Why do we need ^{14}C inter-comparisons?: the Glasgow -^{14}C inter-comparison series, a reflection over 30 years.

E. Marian Scott[1], Philip Naysmith[2] and Gordon T. Cook[2]

[1]School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[2]SUERC, University of Glasgow, East Kilbride, UK

Corresponding author: E Marian Scott: marian.scott@glasgow.ac.uk

Abstract

Radiocarbon measurement is a well-established, routinely used, yet complex series of inter-linked procedures. The degree of sample pre-treatment varies considerably depending on the material, the methods of processing pre-treated material vary across laboratories and the detection of ^{14}C at low levels remains challenging. As in any complex measurement process, the questions of quality assurance and quality control become paramount, both internally, i.e. within a laboratory and externally, across laboratories. The issue of comparability of measurements (and thus bias, accuracy and precision of measurement) from the diverse laboratories is one that has been the focus of considerable attention for some time, both within the ^{14}C community and the wider user communities. In the early years of the technique when there was only a small number of laboratories in existence, inter-comparisons would function on an *ad hoc* basis, usually involving small numbers of laboratories (e.g.Otlet et al, 1980). However, as more laboratories were set-up and the detection methods were further developed (*e.g.* new AMS facilities), the need for more systematic work was recognised. The international efforts to create a global calibration curve also requires the use of data generated by different laboratories at different times, so that evidence of laboratory offsets is needed to inform curve formation. As a result of these factors, but also as part of general good laboratory practice, including laboratory benchmarking and quality assurance, the ^{14}C community has undertaken a wide-scale, far-reaching and evolving programme of global inter-comparisons, to the benefit of laboratories and users alike. This paper looks at some of that history and considers what has been achieved in the past 30 years.

Keywords: quality assurance; radiocarbon laboratory inter-comparisons; accuracy; precision

Introduction

In any applied science that makes use of advanced measurement methods, there has been and remains a need for these measurements to be robust and reliable to ensure that:

- the laboratory methods used (both chemistry and instrumental) are appropriate and properly validated;
- the results are traceable and thus linked to internationally recognised standards;
- the composite data sets brought together for activities such as the international radiocarbon calibration initiative are of the highest quality.

These goals are entirely in keeping with widely accepted principles in analytical science, namely systems of analytical quality control (QC) as the fundamental basis for overall quality assurance (QA) and quality management. Part of quality assurance is concerned with establishing and maintaining primary standards and reference materials (with known activities/concentrations) and then the subsequent routine measurement of those standards and reference materials within the laboratory, supported by participation in inter-laboratory trials or proficiency tests. Proficiency testing is widely used in the analytical chemistry communities, based on proficiency trials (also known as inter-laboratory comparisons). Within the ^{14}C community, as the number of laboratories grew, there was discussion about what the community of laboratories required. Long and Kalin (1990) stressed that it was incumbent on individual radiocarbon laboratories to engage in a formal programme of quality assurance (QA) while Polach (1989) noted that the opportunity for internal checking by individual laboratories in routine ^{14}C measurement was hampered by a lack of suitable quality control (QC) and reference materials. The series of radiocarbon inter-comparisons (SIRI and its predecessors), which are the basis of this review, are examples of proficiency trials, which have developed with the community to address the needs and as is common in the analytical chemistry community, they have followed standard protocols which have evolved dependent on the specific trial design and objectives. Fundamentally, participation in proficiency trials helps ensure the results from a laboratory are meaningful. Participation should also contribute to and enhance a laboratory's reputation and form part of a laboratory-based quality assurance programme and at the same time, support the development of reference materials.

The work reported here describes our efforts over a 30-year period to deliver suitable reference materials and to provide the radiocarbon community with regular opportunities to benefit from independently organised laboratory inter-comparisons. These inter-comparisons have allowed the evaluation of any laboratory bias or offsets (the difference between the measured and 'true' ages) and also provided an independent measure of laboratory precision (the scatter in the measured ages) relative to the known or estimated ages for the

reference samples, as well as providing laboratories with a benchmarking activity as a part of their quality assurance.

While the major emphasis in their organisation has been to deliver benefit to the laboratory, at the same time, though not with the equal emphasis, they have also delivered indirect benefits to the users of ^{14}C measurements who have perceived a need for assurance of the comparability and quality of dates, especially as increasingly sophisticated modelling of assemblages of dates is being used to provide insights into cultural interactions and past environments. For many users, the laboratory procedures are complex and 'black box', and as a result, their quality may be difficult to judge. In this way, laboratory participation in such trials offers users an independent and verifiable performance check.

What does an inter-comparison deliver and equally important, what does it not deliver?

First, an inter-comparison provides a snapshot in time and an independent check on laboratory performance relative to comparable facilities, and to any absolute standards that might exist. However, since they are not organised continuously (primarily so as not to burden the laboratory), they do not provide a continuous quality assurance check and may therefore detect only issues current at the time of the trial. This is the major difference in terms of quality control (within the laboratory and which should be a routine and continuous process) and external quality assurance. While many laboratories have participated over the years in the entire radiocarbon inter-comparison series (which occur at 3-4 year intervals), and so their performance could be studied over time, this is not a key objective, but more an additional benefit.

An inter-comparison also has the potential to create new reference materials. While the radiocarbon community has devoted considerable care and effort to establishing and maintaining primary standards (currently SRM-4990C - Oxalic Acid ll) and reference materials such ANU-sucrose (also known as IAEA-C6), Chinese-sucrose and the IAEA C1-C6 series (Rozanski *et al.*, 1992), augmented by additional oxalic acid samples (IAEA C7 and C8) (Le Clerq et al, 1998), our inter-comparisons have generated a suite of natural materials such as humic acid and cellulose which are routinely dated, and whose ages have been estimated from the large numbers of measurements made by many laboratories (Scott et al, 2004). These materials have then been made available to new laboratories, or for the commissioning of new systems, which again provides an additional benefit to the ^{14}C community.

In none of the trials have we created "a league table of laboratories". We have also resisted removing the anonymity in our reporting since our ethos has always been that our work is for the benefit of the laboratories. However, we have always published a list of participating

laboratories and this will continue, so that users are made aware of which laboratories have taken part, and thus armed with this information they can seek information from any laboratory they wish to commission to date samples. We have also written and spoken quite widely about the quality assurance questions that users should ask a laboratory. In this way, we believe that the laboratory and user relationship is strengthened. In other scientific communities, laboratory accreditation schemes have been introduced which formalise criteria for evaluating laboratory performance, and often involve considerable paperwork and bureaucracy. This has not been our intention.

Over a 30-year period, there have been significant changes in the community of laboratories and in the technical challenges (smaller and older samples being dated), so we are able to reflect on how our programme of work has evolved and responded to these changes. In the following sections we review the key aspects of a proficiency trial, namely; materials used, design of the trial, analysis of the results, and definition of a consensus value and its uncertainty. Finally, we look forward to what remains to be done in this ongoing QA process.

What makes a proficiency trial?

