sum of catechin and epicatechin). If your analytical method gives you the “right” answer for catechins in SRM 2384 Baking Chocolate, you will have confidence in your ability to reliably measure catechins in your product samples. Figure D.1 displays (dry lab) “results” for seven measurement processes, each set of results consisting of three independent measurements of total catechin monomers in single units of SRM 2384. The individual data and the summary statistics for the seven methods are provided in Appendix IV. The results are labeled “A” through “G.” Your results are likely to be similar to one of these examples.

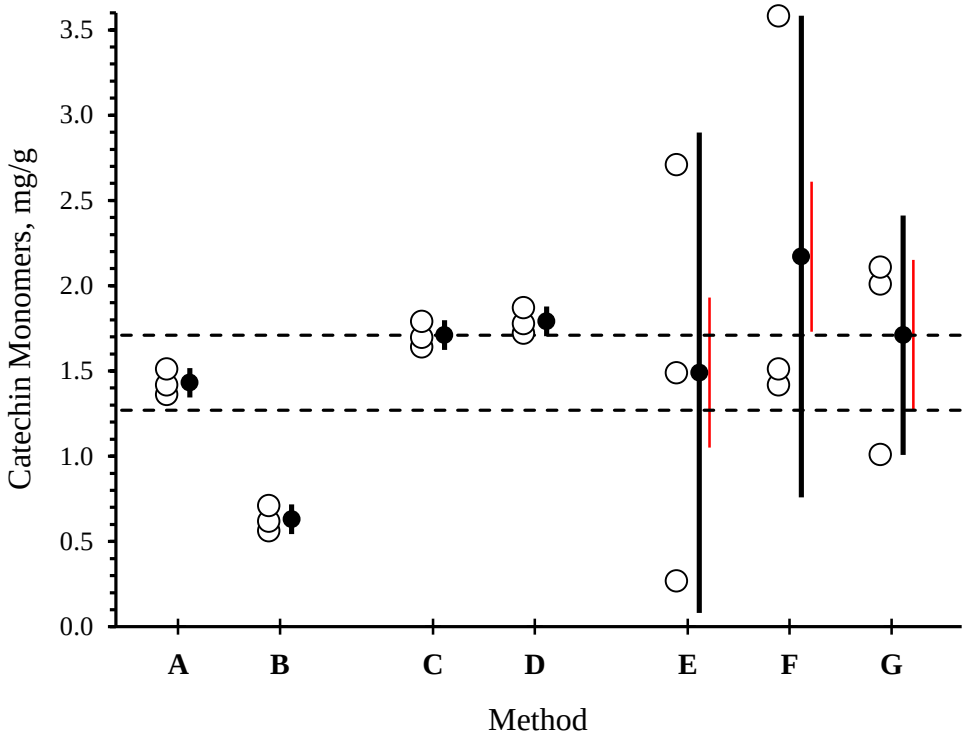


Figure D.1. Graphical Evaluation of Method Trueness

The horizontal dashed lines in Figure D.1 bound the NIST-certified $x_{\mathrm{NIST}} \pm U_{95}(x_{\mathrm{NIST}})$ interval of 1.49 mg/g ± 0.22 mg/g. The open circles represent individual independent measurement results, $x_{\mathrm{lab},i}$. The solid circles represent the arithmetic mean of your measurements,

$$\bar{x}_{\mathrm{lab}} = \left(\sum_{i=1}^{n} x_{\mathrm{lab},i}\right)/n,$$

where *n* is the number of your measurements. The thick vertical lines through those circles represent approximate 95 % level of confidence intervals about your mean

$$U_{95}(\bar{x}_{\mathrm{lab}}) = k \cdot s(\bar{x}_{\mathrm{lab}})$$

$$s(\bar{x}_{\mathrm{lab}}) = s(x)/\sqrt{n}$$

$$s(x) = \sqrt{\sum_{i=1}^{n} \frac{\left(x_{\mathrm{lab},i} - \bar{x}_{\mathrm{lab}}\right)^2}{(n-1)}}$$

where $s(\bar{x}_{\mathrm{lab}})$ is the standard deviation of the mean (also known as the standard error of the mean), *s*(*x*) is the standard deviation of your measurements, and *k* is an appropriate expansion factor. When the number of independent measurements is not large, say less than 10, then *k* should be estimated using Student's *t* with the appropriate number of degrees of freedom. When the number of measurements is large or you have confidence in the estimate based on other evidence, it is typically asserted that you have "large" effective degrees of freedom and the value for *k* is 2 [23]. For convenience, Figure D.1 uses *k* = 2. The thin red vertical lines to the immediate right of the solid circles for methods E through G represent the approximate 95 % level of confidence intervals after a number of additional measurements have been made.

Without any further analysis, it can be safely concluded that method A is unbiased and that method B is biased. The independent measurement results within each of these methods are very self-consistent, with a $U_{95}(\bar{x}_{\mathrm{lab}})$ of 0.087 mg/g that is much less than the 0.22 mg/g of $U_{95}(x_{\mathrm{NIST}})$. For method A, the entire $\bar{x}_{\mathrm{lab}} \pm U_{95}(\bar{x}_{\mathrm{lab}})$ interval is well inside the $x_{\mathrm{NIST}} \pm U_{95}(x_{\mathrm{NIST}})$ certified interval, indicating that results from this method are reliably unbiased. For method B, the entire $\bar{x}_{\mathrm{you}} \pm U_{95}(\bar{x}_{\mathrm{lab}})$ interval is well outside the certified interval, indicating that results from this method, while self-consistent, are consistently biased. Additional measurements would be unlikely to change these conclusions.

Methods C and D also produce reliably consistent results and have $U_{95}(\bar{x}_{\mathrm{lab}})$ of 0.087 mg/g. The $\bar{x}_{\mathrm{lab}}$ of the method C results is just at the upper edge of the certified interval; assuming the results are roughly normally distributed, there is at least a 50 % probability that the "very many ... long term" mean is actually within the certified interval and the method can be considered unbiased. For method D, it is the lower limit, $\bar{x}_{\mathrm{lab}} - U_{95}(\bar{x}_{\mathrm{lab}})$ that is aligned with the upper edge of the certified interval and thus there is only about a 2.5 % chance that more measurements would yield an $\bar{x}_{\mathrm{lab}}$ that is within the certified interval. Therefore it is rather likely that method D provides results that are somewhat biased (although whether that bias is significant depends on your needs.)

But by NIST's definition of the certified interval, the true value of *X* in about 2.5 % of the SRM units may be equal to or a bit greater than the upper dotted line. If in this situation, a rigorous bias evaluation is required and more measurements are to be made, they would best be made on a different bar of the chocolate SRM to better identify whether the slight bias reflects the intrinsic

characteristics of the measurement processes or the true value of *X* in the particular bar that was analyzed. (Many food and dietary supplement SRMs contain several packets or bottles per SRM unit; analyzing another packet does not necessarily mean that you must buy another unit of the SRM.)

In contrast to methods A through D, the relatively large intervals for methods E, F, and G are attributable to widely dispersed individual results. Of course, a pragmatic first step would be to improve the measurement process to reduce the overall variability. However, for this illustrative case, more measurements on the *same* SRM unit will improve the confidence in the trueness assessment for methods E, F, and G. The $\bar{x}_{\text{lab}} \pm U_{95}(\bar{x}_{\text{lab}})$ intervals for all three methods include all of the $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}})$ interval. Thus the available evidence is that the methods could be unbiased; however, the intervals are too large to assert that conclusion with much confidence. As identified in [22], a meaningful bias assessment requires that the $\bar{x}_{\text{lab}} \pm U_{95}(\bar{x}_{\text{lab}})$ interval shouldn't be much wider than the $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}})$ interval. How can that be achieved?

Assuming that the results for a measurement procedure are more or less symmetrically distributed around the true value of the quantity of *X* in your packet of the SRM, as indicated by the thin red lines in Figure D.1, your knowledge of this true value will improve as additional independent measurements are made. The rate of improvement is proportional to the square root of the number of measurements, $\sqrt{n}$. The next sections discuss how to estimate the minimum number of measurements needed.

D.2 What n is needed? (Estimate from literature or your own historical data).

If you have a reliable estimate of the expected measurement standard deviation, *s*(*x*), for given amounts of analyte, then estimating the *n* needed to get an appropriately small $U_{95}(\bar{x}_{\text{lab}})$ is fairly straightforward. You can estimate generic values for *s*(*x*) from historical data on replicate analyses of unknowns and/or quality control materials [24]. If your measurement method is the same or very similar to one characterized in the literature, then a usable estimate of *s*(*x*) may be available from robustness evaluations or interlaboratory reproducibility studies [25]. Standards Development Organizations such as AOAC International generate such estimates in the process of evaluating their Official Methods of Analysis. These estimates are often given in the form of relative standard deviations expressed as percent, $CV = 100 \cdot s(x)/\bar{x}$, and would then need to be converted back into the standard deviations for each particular *x*: $s(x) = \bar{x} \cdot CV/100$.

