Radiocarbon, Vol 61, Nr 5, 2019, p 1159–1168 DOI:10.1017/RDC.2019.10
Selected Papers from the 23rd International Radiocarbon Conference, Trondheim, Norway, 17–22 June, 2018

LIFE AFTER SIRI—WHERE NEXT?

E Marian Scott[1]* · Philip Naysmith[2] · Gordon Cook[2]

[1]University of Glasgow - School of Maths and Statistics, University of Glasgow, Glasgow, G12 8QQ, Scotland, United Kingdom
[2]Scottish Universities Environmental Research Centre, SUERC Radiocarbon Dating Laboratory, East Kilbride, South Lanarkshire, Scotland, United Kingdom

ABSTRACT. Radiocarbon (^{14}C) dating is routinely used, yet occasionally, issues still arise surrounding laboratory offsets, and unexpected and unexplained variability. Quality assurance and quality control have long been recognized as important in addressing the two issues of comparability (or bias, accuracy) and uncertainty or variability (or precision) of measurements both within and between laboratories (Long and Kalin 1990). The ^{14}C community and the wider user communities have supported interlaboratory comparisons as one of several strands to ensure the quality of measurements (Scott et al. 2018). The nature of the intercomparisons has evolved as the laboratory characteristics have changed s. The next intercomparison is currently being planned to take place in 2019–2020. The focus of our work in designing intercomparisons is to (1) assist laboratories by contributing to their QA/QC processes, (2) supplement and enhance our suite of reference materials that are available to laboratories, (3) provide consensus ^{14}C values with associated (small) uncertainties for performance checking, and (4) provide estimates of laboratory offsets and error multipliers which can inform subsequent modeling and laboratory improvements.

KEYWORDS: accuracy, intercomparison, precision.

INTRODUCTION

The radiocarbon (^{14}C) community has a long experience of participating in intercomparisons. Even from the early days of the technique, it was not uncommon for a small number of laboratories to exchange samples (e.g., Otlet et al. 1980). However, as the community of laboratories grew then so did the scale of the intercomparisons, so that in the past 30 years, there have been five wide-scale intercomparisons, covering many different types of samples and many thousands of measurements (Scott et al. 2018). Questions that are often asked are whether the efforts are worthwhile? And what are the benefits and for whom?

One of the early objectives for our work in designing the sequence of intercomparisons (Cook et al. 1990; Scott et al. 1989; Scott Aitchison et al. 1990a; Scott et al. 1990b; International Collaborative Study [ICS], Third, Fourth, Fifth, and Sixth international radiocarbon intercomparisons [TIRI, FIRI, VIRI, and SIRI respectively] Harkness et al. 1989; Gulliksen and Scott 1995; Scott et al. 1997, 1998, 1991, 1992, 2004a, 2004b, 2007, 2010a, 2010b, 2010c; Bryant et al. 2000; Boaretto et al. 2002; Thompson et al. 2006) was to gain a better, empirically based understanding of the uncertainties associated with a measurement. Could we, by experimental means, quantify the variability observed and thus provide justification for the routinely quoted errors? A well-designed intercomparison is able to quantify both the variability we observe within a laboratory but more importantly the variability among laboratories. In this latter context, however, we need to carefully distinguish two forms of variability, one which is systematic and where we would typically use the language of offsets, i.e. systematic differences (usually expressed in terms of the average) between laboratories, and random, which would be the scatter of results from different laboratories (remembering that we would not expect identical results from different laboratories even when measuring identical samples).

However, there are other significant benefits which a laboratory can access from participation in an intercomparison. Four identified benefits are (1) experimentally quantifying uncertainty on measurements, (2) benchmarking laboratory performance and while an intercomparison is a

*Corresponding author. Email: Marian.Scott@glasgow.ac.uk.

snapshot in time, it provides an independent check which supplements routine QC procedures, (3) access to well characterized materials, typical of the routine dating material and spanning the age range, which form a catalog of reference materials, and (4) quantifying reproducibility, and comparability (for laboratory and user).