According to Thomson et al. (2006), the preparation and validation of test material is the first step in devising a proficiency trial. Then, the participants are recruited and the materials distributed to them for analysis. Subsequently, the results are reported back to the organisers who then statistically analyse them and inform participants of the results. In the following sections we describe some ^{14}C specific considerations of these three stages and how they have changed over time in the inter-comparisons we have organised.

Sample (test material) choice

While the format of inter-comparison studies organised by the Glasgow University group has changed over the last 30 years, the main criterion for selecting the samples has remained constant - namely wherever possible, they should be natural and routinely dated materials that have the potential to become internationally recognised reference materials (Scott et al, 2004). For quality control purposes, a reference material is commonly a natural material that behaves as similarly as possible to the routinely measured samples. Therefore, we have worked to create a suite of reference materials that are representative of routinely dated materials and whose expected ages cover the full range of the applied ^{14}C timescale. Once such materials have been through the inter-comparison process they have a ‘certified’ result that can be used in future laboratory quality control programs.

For our programmes, the main criteria for selecting potential materials are shown below. These are ideal and not always achievable, but we have strived to meet them wherever possible:

1) There should be a sufficient quantity of material available to meet current and future requirements.
2) They should be of archaeological and/or geological interest.
3) They should cover the broad spectrum of laboratory experience (age, sample type, etc).
4) They should satisfy rigorous homogeneity testing.
5) Wherever possible, known age material should be used.

The majority of the samples used in our inter-comparison studies have been identified, collected and provided to us by members of the ^{14}C scientific community, to whom we are extremely grateful. When we are deciding on what samples are to be included in any inter-comparison study, one of the main requirements is homogeneity. Homogeneity of the reference material refers to the variability observed in true replicates (sub-units). Therefore, for some materials, this may require the material to be homogenized and for many of the samples we have selected, homogenisation has been a major undertaking, given the quantity of material needed to fulfil criterion 1) above (we have described these procedures in detail in the following section). Our proficiency trials have included the following sample types: peat (whole peat, humin and humic acid extracts), bone, wood (whole wood and cellulose), carbonate and grain. Their ages have spanned the period from modern to background. In only one (a very early study, the ICS (Scott et al, 1990)), we created artificial samples of carbonate and benzene of known activities. One major challenge in the earlier studies was that the majority of participating laboratories were radiometric (using either liquid scintillation or gas proportional counting) and for such laboratories, the mass of sample required for a single analysis was of the order of several grams, which meant that a large quantity of material needed to be sourced to meet current and future needs. In such cases, the issue of homogeneity was of particular concern and consequently, pre-treatment had to be carried out on a large scale. We were sometimes challenged (notably in an algal carbonate sample) to demonstrate homogeneity. In other cases, the sampling of the material in the environment was constrained e.g. in the case of peat samples from Iceland and Scotland, these were sampled from narrow depth horizons to provide a sample of restricted age range and thereafter we relied on grinding to a relatively uniform particle size and extensive mixing to achieve homogenisation. In the case of humic acid, grinding and mixing was not required to the same degree as the humic acid was extracted in solution and was subsequently precipitated (the solution phase providing the homogenisation). Wood is

a material of choice, especially dendro-dated tree-rings. Typically, finite-age wood samples came from a known set of tree–rings (10 or 20), while for infinite age wood samples, the number of rings might be considerably more than 20. In some cases, we have used cellulose (which has gone through substantial mixing to produce a homogenised sample) as the preferred form for distribution to laboratories. As the proportion of AMS laboratories participating in the trials has grown, we have also introduced single ring samples. For these and grain (again a single year of growth), we have used no homogenisation other than careful mixing. Bone samples were derived from single bones that were sub-sampled, so we would expect them to be homogeneous. Shell, charcoal, archaeological grain and algal carbonates have all proved more challenging, and we have relied very much on the provenance of the samples to constrain any heterogeneity with lesser or greater success. In one exercise (FIRI), we were able to carry out a series of formal homogeneity tests before the trial began, and it is also the case that routinely, a number of measurements are made on each material in the SUERC laboratory before they are accepted for inclusion in the trial. With regard to samples of independently known age, we have been very fortunate in being able to access tree-ring samples that have been dendro-dated, thereby providing an independent age control. Where independently known age material was not available, in some cases we have also had access to samples that have been previously dated. We have also been extremely fortunate in accessing materials that are of general scientific interest, e.g. a charcoal sample from the European Palaeolithic site, Chauvet Cave, that was used in SIRI (Scott et al, 2017).

Regarding the labelling of samples; in the very early studies, such as ICS, each laboratory received a randomly labelled set of samples (in this way laboratories could not immediately identify samples in common), and laboratories were also randomly labelled. In later trials, and given the strong imperative to create internationally recognised reference materials, each sample received a unique id (typically a letter of the alphabet prefixed by the trial acronym).

Table 1 summarises all the samples used in the past 4 inter-comparisons, whether they have been used in previous inter-comparisons and the pre-treatment procedure used.

Study	**Sample Code**	**Sample Type**	**Pre-treatment**
TIRI	Sample A	Barley mash	Air dried (35°C) and mixed
	Sample B	Belfast pine Q7780	None
	Sample C	IAEA Cellulose **(IAEA C3)**	None
	Sample D	Hekla peat Iceland	Air dried (35°C) and mixed
	Sample E	Ellanmore peat humic acid fraction	Humic acid extraction

	Sample F	Icelanadic doublespar	None
	Sample G	Fugla Ness wood	Acid, alkali,alkali,acid
	Sample H	Ellanmore whole peat	Air dried (35°C) and mixed
	Sample I	Caerwys Quarry Travertine	None
	Sample J	Buiston Crannog wood	None
	Sample K	Turbidite carbonate	Oven dried (50°C), ground and mixed
	Sample L	Whalebone (Norway)	None
	Sample M	Icelandic whole peat	Air dried (35°C) and mixed
FIRI	Sample A	Kauri wood New Zealand	None
	Sample B	Kauri wood New Zealand	None
	Sample C	Marine Carbonate (**TIRI K)**	Oven dried (50°C), ground and mixed
	Sample D	Belfast wood Q7780	None
	Sample E	St Bees peat humic acid fraction **(FIRI E)**	Humic acid extraction
	Sample F	Belfast wood Q7780	None
	Sample G	Barley mash	Air dried (35°C) and mixed
	Sample H	German wood	None
	Sample I	Belfast cellulose Q7780	Cellulose extraction
	Sample J	Barley mash	Air dried (35°C) and mixed
VIRI	Sample A	Barley mash	Air dried (35°C) and mixed
	Sample B	Grain (Israel)	None
	Sample C	Barley mash **(FIRI G)**	Air dried and mixed
	Sample D	Grain Israel	None
	Sample E	Mammoth bone	None
	Sample F	Horse bone (Siberia)	None
	Sample G	Human bone	None
	Sample H	Whale bone	None
	Sample I	Whale bone	None
	Sample J	Humic acid (Siberia)	Humic acid extraction
	Sample K	Wood (Hohenheim)	None
	Sample L	Wood (Belfast)	None
	Sample M	Wood (Loch Tay)	None
	Sample N	Wood (Loch Tay)	None
	Sample O	Wood (Cambridge)	Cellulose extraction
	Sample P	Charcoal (Mexico)	None
	Sample Q	Charcoal (Iceland)	None
	Sample R	Murex Shell (Israel)	None
	Sample S	Barley mash **(VIRI A)**	Air dried (35°C) and mixed
	Sample T	Scottish Peat humic acid fraction	Humic acid extraction
	Sample U	St Bees peat humic acid fraction **(FIRI E)**	Humic acid extraction
SIRI	Sample A	Wood (Miocene Hohenheim Germany)	None
	Sample B	Mammal bone (North Sea)	None
	Sample C	Mammoth Bone (LQL4)	None
	Sample D	Barley mash	Air dried (35°C) and mixed