Using the method C example, assume that about half of the $\bar{x}_{\text{lab}} \pm U_{95}(\bar{x}_{\text{lab}})$ interval must fit within $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}})$; that is $U_{95}(\bar{x}_{\text{lab}})/2 \leq U_{95}(x_{\text{NIST}})$. Since we are assuming here that $U_{95}(\bar{x}_{\text{lab}})$ has been estimated from a confident estimate of $s(\bar{x}_{\text{lab}})$,

$$s(\bar{x}_{\text{lab}}) = U_{95}(\bar{x}_{\text{lab}})/2$$

and so

$$s(\bar{x}_{\text{lab}}) \leq U_{95}(x_{\text{NIST}}).$$

Given $s(\bar{x}_{\text{lab}}) = s(x)/\sqrt{n}$, then

$$\sqrt{n} = s(x)/s(\bar{x}_{\text{lab}}) \approx s(x)/U_{95}(x_{\text{NIST}})$$

and thus

$$n \approx (s(x)/U_{95}(x_{\text{NIST}}))^2 .$$

The $s(x)$ for methods E and F is 1.22 mg/g; assuming that this value is actually known with high confidence, then the *n* required to confidently assess bias is at least $(1.22/0.22)^2 = 5.6^2 \approx 31$. The thin red vertical line to the right of the $\bar{x}_{\text{lab}}$ for the three methods in Figure 2 represents the length of the resulting $\pm U_{95}(\bar{x}_{\text{lab}})$ intervals from a total of 31 measurements. With the additional measurements, method E is seen to be unbiased and method F to be biased.

For most situations, this high number of independent measurements will not be feasible. It is thus questionable whether assessing the bias of very imprecise methods is a useful exercise. However, the extremely high result for method F (3.58) could well reflect some one-off analytical glitch in the procedure rather than intrinsic variability of the method. If the root cause of the extreme value could be determined (and the procedure corrected), this result could be excluded from the estimate of $\bar{x}_{\text{lab}}$ and the method assessed as unbiased – although another measurement or two would be needed to make that conclusion convincing. Thus, it is quite possible that relatively few additional measurements would be required to confidently assess trueness for method F.

Although the $\bar{x}_{\text{lab}} \pm U_{95}(\bar{x}_{\text{lab}})$ for method G spans the entire certified interval, there is less than a 50 % chance that the $\bar{x}_{\text{lab}}$ is within $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}})$. If the method's *s*(*x*) of 0.61 mg/g is known with high confidence then $n = (0.61/0.22)^2 = 2.8^2 \approx 8$ measurements are needed to confidently establish whether $\bar{x}_{\text{lab}}$ is more likely than not to be within the certified interval.

D.3 What n is needed? (Estimation from SRM data)

If, for any reason, *s*(*x*) must be estimated only from measurements on the SRM itself, the process is a bit more complicated. Rather than using a constant expansion factor of 2 to estimate the uncertainty from the standard deviation from just a few measurements, you should use the Student's *t* distribution at the 0.95 confidence level for *n*-1 independent analyses, $t_{95,n-1}$. Under the same assumption that half of your interval must fit within the certified uncertainty, $t_{95,n-1} \cdot s(\bar{x})/2 \leq U_{95}(x_{\text{NIST}})$, then

$$n \approx \left(t_{95,n-1}/2\right)^2 (s(x)/U_{95}(x_{\text{NIST}}))^2.$$

Since the value of $t_{95,n-1}$ depends on *n*, you can estimate *n* most easily with statistical software (see Appendices 3 and 4) or from a plot of *n* as a function of the ratio $s(x)/U_{95}(x_{\text{NIST}})$. Figure D.2 displays ratios corresponding to *n* from 2 to 32 when $s(x)$ is well known (open diamonds) and when it is estimated from the same results used to estimate the mean (solid diamonds). The solid vertical lines represent the ratio values for the 1.22 mg/g $s(x)$ of methods E and F and the 0.61 mg/g $s(x)$ of method G.

The estimates are quite similar for ratios greater than about 5 (and less than about 0.5, because a minimum of two values is required to estimate a standard deviation). But for ratios between these values, you must make a few more measurements to confidently assess trueness than is indicated using a fixed expansion factor of 2. The Student's *t* expansion suggests that 11 rather than 8 results are needed when the ratio is 2.8 (method G). Both estimates suggest that more than 30 results would be required when the ratio is 5.6 (methods E and F).

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.SP.260-181

Remember that these values are appropriate when you require that there be at least a 50 % overlap of the 95 % level of confidence interval about your estimated mean with the certified interval. The extent of overlap required for a particular purpose reflects value judgments about the relative costs associated with false positive and false negative risks. The smaller the overlap, the greater the risk of accepting a biased method as unbiased. The greater the overlap, the greater the number of independent results that will be needed. It's up to you to decide what risks and costs are fit for your purpose, and thus the degree of overlap that is acceptable [26,27].

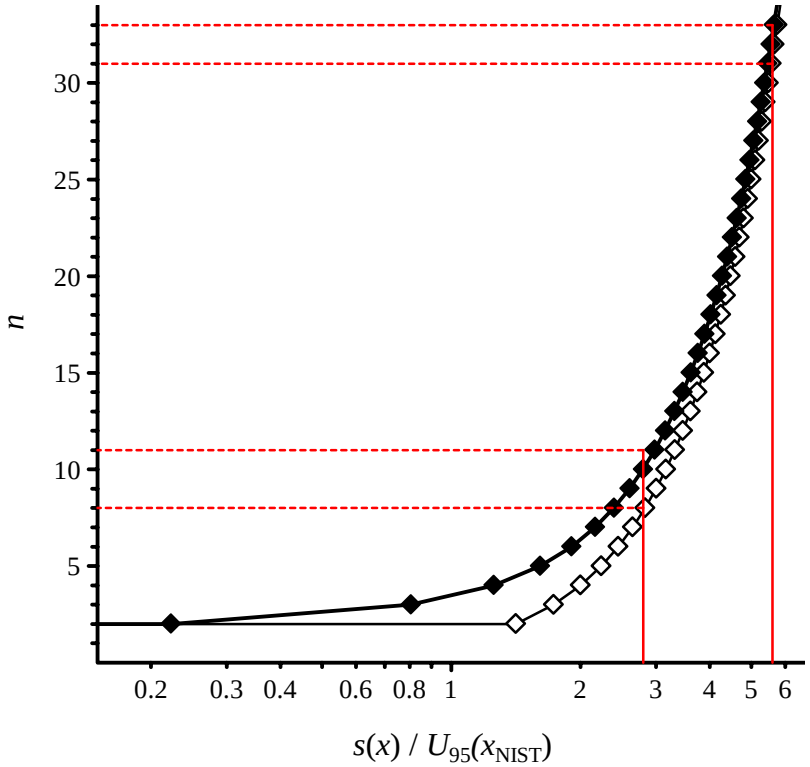


Figure D.2. Number of results needed to confidently assess trueness

Open diamonds (◊) represent the function when $s(x)$ is well known.
Closed diamonds (♦) represent the function when $\bar{x}$ and $s(x)$ are estimated from the same data.

D.4 Quantitative estimation of trueness (bias)

The graphical analysis described above is often sufficient to establish whether or not your measurement process is capable of providing fit-for-purpose results. Assuming that $s(\bar{x})$ is at least as small as $U_{95}(x_{\mathrm{NIST}})$, the difference between the assigned value for the SRM and your mean value is a good quantitative summary of your method's bias:

$$Bias = \bar{x}_{\mathrm{lab}} - x_{\mathrm{NIST}} \, .$$

The combined standard uncertainty (i.e., the standard deviation) associated with *Bias* is:

$$u(Bias) = \sqrt{s^2(\bar{x}_{\mathrm{lab}}) + u^2(x_{\mathrm{NIST}})} \; ; \; u(x_{\mathrm{NIST}}) = U_{95}(x_{\mathrm{NIST}})/2 \, .$$

The 95 % level of confidence expanded uncertainty estimate on the *Bias* is again

$$U_{95}(Bias) = k \cdot u(Bias) \, .$$

Since $u(x_{\mathrm{NIST}})$ is known with high confidence, the number of degrees of freedom associated with $u(Bias)$ will, in general, be the number associated with $u(\bar{x}_{\mathrm{lab}})$. However, if $s(\bar{x}_{\mathrm{lab}})$ is estimated from a small number (n) of independent measurements and much smaller than $u(x_{\mathrm{NIST}})$, the effective number of degrees of freedom for k, $k = t_{95,\nu_{\mathrm{eff}}}$, can be calculated using the Welch-Satterthwaite formula [23]:

$$\nu_{\mathrm{eff}} = \frac{\left(s^2(\bar{x}_{\mathrm{lab}}) + u^2(x_{\mathrm{NIST}})\right)^2}{s^4(\bar{x}_{\mathrm{lab}})/(n-1) + u^4(x_{\mathrm{NIST}})/60} \, .$$

The "60" is an arbitrary stand-in for the "large" number of degrees of freedom associated with an SRM's certified value, but it is convenient in that $t_{95,60} = 2.000$ and so $t_{95,\nu_{\mathrm{eff}}}$ will not be less than 2.