It is also worth reflecting that there are benefits which accrue indirectly to the user. Laboratory participation (and performance) in an intercomparison may be invisible to a user (we have always published a list of which laboratories participated but not identified their individual performance) since we have always argued that the laboratory-user relationship is an important one, and recommended users should ask the laboratories they wish to engage with what their QA/QC procedures are and how they perform in these global studies (or indeed in others). Of course, participating in an intercomparison is a time-limited and specific assessment, but it is indicative of a laboratory that is taking care of its performance. However, to the user, many of the benefits are indirect, and the laboratory efforts may indeed be invisible. Nevertheless, users should be aware of the efforts that laboratories make to deliver QA and QC and it would seem highly appropriate that the laboratory-user relationship should include the exchange of such information.

In this paper, we identify, develop and illustrate the four laboratory benefits and consider the future of the program. The structure of the paper is as follows: first, we develop each benefit in turn and give examples of those benefits, next we outline some specific aspects, and finally we reflect on the future.

BENEFITS FROM PARTICIPATING IN AN INTERCOMPARISON

Many benefits of participation in intercomparisons are derived from the quality assurance (QA) and quality control (QC) principles and policies adopted by laboratories, which intercomparisons support. The four benefits that intercomparisons provide are directly related to laboratory performance and the uncertainties associated with laboratory measurements (Thompson et al. 2006). In each of the following sections, we will describe the background to the benefit and how it manifests through the intercomparison.

First, it is worth considering the steps in the design and implementation of a study. The key steps are to:

- Identify the specific objectives for the study, and design the study accordingly to ensure that key characteristics can be estimated and modeled.
- Identify the samples that are required—sample criteria include
  - sufficient quantity
  - homogeneous
  - natural
  - interesting
  - spanning the ^{14}C activity range
- Invite laboratories to participate, and set a deadline for results
- Collate and analyse the results, focusing on uncertainty quantification—offsets
- Provide feedback to ensure that participation is helpful to the laboratory
- Characterize the materials so that they can be archived and made available as certified reference materials.

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

Quantifying Uncertainty and Variability

Uncertainty, variability, and errors are terms that are often used when discussing ^{14}C dating and intercomparisons. Uncertainty quantification is the natural place to start the discussion and that begins with the laboratory quoted error, which is calculated based on a number of considerations, but in essence represents the uncertainty in the measurement. This is quantified based on the variability we observe and is interpreted as the variability we would expect to observe if we were able to repeat the measurement. This also accounts for variability in standards, backgrounds and other laboratory processes. Typically, in many laboratories, this will be based in part on the observed standard deviation from a set of measurements (variation) made on an in-house well-characterized material (see Naysmith et al. 2019, this volume). In general terms, uncertainty, variability and error also relate to precision (precise results show small variation) and to the concept of repeatability where "repeatability of a measurement is the closeness of agreement between successive measurements carried out under the same conditions e.g., same location, same person, same measurement procedure." Another key property is that of reproducibility which refers to "the closeness of agreement between the results of experiments conducted under changed conditions, including different laboratories and instruments" (Taylor and Kuyatt 2001).

While these descriptions focus on "the variation in the final ^{14}C results," we can also consider how to decompose this variation into the contributions from the component stages of providing a ^{14}C date through designing certain aspects into the intercomparison or in-house. This is an attribution process that allows a laboratory to complete an accounting of where the variability is coming from and which sources contribute most. For an individual laboratory, the pretreatment chemistry used, the graphitization process, the measurement procedure (background and standards used), the stability of the AMS, and indeed the human aspects, all will make a contribution to the overall variability. How can we estimate, and quantify those uncertainties (i.e. attribute and quantify the variability due to these factors)? It is true that a full decomposition of the variation in a set of results is best achieved within the laboratory (Scott et al. 2019, this volume) using a designed experiment approach, but it is also possible to quantify some components of within and between laboratory variability in an intercomparison, if designed appropriately. In the intercomparison designs we have developed, this includes replication at different stages in the measurement process (e.g. duplicate samples), a hierarchical approach including both a pretreated and untreated material, and repeat of materials over time, providing linking samples to provide the temporal aspects. To achieve these designs, we need large quantities of sample material which must be shown to be homogeneous (see FIRI homogeneity testing [Scott 2003]). For the user, who may be interested in a compendium of dates e.g. for a specific site, they may be using results from multiple laboratories (focussing thus on reproducibility), so that the individual laboratory is also a potential source of variability.