	Sample E	Wood (New Zealand)	None
	Sample F	Wood (Belfast)	None
	Sample G	Wood (Belfast)	None
	Sample H	Wood (Belfast)	None
	Sample I	Wood (Arizona)	None
	Sample J	Charcoal	None
	Sample K	Doublespar (Iceland)	None
	Sample L	Wood (Arizona)	None
	Sample M	Wood (Scottish Crannog)	None
	Sample N	Scottish peat humic acid fraction (**VIRI T**)	Humic acid extraction

Table 1. Inter-comparison sample types, and pre-treatment (pre-treatments are discussed below)

Pre-treatment procedures to prepare bulk trial samples

Pre-treatment is a critically important part of the dating method, its main function being to remove extraneous, non-contemporary carbon. Methods of pre-treatment vary across materials and in laboratory practise. A brief description of the pre-treatment methods used in the preparation of our reference materials is given below.

Whole Peat/Humic acid extraction. Well-humified peat samples were collected from freshly cut exposures (about 20 cm depth to provide limited age variation). The raw samples were air dried and sieved through a 3 mm mesh to remove large root fragments, oven dried and mixed by several passages through a grinding mill. If whole peat and humic acid were required then half of the product was retained in this form and mixed further. To obtain the humic acid fraction, the remainder was subjected to successive digestions in 2M potassium hydroxide and the alkali-soluble humic acid extracts were removed by filtration and combined. The humic acid was then precipitated from the bulk solution by adjusting to pH3 with sulphuric acid. The resulting humic acid slurry was separated by centrifugation, re-bulked, washed several times with distilled water and oven dried at 70^0C. The resultant granules were washed with warm distilled water, filtered and dried to constant weight. The final product was again subjected to physical mixing. The alkali-insoluble (humin) residues from the extraction were also recovered and retained for future reference (Harkness et al., 1989).

Whole Wood: Many of the samples came from dendro laboratories and were simply cut into suitable sized fragments for distribution. For others, the samples were digested in 0.5M KOH at 80^0C, soaked in distilled water to remove excess alkali and then digested in hot 2M HCl.

Finally, the wood was again soaked in distilled water to remove excess acid and dried to a constant weight in a vacuum oven.

Cellulose: The wood was either chopped into small pieces, or shavings were produced using a power plane. The material was then subjected to repeated digestion in 2M potassium hydroxide, washing, acidification and bleaching in sodium chlorite/hydrochloric acid solution. The fibrous extract was washed free of chlorite with distilled water, oven dried at 40^0C and thoroughly mixed by tumbling.

Barley mash: Bulk samples were taken from single fermentation vats, and therefore, were already very well mixed in the industrial process. The material was immediately oven dried to avoid the possible development of mould growths and was finally subjected to physical mixing.

Reference materials

One of the most valuable outcomes from the trial is the archival materials, which are now well characterised by the community and can thus function as new reference materials. A reference material is commonly a natural material that behaves as similarly as possible to the samples being measured. To ensure the widest possible practical advantages, the materials should also be representative of routinely dated materials and their ages should span the full range of the applied ^{14}C timescale. These materials are typically certified on the basis of a laboratory inter-comparison, therefore, when selecting samples for an inter-comparison, their dual purpose must be considered. Given the importance of ^{14}C dating in chronology construction, ideally some of the samples should be independently dated. The most appropriate material for this purpose is dendrochronologically dated tree-ring sequences, which are already used to underpin the absolute calibration of the conventional ^{14}C timescale back to approximately 13900 years before present. Also, because of the considerable use of ^{14}C dating within routine archaeological investigations, reference materials of archaeological significance are valuable. Further, given the long history of inter-comparisons in the ^{14}C community, it is also important that samples should link any new inter-comparison to past studies. In this way, continuity of laboratory performance can be assessed. Table 1 shows the extensive set of reference materials and their various forms.

Inter-comparison design

The design of our studies has varied over the years, dependent on the specific scientific questions we have been exploring. Some of the studies have been multi-stage, others only

single stage. In some, we have provided pre-treated samples, as well as the raw material. We have always included *linking* samples so that a small number of test materials have appeared in more than one inter-comparison, providing a thread through time and not just a single snapshot. Occasionally we have provided duplicate samples (but blind to the laboratories) to assess precision (especially with regard to the laboratory's quoted errors). In the hierarchical, multi-stage studies, our goal was to understand and quantify the contribution of the various pre-treatment and sample preparation stages to the variation observed in the results (components of variation). In the single stage studies, the goal was to describe the variation in the final results. In some of the trials, we have focussed on specific materials such as bone, while in others the focus has been on specific age ranges (*e.g.* background or close to background samples).

Always, we have aimed to provide consensus values for the materials, with an estimate of uncertainty on that value, thus enhancing the future value of the materials to the laboratories.

Anonymity or not

A conscious decision was made from the very first trial (ISG, 1982), that we would not identify individual laboratory results; however, we agreed that we would publish a full list of participating laboratories. Our reasoning was clear, namely that our programme initially would be for the benefit of the participating laboratories, and the benefits gained would then indirectly benefit the user community. Through publishing a list of participating laboratories, users and funders would be able to quickly see which laboratories had participated and would then be able to work first hand with the laboratory to understand results and the evidence of quality and confidence in the results.

Laboratory pre-treatment procedures and other ancillary information

Where we have provided the natural form of the material, we have not been prescriptive in the method of pre-treatment to be used, we have however requested that the laboratory provide details of the methods used. Similarly, we have also asked for information concerning the laboratory primary standard and background materials used, and other relevant laboratory information. We have always requested that $\delta^{13}C$ values be reported and for bone samples we asked for the $\delta^{15}N$ value and the C/N ratio. These additional pieces of information have formed part of the subsequent analyses to explore sources of variation in the results.

The time line of trials

There have been many small scale inter-comparisons typically involving a small number of laboratories or on a very specific topic (eg cremated bone, Naysmith et al., 2007, infinite age bones, Cook et al., 2012) and tree rings (Hogg et al., 2013). However, in this section we focus on the global inter-comparisons where an open request to all laboratories to participate was issued, starting from the International Study Group (ISG (1982, 1983) and the International Collaborative study (ICS) in 1988 (Cook et al., 1990, Harkness et al., 1989, Scott et al., 1989, 1990, 1991)). Here, we consider the different study designs used.