If zero is contained within the interval [$Bias - U_{95}(Bias)$ to $Bias + U_{95}(Bias)$], then with about a 95 % level of confidence your method is not significantly biased. If zero is not contained within the interval, you should investigate why your method appears biased and take corrective action.

D.5 The perils of checking trueness with two or more validation materials

When two or more appropriate validation materials are available, it is possible to get "mixed signals": the *Bias* determined relative to one material may be significant while the *Bias* for another is insignificant. Reference [28] reviews various mathematical ways that have been proposed to deal with such ambiguous situations by including the uncorrected bias (and its estimated uncertainty) into the estimated uncertainty. This approach should be avoided since (in addition to increasing the uncertainty beyond what should be attainable) it provides a disincentive to discovering why your method appears biased with some samples and not with others.

It is also possible that one or another of the natural-matrix SRMs has degraded or even been incorrectly value assigned; if you have strong evidence that this may be the case, you can e-mail the SRM's Technical Contact. This person's name is listed on the SRM's page on the NIST website, www.nist.gov/srm; enter the SRM number and you'll be taken to a second page, where you can select the link to the SRM in which you are interested.

Step E. Use of an SRM for Characterization of an In-House Quality Control Material

SRMs are expensive and are prepared in limited quantities. While you need to have quality control (QC) materials that you analyze on a regular basis to document that your measurement process is stable (and to discover instabilities quickly if it isn't), using an SRM just for QC wastes scarce resources. You should consider preparing one or more in-house QC materials.

A QC material must be stable and homogeneous and have a composition similar to that of the samples you will be analyzing. For the protein bar example, you might want to grind and package a sufficient quantity of a single batch of the protein bar itself. Grinding will help to ensure homogeneity. Packaging in single-use portions (under argon or nitrogen if possible) and cold storage (frozen, preferably at -80 °C) will enhance long-term stability.

You will need to characterize your in-house QC material(s) by analyzing them and the SRM(s) together at least 10 times over a relatively long period of time. You must prepare all samples independently from start to finish: you cannot merely analyze a single preparation multiple times. The results for the SRM are used to establish that your measurement process is acceptably stable. Assuming that the results for the candidate QC material(s) are acceptably precise, the mean and standard deviation of these results can be confidently used to establish action limits [17].

Figure E.1 displays a comparison of (dry-lab) results for total catechin monomers in SRM 2384 (solid circles) and your candidate in-house QC material (open circles). These "results" represent one or two measurements of the candidate control material per working day over a period of one month and a few measurements of the SRM per week (data and calculations are in Appendix IV.)

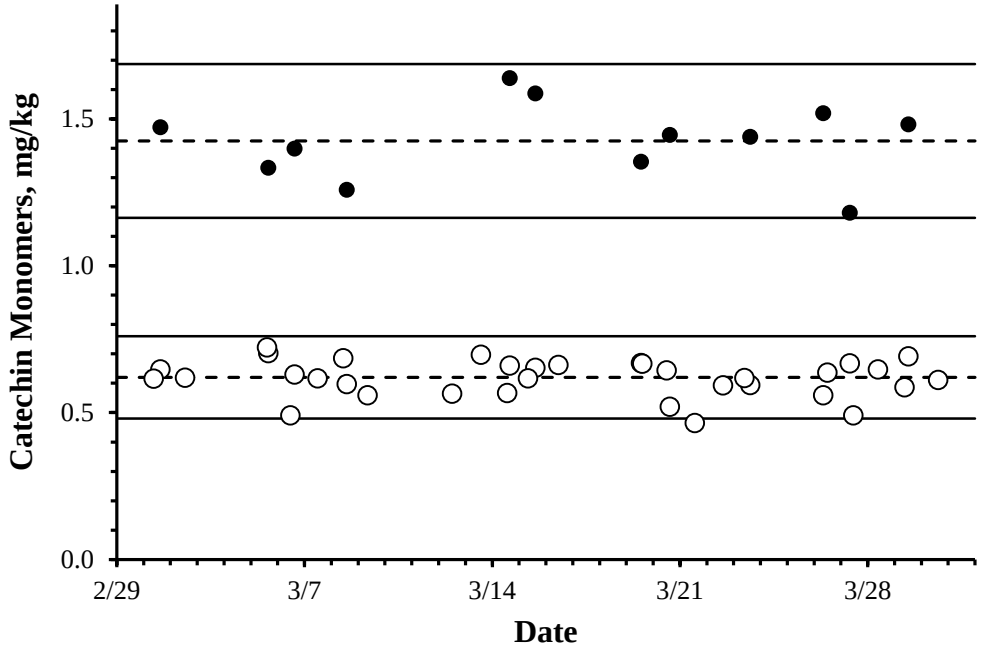


Figure E.1. Comparison of Measurement Results for the SRM and Candidate QC Material

The dashed line in the middle of the SRM results in Figure E.1 represents the mean value of your 12 results for the SRM, $\bar{x}_{\text{lab,SRM}} = 1.43$ mg/kg, with the solid bracketing lines representing an approximate 95 % level of confidence interval on the measurements,

$$U_{95}\left(\bar{x}_{\text{lab,SRM}}\right) = 2 \cdot s\left(\bar{x}_{\text{lab,SRM}}\right) = 0.26 \text{ mg/g}.$$

Note that this interval is based only on the measurements you made. Likewise for the QC material, the dashed and bracketing lines represent the mean and its 95 % level of confidence interval, $\bar{x}_{\text{lab,QC}} \pm U_{95}\left(\bar{x}_{\text{lab,QC}}\right) = 0.62$ mg/g ± 0.13 mg/g, for the 34 measurements that you made on the QC material.

The lack of any apparent trend in either set of results suggests that the measurement process and the QC material are adequately stable. The percent relative standard deviations of the two materials, $CV_{\text{lab,SRM}} = \frac{100 \times 0.26/2}{1.43} = 9.1$ % and $CV_{\text{lab,QC}} = \frac{100 \times 0.13/2}{0.62} = 10.5$ %, are similar enough to conclude that the performance of the measurement process is about the same for the two materials and that the QC material is adequately homogenous. Rather than using the SRM to monitor your measurement process, you can now confidently – and much less expensively – use your own material.

Should future measurement results for the QC material start to diverge from the trend or become significantly less precise, you can identify whether the material or the measurement process itself is at fault by again evaluating the SRM. If you conclude that the QC material has degraded, you will need to prepare a new material and characterize it against the SRM as with the original material.

Step F. Use of an SRM in Value Assigning a Secondary Reference Material

Monitoring the stability of a measurement process requires QC materials that are stable and homogenous, but you don't have to know the true value of *X* in the material in order to monitor the imprecision of the method (see Step E). However, if the measurement process is traceably calibrated (see Step B), the sources of imprecision identified and characterized (Step C), and the trueness of the calibration confirmed by analysis of matrix-matched natural-matrix SRMs (Step D), then the QC material can become a secondary reference material (RM). If sufficiently well characterized, such RMs can be used in place of SRMs for (nearly) all purposes, and with proper documentation [29,30] can be used and even sold as CRMs. To minimize the accumulation of uncertainty and propagation of unrecognized biases, secondary RMs should not be used to value assign other secondary RMs if CRMs of "higher metrological order" are available.

While the control chart analysis presented in Step E is not fully adequate for value assigning a commercial CRM, it may be suitable for within-organization tasks where the assessment of accuracy rather than just imprecision is important. A data and calculation example of value assignment of a QC material as a secondary reference material is provided in Appendix IV.

Recall that the 12 independent measurements of the SRM yielded $\bar{x}_{\text{lab,SRM}} = 1.43$ mg/kg with $U_{95}(\bar{x}_{\text{lab,SRM}}) = 2 \cdot s(\bar{x}_{\text{lab,SRM}}) = 0.26$ mg/g and that 34 measurements on the QC material yielded $\bar{x}_{\text{lab,QC}} = 0.62$ mg/g with $U_{95}(\bar{x}_{\text{lab,QC}}) = 2 \cdot s(\bar{x}_{\text{lab,QC}}) = 0.13$ mg/g. The standard deviations of the mean values are then $s(\bar{x}_{\text{lab,SRM}}) = 0.13/\sqrt{12} = 0.038$ mg/g and $s(\bar{x}_{\text{lab,QC}}) = 0.065/\sqrt{34} = 0.011$ mg/g.

The certified value for total catechin monomers in SRM 2384 is $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}}) = (1.49 \pm 0.22)$ mg/g so $Bias = 1.43 - 1.49 = -0.06$ mg/g, $U_{95}(Bias) = 2\sqrt{0.011^2 + (0.22/2)^2} = 0.22$ mg/g. Since 0.06 mg/g is considerably less than 0.22 mg/g, *Bias* can be considered insignificant and, with yellow-light caution, the measurement procedure considered to be unbiased. The appropriate assigned value for the QC material is then $\bar{x}_{\text{lab,QC}} \pm U_{95}(\bar{x}_{\text{lab,QC}}) = (0.62 \pm 0.13)$ mg/g.