Within laboratory aspects, included in the intercomparison design, have included a hierarchical approach (e.g. the three stages of ICS and FIRI/VIRI) where we designed a series of stages looking at bulk pretreatments, duplicate samples, and different pretreatment methods where variability may be introduced. Thus, in stage 2 of the ICS, we provided homogenized pretreated samples (in duplicate) and in stage 3, the raw material, while in FIRI, we again introduced duplicate samples. We have used two streams of humic samples (and wood) that appear in several studies (see Table 1; Scott et al. 2018, 2019).

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

Table 1 Bone samples used in the intercomparisons.

Study	Sample code	Sample type	Description
TIRI	L	Whalebone	Excavated in Norway in 1992
VIRI	E	Mammoth	The mammoth bone sample comes from a site called Quartz Creek, Dawson City, Yukon Territory.
	F	Horse bone	This sample is from an excavation in 2001 in Siberia at one of the Scythian burial sites.
	G	Human bone	This is a bone sample from a young female buried with a neonate in a waterlogged coffin.
	H	Whale bone	This whale bone sample was submitted to the University of Washington Quaternary laboratory in August 1983 and the laboratory entry reads: QL-1857
	I	Whale bone	This bone sample is from the cranium of a whale, species not determined. It was found partly buried at the surface of coarse beach material on a marine beach 12 m above present sea level on Svalbard in 1997
SIRI	B	Mammoth	From North Sea
	C	Mammoth	LQL4

All of this argues for the need to have a sufficient amount of well homogenized materials as part of the design, to ensure that there is sufficient to provide the laboratories with duplicate samples of raw and pretreated materials. Materials we have used include a bulk grain sample from a single year of growth, a bulk tree ring sample (single and multiple rings; see Scott et al. 2019), including both a cellulose sample but also non-pretreated wood, a bulk humic acid and a raw peat.

Benchmarking

A second significant benefit is participation in the intercomparison gives a laboratory an opportunity to benchmark their own operation and while this is a snapshot in time of performance, this should be considered as supplementing internal laboratory QC procedures. Benchmarking allows the laboratory to compare its performance to an external standard, in this case typically the ^{14}C community. Benchmarking also assists in identifying processes that represent best practices. When results are considered as a whole, then laboratories can identify areas of strength or weakness, and can demonstrate the validity of the results (NIST 2015).

Our more recent work (Scott et al. 2003, 2018) has provided z-scores as a means of benchmarking. The z-score is calculated relative to the sample consensus value and incorporates random and systematic uncertainties where

$$\text{z-score} = (x_m - x_A)/\sigma_p \quad (1)$$

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

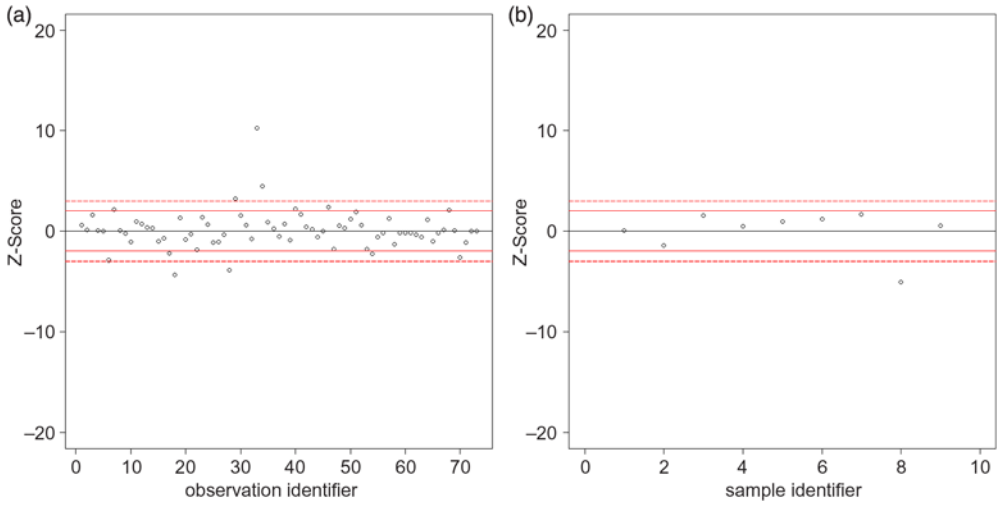

Figure 1 (a) Z-scores for a single sample (D) in SIRI, with warning and action values identified with solid and dashed lines. (b) Z-scores for an individual laboratory with warning and action values identified with solid and dashed lines.

where x_m is the reported result, x_A is the assigned or true value for the material, and σ_p is the target value for standard deviation. This latter value is often set a-priori to reflect the precision needed for a specific application field.