ISG (1982, 1983): this very first Glasgow University led trial started in 1979 and involved 20 laboratories who each received a series of 8 tree-ring samples from a single section dated to 5100 ^{14}C years BP approximately. Samples were identified on the section of wood that was provided, each sample being equivalent to 10 tree- rings, with the entire section spanning two hundred years in total. The goals were to understand the relationships between the observed variation in the results, between the tree-ring blocks (known separation) and across the laboratories, in relation to the routinely quoted age error. Laboratories were asked to extract cellulose from the wood.Three high precision radiometric facilities participated. Laboratory offsets and error multipliers were estimated from the results

ICS (Cook et al, 1990, Harkness et al, 1989, Scott et al, 1989, 1990, 1991): In this trial, one of the goals was the quantitative assessment of variability and its attribution to the processes of counting, sample synthesis and pre-treatment. To achieve this goal, we designed a study with three stages, with duplicate samples provided at every stage. The hierarchical study ran over a 4 year period. Stage 1 investigated only the counting procedures, with prepared carbonate and benzene samples being supplied to laboratories. This was the only study where we used 'artificial' samples. Gas counting laboratories still needed to prepare the counting gas from the carbonate, but for liquid scintillation laboratories, the benzene samples required no further pre-processing in the laboratory. In Stage 2 we provided homogenised, pre-treated natural samples, so that sample synthesis and counting were assessed. Finally, in Stage 3, non-pre-treated materials were supplied in order to assess all aspects of the dating process (sample pre-treatment, sample synthesis and counting). Eighty laboratories were invited to participate, with a total of 52 returning results, of which only 8 were AMS.

Neither the ISG nor ICS created archival reference materials so that the analysis did not report consensus values for the samples but rather focussed on laboratory performance (offsets and error multipliers).

Following the ICS study, **TIRI** (the Third International Radiocarbon Inter-comparison) (Scott et al., 1992, Gulliksen and Scott, 1995, Scott 2003) was organised and commenced in 1991.

This study had one compulsory stage (*i.e.* laboratories received the same suite of samples) and a second stage where laboratories could choose from an optional suite of samples, reflecting the fact that some materials are more specialised than others. 67 sets of results were reported, including 11 from AMS facilities.

Simply stated, the aims of TIRI were:

1. To function as the third arm of the quality assurance procedure.
2. To provide an objective measure of the maintenance and improvement in analytical quality.
3. To assist in the development of a 'self-help' scheme for participating laboratories.

The next study in the sequence was **FIRI** (the Fourth International Radiocarbon Inter-comparison) which was completed in 2000 (Scott et al, 1997, 1998, Bryant et al, 2000, Boaretto et al, 2002). This again was a single stage study, but some samples were provided in duplicate (this fact was blind to participants). Again, there were some more specialised materials provided. The number of AMS facilities participating increased to just over 30, from a total of just over 60 participating laboratories. Our goals were to:

- Demonstrate the comparability of routine analyses carried out by both AMS and radiometric laboratories;
- Quantify the extent of, and sources of, any variation;
- Investigate the effects of sample size, pre-treatment and precision requirements on the results.

The Fifth International Radiocarbon Inter-comparison (**VIRI)** (Scott et al, 2010a,b,c) commenced in 2004 and continued the traditions of TIRI and FIRI but was designed to address some of the criticisms of TIRI and FIRI, including the need for the measurements to be made over a relatively short period of time (hence the workload within the laboratory is compromised)and the fact that they provide only a snap-shot in time. However, VIRI also retained some of their important features, namely, using natural samples and ensuring the anonymity of participating laboratories to prevent the creation of laboratory league tables. VIRI was a 4-year project, with samples distributed in three sets, roughly a year apart. The first suite was grain; two modern and two archaeological samples; the second suite was bone (with several samples close to background in activity), while the third and final suite (12 samples in total) included charcoal, wood, grain, shell and humic acid. For the first time, some samples were included that were specifically for AMS facilities, reflecting the changing laboratory demographics.

Its aims and objectives were:

- to demonstrate the comparability of routine analyses carried out in radiocarbon laboratories
- to investigate the effects of sample pre-treatments
- to quantify the extent and sources of variation in results

A significant concern related to bone, which is an increasingly important material for dating. Therefore a number of bone samples were studied in detail, including the effect of different pre-treatment methods. More than 70 laboratories participated of which over half were AMS.

The most recently completed exercise is **SIRI** (the Sixth International Radiocarbon Inter-comparison), which commenced in 2013 and was completed in 2016. Again, this was a single stage trial, designed predominantly for AMS facilities. A total of 13 samples for AMS laboratories and 5 samples for radiometric laboratories (4 of which were common to AMS samples A, B, D and K) were sourced. A set of single, dendro-dated tree rings were also included. The materials had a range of ages (modern, a few thousand years, approx. 40,000 years and background).

The aims and objectives of SIRI were:

- to demonstrate the comparability of routine analyses carried out in radiocarbon laboratories
- to quantify the extent and sources of variation in results
- through choice of material to contribute to the discussion concerning laboratory offsets and error multipliers in the context of IntCal (the International Calibration Programme).
- to gain a better understanding of differences in background derived from a range of infinite age material types.

Background samples formed a specific focus in SIRI with 5 such samples, including bone.

From these brief descriptions, it is clear that historically the trials have had several objectives in common, but also that there have been nuanced developments. As the laboratory community has become more mature, the trials have tended to become simpler in their organisation (single stage) but more sophisticated in the materials, with specific aspects being examined (background materials), and this reflects the evolution of community need.

In the early studies, when there were very few AMS facilities routinely dating material, they received the same samples as the radiometric laboratories, but were asked to carefully consider the sampling they did from the original bulk material. Latterly, the samples were

prepared with AMS laboratories in mind and only a limited set of samples was prepared for radiometric laboratories- meeting the differing requirements of AMS and radiometric laboratories in a single trial, using a single set of materials, proved very challenging.

Statistical analysis of results

The objectives of the various inter-comparisons have followed a similar evolution to the design. There are some objectives that have not changed, but some which are specific to a given study. Common objectives include (1) an assessment of the comparability of results reported by different laboratories; (2) the extent of any inter-laboratory variability and, where significant, an assessment of the possible cause(s), such as the pre-treatment method used or the material employed to determine background activity and (3) defining consensus values for the test materials to ensure that they can be used more generally as reference materials. Some specific objectives included: a direct comparison of the radiocarbon data generated by radiometric and AMS laboratories (one which was of high profile in the very early studies) and the possible influence of sub-sampling on AMS results, or as in SIRI, understanding the laboratory background. Occasionally, we have also focussed on a specific material such as bone. These objectives then informed the statistical analysis that was carried out.