Conclusion

So there you have it – selection and use of SRMs for quality assurance and establishing traceability in six (perhaps-not-so) easy steps. As a greater variety of materials become available, it becomes easier to find a match to your samples, although there will always be cases in which no good match exists. The analyses themselves require a lot of work, but that work is necessary to assure the quality of all of your results. Fortunately, the symbols in the equations seem more cumbersome than the mathematical ideas and operations themselves, and so once you're to that point, you're pretty much home free. Good luck and happy measuring!

References

1 Nutrition Labeling and Education Act Public Law 101-535 [HR 3562]; Nov 8 1990; http://thomas.loc.gov/cgi-bin/bdquery/z?d101:H.R.3562

2 Dietary Supplement Health and Education Act Public Law 103-417 [S784]; Oct 25 1994; http://thomas.loc.gov/cgi-bin/bdquery/z?d103:S784

3 US Food and Drug Administration (2005) Food CGMP Modernization – A Focus on Food Safety. http://www.fda.gov/food/guidanceregulation/cgmp/ucm207458.htm

4 JCGM 200:2012. International vocabulary of metrology – Basic and general concepts and associated terms (VIM) 3rd edition, JCGM, 2012, http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf

5 Eurachem (2002) The selection and use of reference materials. http://www.eurachem.org/index.php/publications/guides/usingrm

6 Nordic Committee on Food Analysis (1999) Evaluation of results derived from the analysis of certified reference materials (NMKL procedure no 9) Nordic Committee on Food Analysis, Oslo, Norway http://www.nmkl.org/Engelsk/Newsletter/eng46.htm - procedure

7 National Institute of Standards and Technology, Standard Reference Materials, http://www.nist.gov/srm

8 AOAC International, Technical Division on Reference Materials, Reference Material Database, http://www.aoac.org/divisions/tdrm.html

9 COMAR, International Database for Certified Reference Materials, http://www.comar.bam.de/en/

10 Phillips KM, Wolf WR, Patterson KY, Sharpless KE, Amanna KR, Holden JM (2007) Summary of reference materials for the determination of the nutrient composition of foods. Accred Qual Assur 12:126-133

11 Eurachem (1998) The fitness for purpose of analytical methods A laboratory guide to method validation and related topics; First Internet Version, Dec 1998 http://www.eurachem.org/index.php/publications/guides/mv

12 US Department of Agriculture, National Nutrient Database for Standard Reference, http://ndb.nal.usda.gov/

13 21 CFR 101.9 Nutrition Labeling of Food. http://www.gpo.gov/fdsys/granule/CFR-2012-title21-vol2/CFR-2012-title21-vol2-sec101-9/content-detail.html

14 Duewer DL, Parris RM, White E, May WE, Elbaum H (2004) An Approach To The Metrologically Sound Traceable Assessment of the Chemical Purity of Organic Reference Materials. NIST Special Publication 1012, NIST, Gaithersburg (MD), pp. 53 http://www.nist.gov/customcf/get_pdf.cfm?pub_id=901295

15 NIST, Traceability - NIST Policy and Supplementary Materials (2012), http://www.nist.gov/traceability

16 Ellison SRL, King B, Rösslein M, Salit M, Williams A (eds) (2003) Traceability in chemical measurement: a guide to achieving comparable results in chemical measurement. http://www.eurachem.org/index.php/publications/guides/trc

17 Hibbert DB, Gooding JJ (2006) Data Analysis for Chemistry: An Introductory Guide for Students and Laboratory Scientists, Oxford University Press, New York

18 Prichard E, Barwick V (2007) Quality Assurance in Analytical Chemistry, Wiley, Chichester England.

19 Ellison SLR, Barwick VJ, Farrant TJD (2009) Practical Statistics for the Analytical Scientist: A Bench Guide, 2nd Ed, Royal Society of Chemistry

20 Kallner A (2014) Laboratory Statistics: Handbook of Formulas and Terms, Elsevier, Waltham MA, USA

21 Eurachem/CITAC Guide (2007) Use of uncertainty information in compliance assessment http://www.eurachem.org/index.php/publications/guides/uncertcompliance

22 Rukhin AL (2013) Assessing compatibility of two laboratories: formulations as a statistical hypothesis testing problem. Metrologia 50:49-59 http://iopscience.iop.org/0026-1394/50/1/49/

23 JCGM 100:2008. Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM). JCGM. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

24 Thompson M, Wood R (2006) Using uncertainty functions to predict and specify the performance of analytical methods. Accredit Qual Assur 10:471-478

25 Appendix D, Official Methods of Analysis of AOAC International, 17th Edition, AOAC International, Gaithersburg, MD (2000)

26 Rukhin AL (2014) Compatibility verification of certified reference materials and user measurements. Metrologia 51:11-17. http://iopscience.iop.org/0026-1394/51/1/11/

27 AOAC Guidelines for Single Laboratory Validation of Chemical Methods for Dietary Supplements and Botanicals, http://www.aoac.org/dietsupp6/Dietary-Supplement-web-site/DSHomePage2.html

28 Magnusson B, Ellison SLR (2008) Treatment of uncorrected measurement bias in uncertainty estimation for chemical measurements. Anal Bioanal Chem 290, 201-213.

29 ISO Guide 34:2009 General requirements for the competence of reference material producers. ISO

30 ISO Guide 35:2006 Reference materials – general and statistical principles for certification. ISO

Appendix I: Acronyms and Symbols

Bias	The difference between your mean value for X in your unit of the SRM and NIST's certified value
CRM	Certified Reference Material, an RM accompanied by a certificate issued by an authoritative body
CV	Relative standard deviation expressed as percent (aka "coefficient of variation")
$CV_{\text{lab,SRM}}$	*CV* of your measurements of X on the SRM
$CV_{\text{lab,QC}}$	*CV* of your measurements of X on your QC material
i	Index over measurement values
j	Index over subsets (groups) of a set of measurements of X
k	coverage factor for transforming standard uncertainties (u – think standard deviations, where $\pm u$ spans about 68 % of the possible values) into 95 % level of confidence expanded uncertainties (U_{95}, where $\pm U_{95}$ is expected to span about 95 % of the possible values). When u is estimated with great confidence, $k = 2$; when u is estimated from a limited number of measurements, $k = t_{95,n-1}$.
MS_{wth}	Within-subset mean squares (see [17,19] for further information)
MS_{btw}	Between-subset mean squares (see [17,19] for further information)
m	Number of subsets (groupings) within a given set of measurements of X
NIST	National Institute of Standards and Technology
n	Number of data values, x, in a given set of measurements of X
n_j	Number of data values, x, in a given subset (group) of the entire set of measurements of X
n'	Effective number of data values per subset (group) when the n_j are not the same
QC	Quality Control
RM	Reference Material, "material sufficiently homogeneous and stable with reference to specified properties, which has been established to be fit for its intended use in measurement or in examination of nominal properties" (VIM 5.13,[4])
SRM	Standard Reference Material, a CRM provided by NIST
$s(x)$	Standard deviation of your measurements of X
$s(\bar{x})$	Standard deviation of the mean of your measurements of X , aka "standard error of the mean"
$s(\bar{x}_{\text{lab}})$	Standard deviation of the mean of your measurements of X on a material
$s(\bar{x}_{\text{lab,QC}})$	Standard deviation of the mean of your measurements of X on independently prepared subsamples of your QC material
$s(\bar{x}_{\text{lab,SRM}})$	Standard deviation of the mean of your measurements of X on independently prepared subsamples of the SRM
s_{I}	Intermediate imprecision, the standard deviation of measurements made on independently prepared subsamples of a material made by multiple analysts

	within a single laboratory or one analyst of that laboratory over a long period of time.
s_{post}	Post sample preparation imprecision, the standard deviation of a series of measurements made on a single preparation of a sample made by one analyst over a very short period of time
s_{r}	Repeatability imprecision, the standard deviation of measurements on independently prepared subsamples of a material performed over a short period of time by the same analyst
$t_{95,n-1}$	Student's *t* value for expanding *u* to U_{95} given that *u* has been estimated from *n* independent measurements. See "*k*"
v_{eff}	Number of effective degrees of freedom used to estimate *k* when *u* is estimated as a combination of two or more uncertainty components
$u(Bias)$	Combined standard uncertainty (*i.e.*, standard deviation) of the *Bias* estimate
$u(x_{\text{NIST}})$	One-half of NIST's $U_{95}(x_{\text{NIST}})$: $x_{\text{NIST}} \pm 2 \cdot u(x_{\text{NIST}})$ provides the same approximate 95 % level of confidence coverage as $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}})$.
$U_{95}(Bias)$	NIST certified approximate 95 % level of confidence interval about the certified value.
$U_{95}(x_{\text{NIST}})$	NIST certified approximate 95 % level of confidence interval about the certified value.
X	the measurand of interest, where "measurand" is a particular analyte in a defined sample matrix
x	a measured value of X
$x_{\text{lab},i}$	the i^{th} measurement of X that you make under defined experimental conditions
x_{NIST}	NIST certified value for the level of X in a given SRM. With high confidence, $x_{\text{NIST}} \pm U_{95}(x_{\text{NIST}})$ is expected to include the true value of X in your unit of the SRM
$\bar{x}$	Arithmetic mean of your measurements of X
$\bar{x}_{\text{lab}}$	Arithmetic mean of your measurements of X
$\bar{x}_{\text{lab,QC}}$	Arithmetic mean of your measurements of X made on your own candidate quality control material
$\bar{x}_{\text{lab,SRM}}$	Arithmetic mean of your measurements of X made on the SRM

Appendix II: Statistics and Examples

The specifications indicated in a NIST reference material certificate provide the estimated measurand x_{NIST}, i.e., the certificate value, and the expanded uncertainty $U_{95}(x_{\mathrm{NIST}})$. Thus, $x_{\mathrm{NIST}} \pm U_{95}(x_{\mathrm{NIST}})$ is the uncertainty interval for the measurand x. Commonly the interval is written with an expansion factor 2, $U_{95}(x_{\mathrm{NIST}}) = 2\sigma_{\mathrm{NIST}}$.