And interpretation of the z-scores follows typically as

- $|\text{z-score}| \leq 2$ satisfactory
- $2 < |\text{z-score}| < 3$ warning
- $|\text{z-score}| \geq 3$ action

An example of z-scores is given in Figure 1(a), showing the z-score for sample D in SIRI and Figure 1(b) for a specific lab across all samples.

Reference Materials

One key benefit which participation in an intercomparison brings to laboratories and which also provides wider benefits to the community (and one which is not time limited) is access to well characterized samples that are typical of the routinely dated materials and span the applied ^{14}C time-scale, and ultimately become what the community recognize as reference materials.

"A reference material is 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. Uses may include the calibration of a measurement system, assessment of a measurement procedure, assigning values to other materials, and quality control" (NIST 2018). Our consensus values provide both a specified activity/age but also the associated uncertainty. The procedures now used to calculate the consensus values have changed from the early studies since, with the predominance of AMS laboratories, we have introduced a linear mixed model approach (Scott et al. 2017, 2018) which appropriately accommodates shared sources of uncertainty (such as more than one result being reported, or a single AMS facility being used by several laboratories, etc). In SIRI, unlike earlier

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

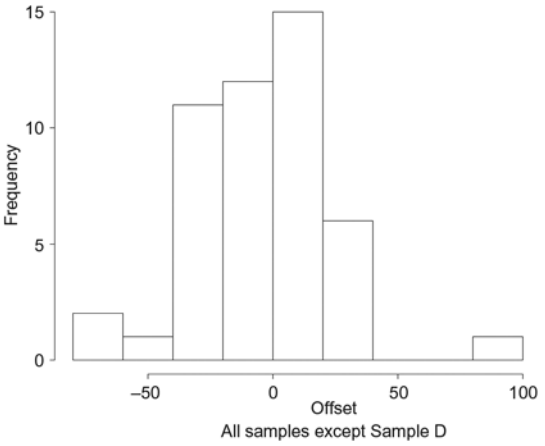

Figure 2 Histogram of the distribution of offsets in yr (BP) in SIRI.

studies, we have used a *random effects model*, this model allows us to include key information on the multiple measurements reported on the same material by each laboratory, which is increasingly relevant since in SIRI, the vast majority of laboratories were AMS.

Historically, we used a robust estimation approach that also takes account of the laboratory quoted error (Rozanski et al. 1992). Examples of some of our current archive of reference materials include: near background bone, background wood, background doublespar, humic acid and barley mash. The values of such reference materials to the community include (1) to help develop new methods of analysis, and (2) to calibrate measurements, institute quality control, and determine performance characteristics. At the end of each study, we have published the consensus values and their uncertainties, and also made known when we have archived material that can be made available.

To Quantify Reproducibility, and Comparability

The fourth benefit is similarly related to benchmarking but explores this in a more specific way, e.g. identifying and quantifying pretreatment effects, different technologies, or sensitivities to effects of background and modern standards. The focus here is on the individual laboratory, but also on the community. In this benefit, we have examined laboratory offsets, identified any significant (statistical) differences that can be attributed to background standards, modern reference standards and pretreatment effects. Laboratory offsets, relative to the consensus values, are expressed as the average laboratory difference from the consensus profile. Figure 2 shows the distribution of laboratory offsets estimated in SIRI (with a mean of –1.5 yr).