Ultimately, the key measurement properties of any method are accuracy and precision and we have used a variety of statistical models to examine these at an individual laboratory level including error multipliers, bias estimates and z-scores. We have also used several models to evaluate the consensus values and uncertainties on those samples chosen to be reference materials. All of the studies have been concerned in one way or another with the underpinning nature of the routinely quoted laboratory age uncertainty and its relationship to the observed variability.

Accuracy is the closeness of agreement between a measurement and the true or reference value. If we imagine a series of measurements, each with the same true value, then if the average of the measurements does not equal (within error) the true value, then the measurement is said to be biased, where the **bias** is the difference between the **expected value** or average of a large series of measurements and the true value. Bias is usually considered to be a systematic error. Thus, in a number of the trials, we have estimated individual laboratory biases (and uncertainties on such quantities) relative to either the known age or the assigned consensus value that has also been estimated for each material, using a standard protocol.

Precision is the closeness of agreement between a series of independent measurements obtained under identical conditions. Precision depends on the **distribution of random**

errors and is commonly estimated by the standard deviation of the results. We have used both duplicate samples (when incorporated in the study design) and the suites of results to provide measures of laboratory precision, typically in the form of error multipliers (in ISG and ICS we introduced internal and external error multipliers (ISG, 1982, 1983, Scott, 1983) which were also used in the IAEA inter-comparison (Rozanski et al., 1992).

The use of a laboratory error multiplier was introduced at a very early stage in the studies (Gulliksen & Scott, 1995, Scott, 2003, Scott et al., 2007) as a very simple means of exploring the variability in a set of results and its relationship to the quoted errors. We used two forms; in the internal error multiplier form (IEM), the laboratory quoted error is increased (or decreased) by a multiple which is estimated, typically from a series of replicate measurements. The error multiplier captures sources of variation in the estimated ^{14}C age that are not accounted for in the quoted error. In the first case, a theoretical Gaussian model for the radiocarbon measurement X is that $X \sim N(\mu, \theta^2\sigma^2)$ where μ, θ, and σ^2 are unknown. A series of replicate ^{14}C measurements are made, denoted by $x_1,\ldots,x_n$ with quoted errors $s_1,\ldots, s_n$ where i=1,…n , the theoretical model is interpreted as meaning that each measurement has the same true ^{14}C age, μ, but that the population uncertainty is $\theta\sigma$, where θ is the error multiplier. The estimate of μ is the weighted average $\hat{\mu}$ (the weights being proportional to s_i) (equation 1) of the measurements, and the estimate of the error multiplier θ (equation 2), $\hat{\theta}$ is given by

(1) $$\hat{\mu} = \frac{\sum_1^n \frac{X_i}{s_i^2}}{\sum_1^n \frac{1}{s_i^2}}$$

(2) $$\hat{\theta}^2 = \frac{1}{n}\sum_1^n \left(\frac{X_i - \hat{\mu}}{s_i}\right)^2$$

Or $$\hat{\theta} = \sqrt{\frac{1}{n}\sum_1^n \left(\frac{X_i - \hat{\mu}}{s_i}\right)^2}$$

It is increasingly common that laboratories will make a long series of replicate measurements on a reference material and assess the standard deviation of the set (closer to the conditions needed for the estimation of the reproducibility standard deviation). If the standard deviation is greater than the quoted errors on the individual estimated ages, then this would suggest an

unaccounted source of variation and a laboratory might chose to quote the larger of the standard deviation of the set and the individual quoted error, or they might use an error multiplier approach to provide a more realistic uncertainty.

This common error multiplier approach uses a multiplicative model. However, as an alternative, an additive model could also be used, such that the overall uncertainty is $\sqrt{(s_i^2 + \tau^2)}$.

Our description of an internal error multiplier can also be developed for the case where we have duplicate samples. In this case, we have also chosen to calculate standardised deviations (as the difference between the duplicate results divided by the error on the difference). Such quantities should have a Normal (or Gaussian) distribution with mean zero and variance 1, which when squared should have a Chi-squared distribution with 1 degree of freedom, **if** the two results are in agreement and the quoted errors represent all sources of variation in the results. This forms the basis for a test of agreement between duplicates.

In the external error multiplier (**EEM**) (more pertinent for the inter-comparison context), the calculations are based on all the results provided by an individual laboratory in the trial (so across all materials). This is described more fully below since it depends on the consensus value.

The analysis of the overall results also follows fairly standard procedures, based on first the evaluation of the assigned or consensus value (the ^{14}C age for each material), followed by laboratory offsets and external error multipliers, and then subsequently, other measures of performance such as the z-scores.

Dealing with outliers

In every trial, we have identified occasional (less than 5% typically in total) anomalous or outlying values. The method of identification is a standard statistical approach, using relatively simple criteria based on the inter-quartile range and distance from the median. More formal analysis could be used but has not been pursued. The influence of such data values depends on the analysis being performed, since any anomalous value will typically increase the overall variability and could also impact the consensus age calculations. It has also been clear that occasionally, the outliers come from the same laboratory, which is then reported to the laboratory.

Calculation of consensus values

Each material needs to be characterised by estimating its activity, which creates a reference value. This value can be considered as the 'known' activity of the material and so future analyses can be compared to this to quantify the accuracy of the measurement. In this way, the material remains useful for laboratory quality assurance.

The procedure used in the calculation of the consensus value is an iterative one, described below and first described in Rozanski et al. (1992).

There are three stages.

- **Stage 1**: Outlying results are removed if they are greater than 3 inter-quartile ranges from the nearest of either the lower or upper quartiles, *i.e.* when a result is either greater than Q3 + 3(Q3-Q1) or less than Q1 - 3(Q3-Q1), where Q1 and Q3 are the lower and upper quartiles of the distribution of results, respectively. The preliminary consensus value is calculated as the median (m) of the remaining results.
- **Stage 2**: Remove results that are at least twice their quoted error (s) from the preliminary consensus value. That is, only keep $|x-m|/s < 2$, where x is the result, m the preliminary consensus value and s the quoted error.
- **Stage 3**: Calculate the final consensus value as a weighted mean of the remaining results, using their s^2 values as the weights.

This approach, when first used, was rather controversial since results from all laboratories were being used in the first stage and only in the second stage were some results removed, based purely on their closeness (relative to their quoted errors) to the preliminary consensus value. This was an approach where neither laboratory reputation nor absolute value of the quoted error contributed to the consensus value estimation.

The uncertainty on the consensus value has also been calculated and reported at 1σ. This is an important measure of the error in the consensus value and should be used in the assessment of laboratory agreement with the consensus value

More recently, we have also adopted a second model (Scott et al., 2017) using a linear mixed effect model that reflects the more common case where AMS facilities are likely to report several determinations for the same sample. This allows the common sources of variation to be correctly attributed. To estimate the consensus value for each material, using the linear mixed (or random effects) model, the total variation in F (or age) around the 'true' age for the material is attributed to two components, the within laboratory (common source of variation), and the between laboratory variation.

With the evaluation of the consensus value for a material, we are able to evaluate a laboratory bias or offset term as the average deviation (weighted) from the consensus value. In the case of known age samples, the offset is estimated relative to that value.