The user's replicated measurements, say, $x_1, \ldots, x_n$ are summarized by the value $\bar{x}$, which is the estimated measurand (typically the sample mean) and s, which estimates the repeatability standard deviation. Sometimes $s/\sqrt{n}$ has the meaning of the uncertainty $u(\bar{x})$ that may include components estimated from information other than the statistical analysis of series of observations in place of or in addition to the statistical analysis of observations. In the following we suppose that s is the sample standard deviation which does not depend on $\bar{x}$. In many situations this independence can be approximately achieved by a suitable transformation of x's.

In the simplest setting accepted here, x_i is a realization of a Gaussian random variable with some mean $x_{\mathrm{NIST}} + \Delta$ and some unknown standard deviation τ. The "no bias" (trueness or compatibility) hypothesis H_0 means that $\Delta = 0$. The error τ represents the repeatability of the lab's measurements.

The commonly used statistical procedure, the classical t-test, rejects compatibility or "compliance" of results to the certified value when

$$|\bar{x} - x_{\mathrm{NIST}}| \geq \frac{t(\frac{\alpha}{2}, n-1)\, s}{\sqrt{n}}.$$

When $\Delta = 0$, the probability of false positives (Type 1 error) by using t-test is exactly α. If $\Delta \neq 0$, the distribution of the ratio $\sqrt{n}(\bar{x} - x_{\mathrm{NIST}})/s$ is known as a noncentral t-distribution with the same degrees of freedom ν and the noncentrality parameter $\sqrt{n}\Delta/\tau$. Besides controlling for the Type 1 error (say, $\alpha = 0.05$), one would like to have the probability (say, β) of false negatives (Type 2 error) as small as possible.

For that purpose you must specify the metrologically important critical bias, $\Delta_c \neq 0$, whose value is large enough to worry about. For the bias to be deemed significant, this value cannot be smaller than $U_{95}(x_{\mathrm{NIST}})$, but realistically Δ_c should not be taken too large. We recommend to limit the choice of Δ_c to the range $U_{95}(x_{\mathrm{NIST}}) \leq \Delta_c \leq 2U_{95}(x_{\mathrm{NIST}})$.

The larger Δ_c, the smaller is the sample size n needed to attain a given Type 2 error, the false negatives probability β, at Δ_c. This balance can be achieved only if there is some information about the unknown τ. Indeed the probability of the Type 2 error is a function of the noncentrality parameter $\sqrt{n}\Delta/\tau$. If τ were known, you could solve for n in the equation, Type 2 error = β, to get the needed sample size n, $n \approx (z(\alpha/2) + z(\beta))^2\tau^2/\Delta_c^2$, where $z(\alpha/2)$ and $z(\beta)$ are the critical values of the standardized normal distribution at the specified probabilities. This value of n is necessary to have the test of power 1 - β.
Since τ is unknown, you must estimate it. For this purpose it may be helpful to compare τ to σ_{NIST} which by itself is not employed in the t-test. Unless $U_{95}(x_{\mathrm{NIST}})$ includes substantially large

components that were not estimated by statistical analysis of a series of measurements, τ must be larger than $U_{95}(x_{\mathrm{NIST}})$. One can take $\tau = B\sigma_{\mathrm{NIST}}$, where the corresponding factor B, say, $1 \le B \le 5$, is determined from your preparatory work.

This factor B also can be described via the *measurement capability index* which here is the ratio of (expected) widths of two coverage intervals, $C_m = 2\sqrt{n}\sigma_{\mathrm{NIST}} / \left(t(\frac{\alpha}{2}, n-1)\,\tau\right)$. Indeed, for the desired measurement capability index C_m, $B = 2\sqrt{n}C_m^{-1} / t(\frac{\alpha}{2}, n-1)$. See [i] for the discussion of other capability characteristics in quality control problems. The recommended value C_m should exceed 1.5 [ii]. According to the rule of thumb, B is about 3 [iii].

Returning to the issue of controlling the Type 2 error, by taking $\tau = B\sigma_{\mathrm{NIST}}$, one gets the estimated value of the noncentrality parameter, $\sqrt{n}\Delta_c/(Bx_{\mathrm{NIST}})$, so that the numerical evaluation of the smallest n such that Type 2 error $\le \beta$, becomes feasible. Modern statistical software, in particular the R language (free software for statistical computing and graphics; http://www.r-project.org/) available in the public domain, offers you several routines to determine numerically the needed sample size for any given values of α, β and Δ_c/τ. Additionally, the software package metRology (http://www.nist.gov/itl/sed/gsg/metrology.cfm) provides specialized functions for statistical metrology as a package for the R software environment, with an option of a convenient user interface using Microsoft Excel (http://www.nist.gov/itl/sed/gsg/metrology-for-microsoft-excel.cfm).

Example 1. Compute power of *t* test

We present here exact calculations for the example 'dry lab' catechin monomer data for seven different methods in Figure 3 in the main text.

Assume that the critical value is $\Delta_c = 2U_{95}(x_{\mathrm{NIST}}) = 0.44$ and $\tau_1 = 0.087/2$ (for methods A, B, C, and D); $\tau_2 = 1.22/2$ (methods E and F); and $\tau_3 = 0.88/2$ (method G). With $d_i = \Delta_c/\tau_i$, you determine $d_{1(\mathrm{A\text{-}D})} = 10.115$, $d_{2(\mathrm{E,F})} = 0.721$, and $d_{3(\mathrm{G})} = 1$ that represents the three grouped methods that have comparable uncertainty.

First, in R (refer to Appendix III for the installation of R and metRology), set your working directory through the **File** – **Change dir** menu dropdown list (this will remain the working directory as long as you maintain the R session). Now, install the Basic power calculations 'pwr' package through the **Packages** – **Install package(s)** menu dropdown list. At the R Console command prompt (>), load the 'pwr' package within the R environment:

(Note: In this document red typeface denotes commands to be entered by the user into the R console and blue typeface is of appropriate outputs.)

```
> library(pwr)
```

Now compute a series of power tests for the three grouped scenarios (d_1, d_2 and d_3) and store the results in the associated data structures (P1, P2 and P3, respectively):

```
> P1=pwr.t.test(d=10.115,power=0.8,sig.level=0.05,type="one.sample",alternative="two.sided")
> P2=pwr.t.test(d=0.721,power=0.8,sig.level=0.05,type="one.sample",alternative="two.sided")
> P3=pwr.t.test(d=1,power=0.8,sig.level=0.05,type="one.sample",alternative="two.sided")
```

Typing P1, P2 or P3 at the command prompt provides a summary of the results for each of the individual power tests:

```
> P1

     One−sample t test power calculation

              n = 2.054403
              d = 10.115
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

> P2

     One-sample t test power calculation

              n = 17.11833
              d = 0.721
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

> P3

     One−sample t test power calculation

              n = 9.93785
              d = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

According to these calculations, about 2 observations are required to have the type 2 error of *t*-test about 0.2 when $\Delta_c/\tau = 10.115$ (P1), while for $\Delta_c/\tau = 1.0$ one needs nearly 10 measurements (P3).

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.SP.260-181

Example 2. Analysis of Variance (ANOVA) for the Catechin Data

We present here an ANOVA for the example catechin monomer validation data that was acquired by five separate analysts in Table 4 in the main text.

For the ANOVA calculations with the following R code, the raw catechin data should be organized into two separate columns, one labeled 'analyst' with values 1 to 5 representing the five different analysts and the second labeled 'mass' with the corresponding catechin mass fraction readings. You can generate the data in Excel and then save as either a .csv or .txt file. Import or 'read' the data (as either a .csv file or a .txt file) and convert it to a data frame in R:

```
> cate = read.csv(file="catechin.csv", head=TRUE) or
> cate = read.table(file="catechin.txt", head=TRUE)
```

Print the data frame and verify that is has the comparable structure:

```
> cate
```

```
   analyst mass
1        1 1.362
2        1 1.388
3        1 1.392
      .....
34       5 1.453
35       5 1.510
36       5 1.573
```

Load the Linear and Nonlinear Mixed Effects Models 'nlme' package through the **Packages** – **Load package** menu dropdown list and then add it to the R environment library:

```
> library(nlme)
```

Construct a new 'cate' data object with the data grouped:

```
> cate=groupedData(mass ~ 1| analyst, data=read.table("catechin.txt",header=TRUE))
```

Generate a summary to present the basic data information within the 'cate' data structure:

```
> summary(cate)
```

```
 analyst      mass
1:10    Min.   :1.357
5: 4    1st Qu.:1.423
4: 4    Median :1.477
3:10    Mean   :1.504
2: 8    3rd Qu.:1.563
        Max.   :1.770
```

Here, the number of measurements for each analyst is displayed as well as the descriptive characteristics of their mass fraction readings.

Calculate the linear mixed-effects model fit of the 'cate' data and print a summary:

```
> cate_lme=lme(mass ~ 1, data=cate, random = ~ 1|analyst)
> summary(cate_lme)

Linear mixed-effects model fit by REML
Data: cate
AIC      BIC   logLik
-67.83977 -63.17373 36.91989
Random effects:
Formula: ~1 | analyst
(Intercept)   Residual
StdDev:  0.08472037 0.06996101

Fixed effects: mass ~ 1
Value  Std.Error DF  t-value p-value
(Intercept) 1.50919 0.03992659 31 37.79912       0

Standardized Within-Group Residuals:
Min         Q1        Med         Q3        Max
-2.06514039 -0.56280795  0.02422057  0.56949594  1.87991441

Number of Observations: 36
Number of Groups: 5
```

Determine the analysis of variance table and a one-way analysis of means for the 'cate' data:

```
> anova(lm(mass ~ analyst, data=cate))

Analysis of Variance Table
Response: mass
           Df         Sum Sq      Mean Sq    F value      Pr(>F)
analyst     4        0.25264      0.06316     12.865  2.772e−06 ***
Residuals  31        0.15220      0.00491

> oneway.test(mass ~ analyst, data=cate)

One−way analysis of means (not assuming equal variances)
data:    mass and analyst
F=8.6657, num df = 4.000, denom df=10.406, p−value=0.002438
```

The analysis of variance table unequivocally rejects the equality of means hypothesis if homogeneity of variances holds. The *p*-value 2.772×10^{-6} is quite small. The same hypothesis is also rejected with the *p*-value 0.002 when variances are not assumed to be equal.

Install the companion to applied regression 'car' package through the Packages – Install Package(s) menu dropdown list. Load the 'car' package within the R environment:

```
> library(car)
```

Compute the Levene's test for homogeneity of variance across the individual groups (i.e., analysts):

```
> leveneTest(mass ~ analyst, data=cate)