Additionally to check measurement uncertainties (where we have not used duplicates), we have used a zeta score defined as:

$$\text{zeta-score} = (x_m - x_A) \,/\, \sqrt{(\sigma_p^2 + \sigma_a^2)} \tag{2}$$

This now incorporates the uncertainty on the reference values σ_a (all other terms are as defined in Equation 1). Interpretation of the zeta-scores is similar to z-scores, and from them, it is

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

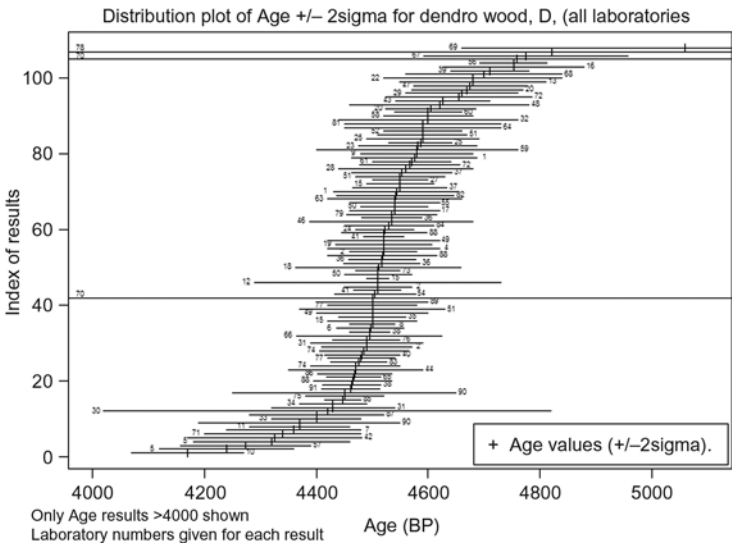

Figure 3 Distribution plot of age ±2 sigma for sample D (FIRI).

possible to evaluate a reduced χ^2 (Steele and Douglas 2006) to quantify any additional variability above that expected given the quoted errors. Visualization is often also shown in a distribution plot such as that illustrated in Figure 3. The shape in such a plot allows us to identify the underlying distribution as well as the spread (or variability).

SPECIFIC STUDIES

In some intercomparisons, we have provided focussed studies including bone (TIRI, VIRI and SIRI) or on background materials (TIRI, FIRI, VIRI, and SIRI). These have allowed laboratories to have access to well described samples in sufficient quantity to benchmark their procedures.

Bone Studies

Substudies have included emphasis on bone samples: apart from the main intercomparisons, we also organized a small cremated bone intercomparison (Naysmith et al. 2007) at the same time as we were running Stage 2 of VIRI. Overall, 8 bone samples have been used, with VIRI stage 2 including only bone samples (see Table 1). Not every laboratory routinely measures bone and there are considerable differences in the pretreatment procedures used. There is also interest in the quality of the bone sample, so that we were also able to study the C/N ratio and pretreatment procedure differences.

Background and Near Background Samples

We have investigated and included a number of both organic and inorganic background and near background samples over the years identified in Table 2. The role that the laboratory background plays in the result and the uncertainty is important (especially in older samples), and this is one area of laboratory practice where there remains considerable divergence in which materials are used and how the background is calculated. We observed

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

Table 2 Background samples used in the intercomparisons.

Study	Sample code	Sample type	Description
TIRI	F	Doublespar	Iceland, from the spar mine, provided from the Museum of Natural History, Reykjavik
FIRI	A	Kauri wood	Subfossil sample from New Zealand
	B	Kauri wood	
VIRI	K	Wood	From Hohenheim (Miocene)
SIRI	A	Wood (VIRI K)	
	C	Mammoth bone	Mammoth bone (Marine Isotope Stage 7; background sample) (Sample LQL4) from Latton Quarry
	K	Doublespar	From Iceland
	L	Wood	From Oregon

instances where the material we provided was better than the in-house background material and also the very clear issues that still exist in how backgrounds are calculated and reported.

For the background samples, we introduced two approaches to summarize the results; In FIRI and VIRI we used a Kaplan-Meier (KM) approach (Scott 2003; Dudley et al. 2016) to estimate the age, dealing with censored age reporting, while in SIRI, we introduced the concept of the limit of background (LoB) (Armbruster and Pry 2008; Scott et al. 2017, 2018). Censored ages mean that the age is reported simply in the form "> BP", and this type of measurement is commonly observed in survival or reliability analysis, where the KM method was first developed, The KM method is a non-parametric method of estimation and the KM survival estimator have been used to estimate the "mean" age (or activity) of the sample whereas in the LoB, we have adopted a different estimation model reflecting the different reporting protocols used by laboratories.

WHERE NEXT?