For assessment of performance for an individual laboratory, we have assumed that the consensus values (or the dendro-dates for the known age samples) can be treated as the 'true' age/activity (equation 3,4). For each laboratory, an offset (or bias) relative to the consensus values and an error multiplier can be calculated. Error multipliers were used in ICS, TIRI and FIRI but not in VIRI.

The measurement model for the results from a laboratory underlying this estimation equation is:

(3) $$X_i \sim N(\mu_l + \alpha, s_i^2)$$

where X_i is the ^{14}C age for sample i, μ_i is the true age and s_i is the quoted error for i=1,…, n the number of reference materials.

In FIRI, the laboratory offset was defined as the average laboratory difference from the consensus profile (μ_i). The model used assumes that for a given laboratory there is a potential systematic offset, α from the consensus profile, which we can estimate. The form for α is that of a weighted average of the standardised deviations.

(4) $$\alpha = (\Sigma(x_i - \mu_i)^2/s_i^2)\Sigma(1/s_i^2)$$

We can then test whether α is plausibly zero (i.e. that the laboratory is, on average, accurate).

In TIRI, a laboratory external error multiplier was estimated for those laboratories where there was no evidence against the null hypothesis that α was plausibly zero (equations 5 and 6).

The measurement model in this case is that

(5) $$X_i \sim N(\mu_i, \theta s_i^2)$$

where X is the ^{14}C age for sample i, μ_i is the true age, s_i is the quoted error and $\sqrt{\theta}$ is the error multiplier.

(6) $$\theta = \Sigma d_i^2 / J$$

where d_i is $X_i - \mu_i$ and J is the number of results submitted by the laboratory.

In VIRI, we introduced a further measure of performance that is more internationally used in the analysis of proficiency trials, namely z-scores (Thompson et al., 2006)

z-scores

The analyses of the results for proficiency tests follow fairly standard procedures; evaluation of the assigned value (e.g ^{14}C age) and measures of laboratory performance, typically based on z-scores derived using one key quantity σ_p. Interpretation of z-scores includes accuracy and precision and 'fitness for purpose'.
For the analysis, we have reported z-scores (equation 7) calculated as

(7) $$Z = (X_M - X_A)/ \sigma_p$$

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 the standard deviation for values of X. We have used the laboratory quoted error for σ_p though hypothetically it is determined by fitness for purpose and represents the amount of uncertainty in the results that is tolerable in relation to the purpose of the analysis. X_A may be known or assessed as the consensus value. Interpretation of the z-score reflects the accuracy achieved and provides a means of making a judgement concerning fitness for purpose.

It is commonly assumed that z should be normally distributed with zero mean and variance 1, where

- A z-score of 0 implies a *perfect* result
- A z-score between –2 and +2 is generally considered as complying with fitness for purpose
- A z-score outside –3 or +3 would be very unusual and further investigation would be needed.

All of the results, whether in the form of laboratory bias, error multipliers, or z-scores, can then be set out visually, in a variety of ways to the individual laboratory or as a summary of the 'population' of laboratories.

Background samples

Background (or blank) samples have always played an integral part in our inter-comparisons but they have also proved challenging for laboratories for a variety of reasons and for the

subsequent statistical analysis and reporting. They are especially critical as the radiocarbon timescale is pushed back beyond 45,000 years. Background samples are those for which there is no measurable ^{14}C activity and therefore, to find natural background samples of bone, wood, shell, peat etc is difficult, but we have been successful on a number of occasions, *e.g.* with the doublespar (carbonate), Hohenheim woods and mammoth bone (Table 1). We have also provided close to background samples on a number of occasions (eg Kauri wood). Background and close to background samples provide a measurement challenge to the laboratories as small residual contaminants may result in diverging values, and there is the additional challenge in the reporting of background and close to background results. Radiocarbon reporting and calculation conventions for background samples are less well defined than for the routine reporting of a ^{14}C age and as a result, laboratories have reported background or near background results in a variety of formats, taking into account different measurement and laboratory aspects. In SIRI, we provided a calculation template for the 4 background samples and asked laboratories to report three quantities- F_m ,the measured fraction modern with fractionation applied to both the sample and standard, **but no correction for background,** f the measured fraction modern of a background sample and finally F, which is F_m corrected for background. The actual format of reporting results for the 4 background samples varied considerably across the laboratories: some simply quoted an age limit, some provided F and an estimate of sigma, but not the other terms, some reported all values as requested, some reported limit of detection values and some simply quoted a value of 0. Some laboratories commented that the SIRI samples were better than their own in house background samples (hence the issue with negative reporting). This variation reflects a variety of understandings of background samples, and the challenge that laboratories face in background evaluation. As a result, consensus values are difficult to evaluate and also potentially not useful in the standard form. For close to background samples we encountered further challenges as the results are often reported as simply "> age" which has required different statistical tools to be developed and applied. In the assemblage of results, we may have both finite and >age reported and if finite, laboratories may quote results with asymmetric errors. There is a wide literature on dealing with 'limit of detection' values, most notably the seminal paper by Currie (1968).

The standard proficiency trial analysis of samples whose activities are close to background was modified to handle the *censored* values. Non-parametric methods of estimation of the mean age used commonly in Survival or Reliability analysis, in particular the *Kaplan-Meier* survival estimator have been used to estimate the 'mean and median' age of the sample (Scott, 2003). Reliability plots (or survivor curves) display the 'survival' probabilities versus time, which in this context is the probability that the sample is greater than age t. However,

we introduced a new approach for SIRI to summarising performance (namely the Limit of Blank or LoB), (Scott et al., 2017). In the first instance, we have focussed on F since most laboratories quoted F. For the SIRI samples, we used the classic Currie (1968) paper on limits of detection and an additional quantity, the limit of blank (LoB). "LoB is the highest *apparent* analyte concentration expected to be found when replicates of a blank sample containing no analyte are tested**"** (Armbruster and Pry, 2008).

The preceding sections have reflected on the common and also less common aspects of the ^{14}C proficiency trials that need to be considered, and also offered some insights into the decisions we have made and the design criteria adopted. In the final section, we reflect on what was discovered and how it has impacted the ^{14}C community.

Findings

As we reflect on the proficiency trials we have conducted over the 30 years period, there are some common and unique features, as one would expect, associated with a technique that was still very innovative and challenging in the 1970's, to one which has become much more routine in the 2000's, .

Overall, the evidence overwhelmingly supports the fact that radiocarbon laboratories are generally accurate and precise but that notwithstanding internal QA procedures, some problems still occur that can best be detected by participation in independent inter-comparisons such as FIRI, VIRI and SIRI, where the results allow individual laboratories to assess their performance and to take remedial measures. Further, inter-comparisons are an important means to the creation of reference materials that can be used by laboratories for internal QA.