Levene's Test for Homogeneity of Variance (center=median)
       Df    F value    Pr(>F)
group   4     1.4716    0.2347
       31
```

This provides the estimates of the parameters in a random effects model which assumes a random bias presence. Namely, the measurements x_{ij} of the analyst $i = 1, \ldots, 5$, $1 \le j \le n_i$, $n_1 = 10$, $n_2 = 8$, $n_3 = 10$, $n_4 = 4$, $n_5 = 4$, are supposed to have the form

$$x_{ij} = \mu + \Delta_i + \varepsilon_{ij}.$$

Here μ is the common true mean, (estimated as 1.506 with uncertainty of 0.038), ε_{ij} are independent normal disturbances with zero mean and unknown variance σ_i^2 depending on the analyst i. The between-analyst effect values, Δ_i, which can be viewed as realizations of a random bias supposed to have zero mean and some unknown standard deviation which is determined to be 0.080 (Intercept Stdev).

Individual σ_i can be found from parameter estimates which give the ratios, σ_i/σ_1 with $\sigma_1 = 0.043$ (Residual StDev). The Δ_i predicted values can be found as $\bar{x}_i - \mu$, e.g. $\Delta_1 = 1.425 - 1.506 = -0.081$, $\Delta_2 = 1.649 - 1.506 = 0.143$, $\Delta_3 = 1.467 - 1.506 = -0.041$, $\Delta_4 = 1.533 - 1.506 = -0.026$, and $\Delta_5 = 1.473 - 1.506 = -0.033$. Thus in this example there was just one analyst (2) with a high bias, and all others were negatively biased. However, Levene's test of variances equality does not reject the homogeneity of variance hypothesis at the 95 % confidence level as the p-value is 0.234.

As additional information, Excel users may want to use metRology for Microsoft Excel, an add-in which allows to call R functions as worksheet functions [iv]. There are several web sites, intended mainly for applications in biostatistics, that also provide the needed sample size calculations [v]. Some additional R codes useful in metrology applications are provided in metRology. Support for metrological applications which can be accessed at http://metrology.sourceforge.net.

References in Appendix II

i NIST (2010) NIST/SEMATECH e-Handbook of Statistical Methods. http://www.NIST.gov/itl/sed/gsg/handbook project.cfm

ii ASME B89.7.3.1-2001 (2002) Guidelines for decision rules: considering measurement uncertainty in determining conformance to specifications. ASME, New York

iii Instone I (1996) Simplified method for assessing uncertainties in commercial, production environment. http://metrology forum.tm.agilent.com/easy.shtml

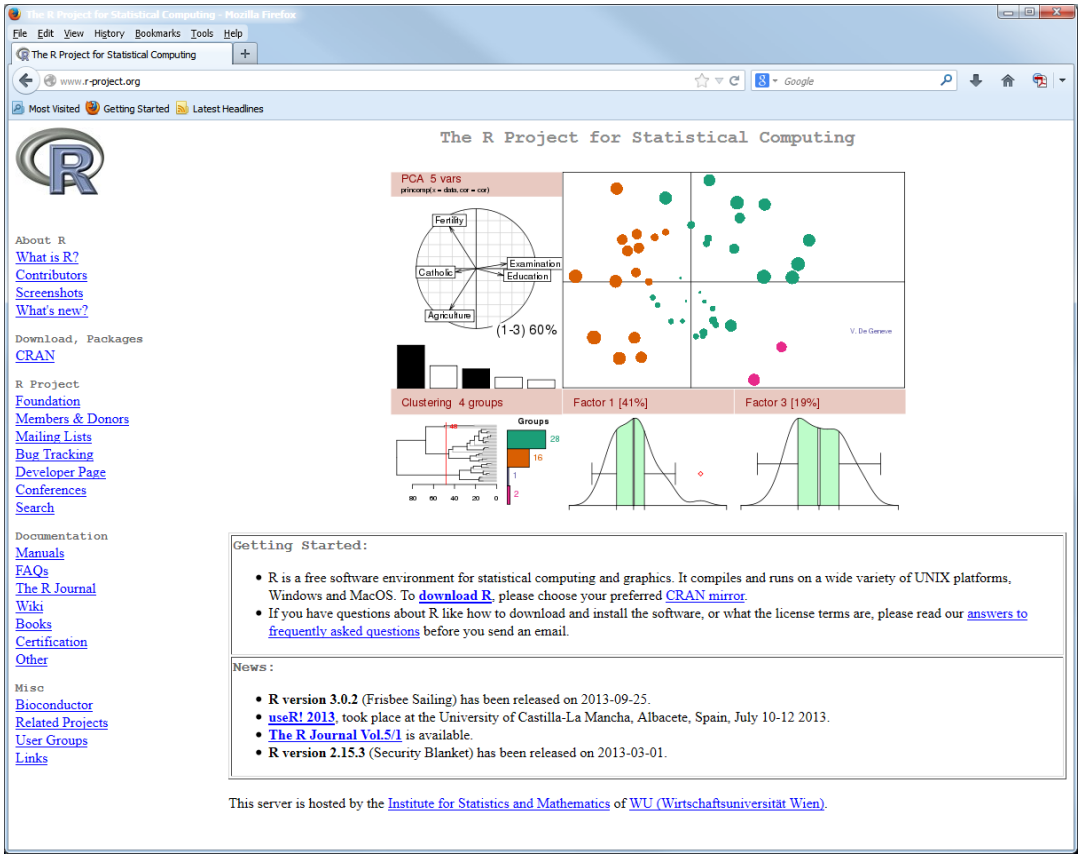
iv Heiberger, R. M., Neuwirth, E. (2009) A Spreadsheet Interface for Statistics, Data Analysis, and Graphics Springer NY

v E.g.: http://hedwig.mgh.harvard.edu/sample size, http://homepage.stat.uiowa.edu/ ˜rlenth/Power, http://calculators.stat.ucla.edu/powercalc.

Appendix III: Installation of R and met*R*ology

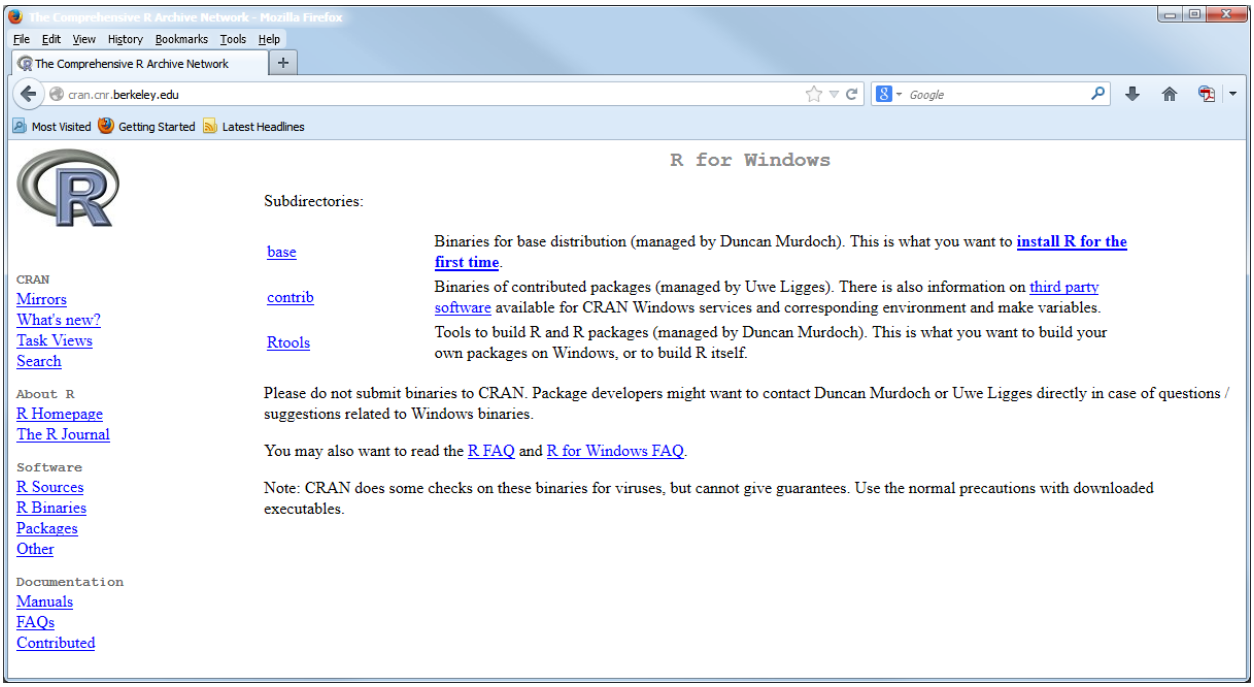
1. Install the newest version of R on your computer (Windows, MacOS or UNIX)

Download the R package from the R project website (http://www.r-project.org/) by either selecting the download R or Comprehensive R Archive Network (CRAN) mirror (CRAN mirror) link under the 'Getting Started' window:

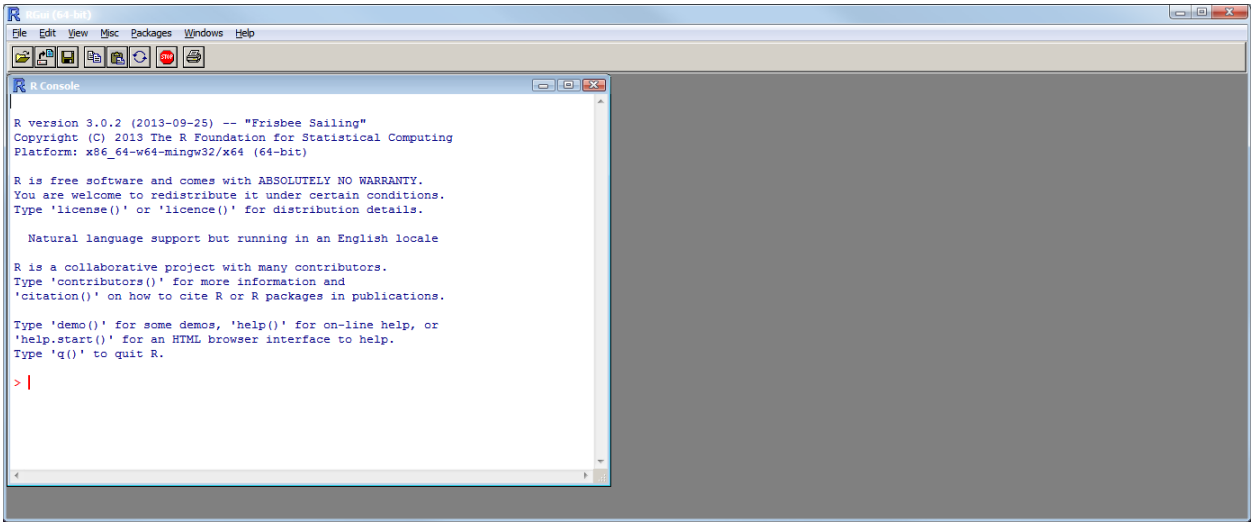


After selecting your CRAN mirror most suitable for your physical location (geographically proximal), select the Download and Install R link for your particular operating system (e.g., Download R for Windows). Install the base subdirectory, which are the binaries for base distribution and are needed to install R for the first time (the links are CRAN mirror-specific):

This publication is available free of charge from http://dx.doi.org/10.6028/NIST.SP.260-181



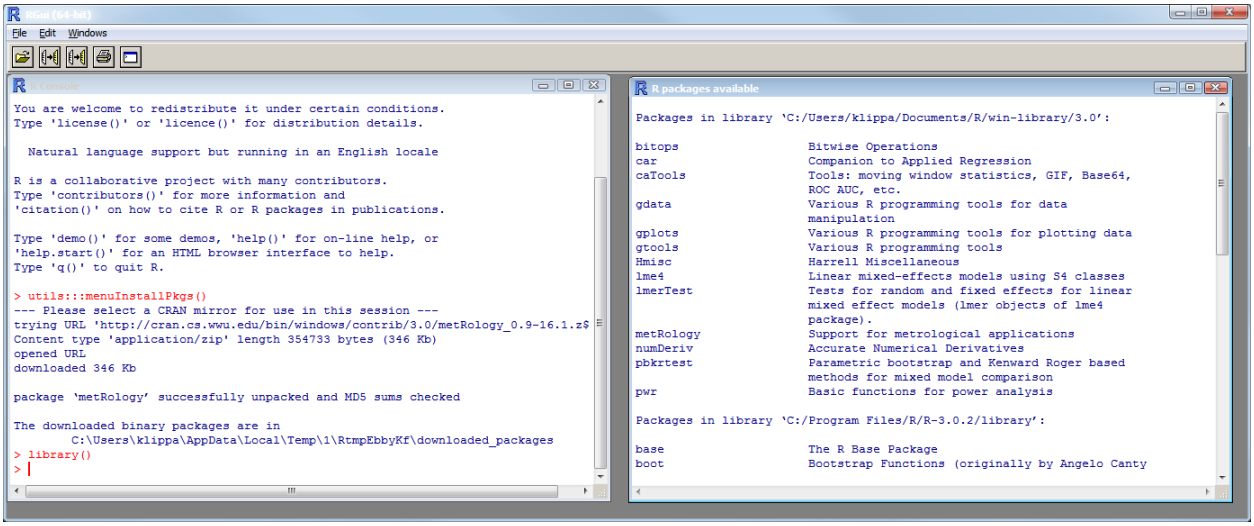
Click on the Download R 3.0.2 for Windows link to download the installation binary file (R-3.0.2-win.exe) to your computer. Save and then run the R executable file (R-3.0.2-win.exe) and select language preferences, directory locations, desktop icon shortcut, etc. during the installation. R is now installed and ready to start. Double-click the (desktop) icon and the initial R console window should appear:



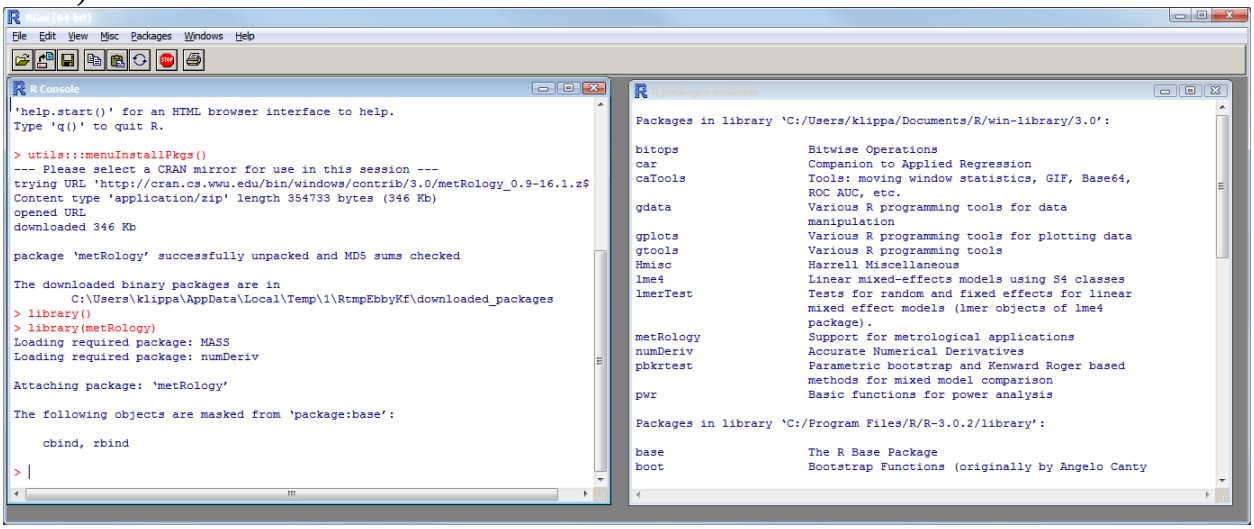
Set your working directory for file storage under the **File** – **Change dir…** menu dropdown list.

2. Install the met*R*ology data analysis package within R

Using the **Packages** Menu, select '**Install package(s)…**' You will then need to select a CRAN mirror. Select the 'metRology' package from the dropdown menu that is automatically displayed and it should install. You can type 'library()' at the command prompt (>) to see the list of packages stored in your library:



You now can load (or attach) the met*R*ology package either through the **Packages** – **Load package…** menu dropdown list or by typing 'library(metRology)' at the command prompt (>):



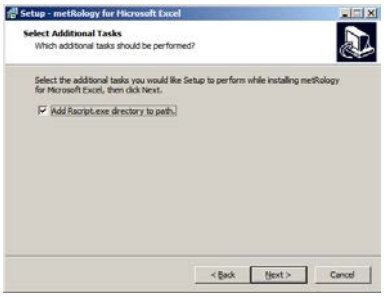
You are now ready to use the functions within the met*R*ology package.