The series of intercomparisons has delivered a greater understanding of the complexities of ^{14}C dating, provided important benefits to the participating laboratories and communities of users, and created a series of reference materials. However, laboratory quality assurance is not something that stops, and as part of that process, a further intercomparison is being planned. This will be similar to SIRI and will be a single stage study (preliminary name G(lasgow)IRI) of up to 10 materials that will be sourced in sufficient quantity to ensure that there is an archive available to the AMS laboratories to use as reference materials. The age range will span modern to >30K, and will include tree rings, humic acid samples, grain and bone.

CONCLUSIONS

In conclusion, the driving focus for the intercomparison program, now spanning 30 years, has been to provide a process to help laboratories monitor, evidence and improve their quality assurance (and NOT to create a league table or laboratory ranking). This paper has helped enumerate and expand on the multiple benefits of intercomparison participation.

https://doi.org/10.1017/RDC.2019.10 Published online by Cambridge University Press

While intercomparisons are limited in that they provide only snapshots in time, they do allow laboratories to assess their performance, in terms of accuracy and precision, and provide a formal mechanism for benchmarking.

Estimation of between-laboratory variation is essential to the user communities. Having access to well-characterized reference materials allows laboratories to adopt the material as an in-house standard to be run routinely, thus contributing to ensuring realistic uncertainty estimates.

ACKNOWLEDGMENTS

The extensive programs of work have been funded from a number of sources including UK research councils (EPSRC and NERC), EU (FP4), NATO, Historic England and Historic Environment Scotland. Colleagues have been immensely generous in their contributions of the materials which are such an important part of the intercomparison. Finally, a huge thanks to the participating laboratories that have contributed several thousand ^{14}C age measurements over the years.

REFERENCES

Armbruster DA, Pry T. 2008. Limit of blank, limit of detection and limit of quantitation. ClinBiochem Rev 29(1):S49–S52.

Boaretto E, Bryant C, Carmi I, Cook G, Gulliksen S, Harkness D, Heinemeier J, McGee E, Naysmith P, Possnert G, Scott M, van der Plicht J, van Strydonck M. 2002. Summary findings of the Fourth international Radiocarbon Intercomparisons (1998–2001). Journal of Quaternary Science 17(7):633–639.

Bryant C, Carmi I, Cook G, Gulliksen S, Harkness D, Heinemeier J, McGee E, Naysmith P, Possnert G, Scott M, van der Plicht J, van Strydonck M. 2000. Sample requirements and design of an inter-laboratory trial for radiocarbon laboratories. Nuclear Instruments and Methods in Physics Research B 172:355.

Cook GT, Harkness DD, Miller BF, Scott EM, Baxter MS, Aitchison TC. 1990. International collaborative study: structuring and sample preparation. Radiocarbon 32(3):267–270.

Dudley WN, Wickham R, Coombs N. 2016. An introduction to survival statistics: Kaplan-Meier analysis. Journal of the Advanced Practitioner in Oncology 7:91–100.

Gulliksen S, Scott EM. 1995. TIRI report. Radiocarbon 37(2):820–821.

Harkness DD, Cook GT, Miller BF, Scott EM, Baxter MS. 1989. Design and preparation of samples for the international collaborative study. Radiocarbon 31(3):407–413.

Long A, Kalin RM. 1990. A suggested quality assurance protocol for radiocarbon dating laboratories. Radiocarbon 32(3):329–334.

Naysmith P, Scott EM, Cook GT, Heinemeier J, van der Plicht J, van Strydonck M, Ramsey C, Grootes PM, Freeman SPHT. 2007. A cremated bone intercomparison study. Radiocarbon 49(2):403–408.

Naysmith P, Scott EM, Dunbar E, Cook G. 2019. Humics—their history in the radiocarbon intercomparisons studies. Radiocarbon this volume. doi: 10.1017/RDC.2019.11

NIST. 2015. How do you harness the power of benchmarking? Available at https://www.nist.gov/blogs/blogrige/how-do-you-harness-power-benchmarking.

NIST. 2018. SRM definiitions. Available at https://www.nist.gov/srm/srm-definitions.

Otlet RL, Walker AJ, Hewson AD, Burleigh R. 1980. ^{14}C interlaboratory comparison in the UK: experiment design, preparation and preliminary results. Radiocarbon 22(3):936–947.

Rozanski K, Stichler W, Gonfiantini R, Scott EM, Beukens RP, Kromer B, van der Plicht J. 1992. The IAEA ^{14}C intercomparison exercise 1990. Radiocarbon 34(3):506–519.