In every trial, there has always been a small percentage (typically less than 5%) of results that would be identified as anomalous. Every laboratory can produce an anomalous date, but without such proficiency trials backed up with detailed and meticulous laboratory quality control, these anomalous dates can go unrecognised. In a very few cases these have been identified as being a systematic effect, *e.g.* linked to pre-treatment or the standard used. Laboratories have often been able to trace the anomalous results to a specific event or effect within the laboratory and correct. In other cases, they may remain unexplained.

Pre-treatment methods vary enormously by material and by laboratory but are essential even though we might argue that they may introduce a further small random variation to the result, but better that than a systematic bias due to residual extraneous carbon. Pre-treatment methods such as bone pre-treatment remains an active area of research. In the early studies of TIRI and FIRI, we found little evidence of a pre-treatment effect.

Background samples remain a challenge for every laboratory and in all the trials where background or near background samples have been included, we have observed divergence in how the results are reported, and systematic differences in different materials. We have argued that a background material for each routinely dated material type (*e.g.* bone, wood, carbonate) and one which can undergo the same full set of laboratory procedures, including the pre-treatment, is key. It would also be productive to develop a clear reporting protocol.

There has been a massive shift in laboratory demographics. Thirty years ago, there was only a small handful of AMS laboratories (single figures) while now the shift is to AMS with only small numbers of radiometric laboratories, however, they have become smaller and more affordable. We require much smaller quantities of samples for AMS analysis, which makes them easier to source, but they are potentially more challenging in terms of homogeneity requirements. Increasingly, as we measure smaller and smaller samples, as we focus our attention on older time spans (>40,000 years), and as our users require better and better precision, laboratory quality assurance becomes more and more critical.

The legacy of TIRI, VIRI, FIRI and SIRI is clear in the archival material that is well characterised and catalogued in Table 2. These materials have been made available to the ^{14}C community on request and free of charge. We have also provided updates and laboratory performance assessments, as well as advice to users.

Table 2: Consensus values and 1 sigma uncertainty for TIRI, VIRI, FIRI and SIRI inter-comaprisons. (expressed as pMC or age BP dependent on the material)

Study	**Sample Code**	**Sample Type**	**Consensus value with 1 sigma uncertainty,**	**Consensus value with 1 sigma uncertainty,**
			pMC or LoB	**Age (BP)**
TIRI	Sample A	Barley mash	116.35 ± 0.0084	
	Sample B	Belfast pine Q7780		4503 ± 6
	Sample C	IAEA Cellulose **(IAEA C3)**	129.7 ± 0.08	
	Sample D	Hekla peat Iceland		3810 ± 7
	Sample E	Ellanmore peat humic acid fraction		11129 ± 12
	Sample F	Icelanadic doublespar	0.18 ± 0.0006	
	Sample G	Fugla Ness wood		39784 ± 620
	Sample H	Ellanmore whole peat		11152 ± 23
	Sample I	Caerwys Quarry Travertine		11060 ± 17
	Sample J	Buiston Crannog wood		1605 ± 8
	Sample K	Turbidite carbonate		18155 ± 34
	Sample L	Whalebone (Norway)		12788 ± 30
	Sample M	Icelandic whole peat		1682 ± 15
FIRI	Sample A	Kauri wood New Zealand	0.24	

	Sample B	Kauri wood New Zealand	0.24	
	Sample C	Marine Carbonate (**TIRI K)**		18176 ± 10.5
	Sample D	Belfast wood Q7780		4508 ± 3
	Sample E	St Bees peat humic acid fraction **(FIRI E)**		11780 ± 7
	Sample F	Belfast wood Q7780		4508 ± 3
	Sample G	Barley mash	110.7 ± 0.04	
	Sample H	German wood		2232 ± 5
	Sample I	Belfast cellulose Q7780		4485 ± 5
	Sample J	Barley mash	110.7 ± 0.04	
VIRI	Sample A	Barley mash	109.1 ± 0.04	
	Sample B	Grain (Israel)		2820 ± 4
	Sample C	Barley mash **(FIRI G)**	110.7 ± 0.04	
	Sample D	Grain Israel		2836 ± 4
	Sample E	Mammoth bone		39305 ± 121
	Sample F	Horse bone (Siberia)		2513 ± 5
	Sample G	Human bone		969 ± 5
	Sample H	Whale bone		9528 ± 7
	Sample I	Whale bone		8331 ± 6
	Sample J	Humic acid (Siberia)		43231 ± 141
	Sample K	Wood (Hohenheim)		60005 ± 846
	Sample L	Wood (Belfast)		2234 ± 17
	Sample M	Wood (Loch Tay)		2430 ± 16
	Sample N	Wood (Loch Tay)		2437 ± 17
	Sample O	Wood (Cambridge)		125 ± 16
	Sample P	Charcoal (Mexico)		1747 ± 18
	Sample Q	Charcoal (Iceland)		637 ± 17
	Sample R	Murex Shell (Israel)		2941 ± 17
	Sample S	Barley mash **(VIRI A)**	109.96 ± 0.0417	
	Sample T	Scottish Peat humic acid fraction		3360 ± 16
	Sample U	St Bees peat humic acid fraction **(FIRI E)**		11778 ± 18
SIRI	Sample A	Wood (Miocene Hohenheim Germany)	0.00381 LoB	
	Sample B	Mammal bone (North Sea)		38671 ± 72
	Sample C	Mammoth Bone (LQL4)	0.00895 LoB	
	Sample D	Barley mash	103.9 ± 0.63	
	Sample E	Wood (New Zealand)		10843 ± 6
	Sample F	Wood (Belfast)		363 ± 3
	Sample G	Wood (Belfast)		377 ± 5
	Sample H	Wood (Belfast)		386 ± 3
	Sample I	Wood (Arizona)		9995 ± 5
	Sample J	Charcoal		32002 ± 33
	Sample K	Doublespar (Iceland)	0.00465 LoB	
	Sample L	Wood (Arizona)	0.00468 LoB	
	Sample M	Wood (Scottish Crannog)	No Result*	
	Sample N	Scottish peat humic acid fraction (**VIRI T**)		3369 ± 4

* Provided only to a small number of radiometric laboratories, so no consensus value reported.

Final reflections

It is our opinion that the requirement for proficiency tests remains strong. Over the past 30 years we have seen a massive shift away from radiometric laboratories to AMS. Now the focus is on developing smaller AMS instruments, while positive ion AMS analysis is on the horizon and the latter will require significant testing to determine the potential accuracy and precision before it becomes an accepted technique. In addition, the ability to measure increasingly small samples with the most modern instrumentation has led to increasing research into measuring single compounds or classes of compounds following chromatographic separations. While the number of laboratories that will undertake this work will be limited, the very complex separation/pre-treatment procedures have the potential to lead to much more specialised, small scale proficiency tests in the future. Nevertheless, tests such as those we have been involved in over the last 30 years will still have their place in radiocarbon science.