More details regarding the met*R*ology software and its ongoing development are described on the NIST website: http://www.nist.gov/itl/sed/gsg/metrology.cfm and at: http://cran.at.r-project.org/web/packages/metRology/.

3. Install the met*R*ology add-in for Microsoft Excel

Detailed instructions on the installation of the met*R*ology add-in for Microsoft Excel are provided on the NIST website: http://www.nist.gov/itl/sed/gsg/upload/metRology-for-Microsoft-Excel-installation-instructions.pdf . You are encouraged to open these instructions and use them as a guide for the installation and testing of the add-in.

In brief, the installation file (metRology-for-Microsoft-Excel-v1-03-setup.exe) for the metRology add-in for Microsoft Excel can be downloaded from the NIST website: http://www.nist.gov/itl/sed/gsg/metrology-for-microsoft-excel.cfm. Save and then run the R executable file (metRology-for-Microsoft-Excel-v1-03-setup.exe). Prior to installation, make sure that it is to be installed in the following directory: C:\Program Files\metRology
for Microsoft Excel. If you have a 64-bit operating system, you may need to modify the destination location to C:\Program Files\ by deleting "(x86)". It is also important the Rscript.exe directory is added to the path:



Lastly, open the appropriate test file (e.g., http://www.nist.gov/itl/sed/gsg/upload/test-metRology-for-Microsoft-Excel.xlsm) and ensure that clicking on the 'Compute Uncertainty' button generates a result.

Appendix IV: Data for the Examples

Step D. Use of SRMs to Establish Trueness of Results

Data for Figure D.1: Catechin Monomers, mg/g:

	Catechin Monomers, mg/g			Summary Statistics, mg/g				
Method	Rep_1	Rep_2	Rep_3	n	$\bar{x}$	$s(x)$	$s(\bar{x})$	$2 \cdot s(\bar{x})$
A	1.362	1.419	1.511	3	1.431	0.075	0.043	0.087
B	0.562	0.619	0.711	3	0.631	0.075	0.043	0.087
C	1.641	1.698	1.790	3	1.710	0.075	0.043	0.087
D	1.722	1.779	1.871	3	1.791	0.075	0.043	0.087
E	2.710	1.490	0.270	3	1.490	1.220	0.704	1.409
F	3.582	1.419	1.511	3	2.171	1.223	0.706	1.412
G	2.010	2.110	1.010	3	1.710	0.608	0.351	0.702

Data for Figure D.2: Number of results needed to confidently assess trueness

		$s(x)/U_{95}(x_{\mathrm{NIST}}) = 2\sqrt{n}/k$	
n	$t_{95,n-1}$	$k = 2$	$k = t_{95,n-1}$
2	12.706	1.414	0.223
3	4.303	1.732	0.805
4	3.182	2.000	1.257
5	2.776	2.236	1.611
6	2.571	2.449	1.906
7	2.447	2.646	2.163
8	2.365	**2.828**[a]	2.392
9	2.306	3.000	2.602
10	2.262	3.162	2.796
11	2.228	3.317	**2.977**[a]
12	2.201	3.464	3.148
13	2.179	3.606	3.310
14	2.160	3.742	3.464
15	2.145	3.873	3.612
16	2.131	4.000	3.753
17	2.120	4.123	3.890
18	2.110	4.243	4.022
19	2.101	4.359	4.150
20	2.093	4.472	4.273
21	2.086	4.583	4.394
22	2.080	4.690	4.511
23	2.074	4.796	4.625
24	2.069	4.899	4.736
25	2.064	5.000	4.845
26	2.060	5.099	4.952
27	2.056	5.196	5.056
28	2.052	5.292	5.158
29	2.048	5.385	5.258
30	2.045	5.477	5.356
31	2.042	**5.568**[b]	5.453
32	2.040	5.657	5.547
33	2.037	5.745	**5.640**[b]
34	2.035	5.831	5.732
35	2.032	5.916	5.822
36	2.030	6.000	5.911
37	2.028	6.083	5.999
38	2.026	6.164	6.085
39	2.024	6.245	6.170
40	2.023	6.325	6.254

a First ratio greater than 2.8
b First ratio greater than 5.6

Step E. Use of an SRM for Characterization of an In-House Quality Control Material

Data for Figure E.1: Comparison of Measurement Results for the SRM and the QC Material

	Total Catechin Monomers, mg/g	
Date and Time	QC Material	SRM 2384
3/1/2012 9:15	0.615	
3/1/2012 15:09	0.647	1.471
3/2/2012 13:10	0.619	
3/5/2012 14:40	0.722	
3/5/2012 15:34	0.702	1.333
3/6/2012 11:26	0.490	
3/6/2012 15:18	0.629	1.399
3/7/2012 11:48	0.616	
3/8/2012 10:40	0.685	
3/8/2012 13:53	0.597	1.258
3/9/2012 8:36	0.558	
3/12/2012 12:18	0.565	
3/13/2012 13:53	0.697	
3/14/2012 13:25	0.566	
3/14/2012 15:42	0.660	1.639
3/15/2012 8:03	0.616	
3/15/2012 14:36	0.652	1.587
3/16/2012 11:20	0.662	
3/19/2012 13:08	0.669	1.354
3/19/2012 14:26	0.666	
3/20/2012 12:14	0.643	
3/20/2012 14:55	0.519	1.446
3/21/2012 13:13	0.464	
3/22/2012 14:35	0.593	
3/23/2012 9:54	0.618	
3/23/2012 14:49	0.594	1.439
3/26/2012 8:17	0.559	1.519
3/26/2012 11:55	0.636	
3/27/2012 8:05	0.667	1.180
3/27/2012 11:04	0.490	
3/28/2012 9:10	0.646	
3/29/2012 8:57	0.585	
3/29/2012 12:21	0.691	1.481
3/30/2012 15:12	0.610	

Statistic	QC Material	SRM 2384
n	34	12
$\bar{x}$, mg/g	0.616	1.426
$s(x)$, mg/g	0.063	0.131
$2 \cdot s(x)$, mg/g	0.126	0.262
CV, %	10.2	9.2

Step F. Use of an SRM in Value Assigning a Secondary Reference Material

Validation of absence of significant bias in SRM measurement values, mg/g

Your Measurements		NIST Certificate		Comparison	
Statistic	Value	Statistic	Value	Statistic	Value
n	12	n	60	v_{eff}	70
$\bar{x}_{\text{lab,SRM}}$	1.426	x_{NIST}	1.49	*Bias*	-0.06
$s(x)$	0.131				
$k = t_{95,11}$	2.20	k	2	k	2
$u(\bar{x}_{\text{lab,SRM}})$	0.038	$u(x_{\text{NIST}})$	0.11	$u(Bias)$	0.12
$U_{95}(\bar{x}_{\text{lab,SRM}})$	0.083	$U_{95}(x_{\text{NIST}})$	0.22	$U_{95}(Bias)$	0.23

QC material Value Assignment, mg/g

Statistic	Value
n	34
$\bar{x}_{\text{lab,QC}}$	0.616
$s(x)$	0.063
$k = t_{95,33}$	2.03
$u(\bar{x}_{\text{lab,QC}})$	0.011
$U_{95}(\bar{x}_{\text{lab,QC}})$	0.02