Scott EM, editor. 2003. The Third International Radiocarbon Inter-Comparison (TIRI) and the Fourth International Radiocarbon Inter-Comparison (FIRI) 1990–2002: results, analyses, and conclusions. Radiocarbon 45(2):135–408.

Scott EM, Aitchison TC, Harkness DD, Baxter MS, Cook GT. 1989. An interim progress report on stages 1 and 2 of the international collaborative program. Radiocarbon 31(3):414–421.

Scott EM, Aitchison TC, Harkness DD, Cook GT, Baxter MS. 1990a. An overview of all three stages of the international radiocarbon intercomparison. Radiocarbon 32(3):309–319.

Scott EM, Baxter MS, Harkness DD, Aitchison TC, Cook GT. 1990b. Radiocarbon: present and future perspectives on quality assurance. Antiquity 64:319–322.

Scott EM, Boaretto E, Bryant C, Carmi I, Cook GT, Gulliksen S, Harkness DD, Heinemeier J, McGee E, Naysmith P, Possnert G, Scott EM, van der Plicht J, van Strydonck M. 2004a. Future needs and requirements for AMS ^{14}C standards and reference materials. Nuclear Instruments and Methods in Physics Research B 223–224:382–387.

Scott EM, Bryant C, Carmi I, Cook GT, Gulliksen S, Harkness DD, Heinemeier J, McGee E, Naysmith P, Possnert G, van der Plicht J, van Strydonck M. 2004b. Precision and accuracy in applied ^{14}C dating: some findings from the Fourth International Radiocarbon Inter-comparison. Journal of Archaeological Science 31:1209–1213.

Scott EM, Cook GT, Naysmith P. 2010a. A report on phase 2 of the 5th international radiocarbon inter-comparison. Radiocarbon 52(2):846–859.

Scott EM, Cook GT, Naysmith P. 2010b. The 5th International Radiocarbon Inter-Comparison (VIRI): an assessment of laboratory performance in stage 3. Radiocarbon 52(2):859–866.

Scott EM, Naysmith P, Cook GT. 2010c. VIRI–summary results and overall assessment. Radiocarbon 52(3):859–865.

Scott EM, Cook GT, Naysmith P. 2017. Should archaeologists care about ^{14}C intercomparisons? Why? A summary report on SIRI. Radiocarbon 59(5):1589–1596.

Scott EM, Cook GT, Naysmith P, Bryant C, O'Donnell D. 2007. A report on phase 1 of the 5th international radiocarbon intercomparison (VIRI). Radiocarbon 49(2):409–426.

Scott EM, Cook GT, Naysmith P, Staff R. 2019. Learning from the wood samples in ICS, TIRI, FIRI, VIRI, and SIRI. Radiocarbon this volume. doi: 10.1017/RDC.2019.12

Scott EM, Harkness DD, Cook GT. 1997. Analytical protocol and quality assurance for ^{14}C analyses: proposal for a further intercomparison. Radiocarbon 39(3):347–350.

Scott EM, Harkness DD, Cook GT. 1998. Interlaboratory comparisons: lessons learned. Radiocarbon 40(1):331–343.

Scott EM, Harkness DD, Cook GT, Aitchison TC, Baxter MS. 1991. Future quality assurance in ^{14}C dating. Quaternary Proceedings 1:1–4.

Scott EM, Harkness DD, Miller BF, Cook GT, Baxter MS. 1992. Announcement of a further international intercomparison exercise. Radiocarbon 34(3):528–532.

Scott EM, Naysmith P, Cook GT. 2018. Why do we need ^{14}C inter-comparisons?: the Glasgow -^{14}C inter-comparison series, a reflection over 30 years. Quaternary Geochronology 43:72–82.

Steele AG, Douglas RJ. 2006. Extending chi-squared statistics for key comparisons in metrology. Journal of Computational and Applied Mathematics 192:51–58.

Taylor BN, Kuyatt CE. 2001. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Available at: http://physics.nist.gov/TN1297 [last accessed 2018/8/31]. Gaithersburg (MD): National Institute of Standards and Technology.

Thompson M, Ellison SR, Wood R. 2006. The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. Pure Applied Chemistry 78(1):145–196.