From a user viewpoint, continuing these tests is fundamental to maintaining their confidence in the technique and the results that they are provided with by their chosen laboratory. In our opinion, it is important that users collaborate with a laboratory that routinely takes part in proficiency tests, has a good quality control regime and is happy to provide the results that demonstrate the accuracy and precision of their measurements. A good relationship between user and laboratory is extremely important and wherever possible should take precedence over analytical cost. This leads on to the question of anonymity. Throughout all the exercises we have organised, we have maintained strict anonymity, with only one of us (EMS) knowing the code associated with each laboratory. Perhaps the time has now come for laboratories to have the choice of declaring their results or staying anonymous when the results are published in the scientific literature.

In conclusion, the ^{14}C community has welcomed the regular inter-comparisons, as evidenced by the widespread participation. Participating laboratories have learned valuable lessons. The archived reference materials offer rich resources for new laboratories and for commissioning new instruments. Users have been reassured by the existence of regular comparisons that the laboratories are striving to ensure highest quality results while at the same time, the laboratories have been able to identify any systematic offsets and additional sources of variation. Indeed, in studies that have used representative samples requiring pre-treatment, chemical synthesis and measurement, it has been possible to identify the

procedure within which problems have arisen and to quantify their relative contributions to the overall variation in the results. Thus, participation in a laboratory inter-comparison has been seen to be a part of a formal QA programme and the resulting reference materials to form a community resource for the benefit of all.

Acknowledgements

The extensive programmes 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 inter-comparison. Finally, a huge thanks to the participating laboratories that have contributed several 1000's of ^{14}C dates.

References

1. Otlet R L, Walker A J, Hewson A D, Burleigh R. 1980. ^{14}C interlaboratory comparison in the UK: experiment design, preparation and preliminary results. Proceedings of 10th International ^{14}C conference, Radiocarbon 22(3), 936-947.
2. 2. Long A, Kalin R M, 1990. A suggested quality assurance protocol for radiocarbon dating laboratories. Radiocarbon 32(3), 329-334.
3. Polach H, (1989). ^{14}Care,Radiocarbon, 31(3), , 422.
4. Rozanski K, Stichler W, Gonfiantini R, Scott E M, Beukens R P, Kromer B, Van der Plicht J, 1992. The IAEA ^{14}C intercomparison exercise 1990. *Radiocarbon* 34(3), 506-519.
5. LeClerq M, van der Plicht J, Groning M (1998). New ^{14}C reference materials with activities of 15 and 15 pMC. Radiocarbon 40 (1), 295-297
6. Scott, E.M. Boaretto, E., Bryant, C., Carmi, I., Cook, G.T., Gulliksen, S., Harkness, D.D., Heinemeier, J., McGee, E., Naysmith, P., Possnert, G., Scott, E.M., van der Plicht, J. and van Strydonck, M. (2004) Future needs and requirements for AMS 14C standards and reference materials. Nuclear Instruments and Methods in Physics Research B, 223-224, 382-387.
7. Thompson M, Ellison S R, Wood R (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. Pure Appl. Chem., 78(1), 145–196.
8. Scott, E.M., Bryant, C., Carmi, I., Cook, G.T., Gulliksen, S., Harkness, D.D., Heinemeier, J., McGee, E., Naysmith, P., Possnert, G., van der Plicht, J. and van Strydonck, M. (2004) Precision and accuracy in applied 14C dating: some findings from the Fourth International Radiocarbon Inter-comparison. Journal of Archaeological Science 31, 1209-1213.

9. Scott E M, Aitchison T C, Harkness D D, Cook G T, Baxter M S, 1990. An overview of all three stages of the international radiocarbon intercomparison. *Radiocarbon* 32(3), 309-319.
10. Scott E M, Cook G T, Naysmith P (2017). Should archaeologists care about 14C inter-comparisons? Why? A summary report on SIRI. Radiocarbon
11. Harkness, DD, Cook, GT, Miller, BF, Scott, EM and Baxter, MS (1989). Design and preparation of samples for the international collaborative study. Radiocarbon 31(3): 407-413.
12. International Study Group (1982). An inter-laboratory comparison of radiocarbon measurements in tree-rings. Nature 298: 619-623.
13. Naysmith, P., Scott, E.M., Cook, G.T., Heinemeier, J., van der Plicht, J., van Strydonck, M., Ramsey, C., Grootes, P.M. and Freeman, S.P.H.T. (2007) A cremated bone inter-comparison study. Radiocarbon 49, 403-408.
14. Cook, G.T., Higham, T.F.G., Naysmith, P., Brock, F., Freeman, S.P.F.T. and Bayliss, A. (2012) Assessment of infinite-age bones from the upper Thames Valley, UK, as 14C background standards. Radiocarbon 54, 845-853.
15. Hogg A, Turney C, Palmer J, Southon, Kromer B, Bronk Ramsey C, Boswijk G , Fenwick P, Noronha A, Staff R, Friedrich M, Reynard L, Guetter D, Wacker L, Jones R, (2013) The new zealand kauri (agathis australis) research project: a radiocarbon dating intercomparison of younger dryas wood and implications for IntCal13 . Radiocarbon, 55(2), 1–14
16. ISG(1983) An international tree-ring replicate study. In Waterbolk, HT and Mook, WG, eds, Proc. Strasbourg, Pact 8: 123-133.
17. Cook G T, Harkness D D, Miller B F, Scott E M, Baxter M S, Aitchison T C (1990) International collaborative study: structuring and sample preparation. Radiocarbon, 32(3), 267-270
18. Scott, E.M., Aitchison, T.C., Harkness, D.D., Baxter, M.S., Cook, G.T. (1989) An interim progress report on stages 1 and 2 of the international collaborative program. Radiocarbon, 31, 414-421.
19. Scott, E.M., Baxter, M.S., Harkness, D.D., Aitchison, T.C., Cook, G.T. (1990) Radiocarbon: present and future perspectives on quality assurance, Antiquity, 64, 319-322.
20. Scott, E.M., Harkness, D.D., Cook, G.T., Aitchison, T.C., Baxter, M.S. (1991) Future quality assurance in 14C dating, Quaternary Proceedings, 1, 1-4.
21. Scott E M, Harkness D D, Miller B F, Cook G T, Baxter M S, 1992. Announcement of a further international intercomparison exercise. Radiocarbon 34(3), 528-532.

22. Scott, E M (ed), (2003). The third international radiocarbon inter-comparison (TIRI) and the fourth international radiocarbon inter-comparison (FIRI) 1990-2002: results, analyses, and conclusions, *Radiocarbon*, 45, 135-408.
23. Gulliksen S and Scott E M. 1995. TIRI report, *Radiocarbon* 37(2), 820-821.
24. Scott, E.M., Harkness, D.D. and Cook, G.T. (1997) Analytical protocol and quality assurance for 14C analyses: Proposal for a further intercomparison. Radiocarbon, 39, 347-350.
25. Scott E M, Harkness D D, Cook G T, (1998) Inter-laboratory comparisons; lessons learned. Radiocarbon. 40(1), 331-343.
26. 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, Sample requirements and design of an inter-laboratory trial for radiocarbon laboratories. NIM (B), 172, 2000, 355.
27. 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, J Summary findings of the Fourth international Radiocarbon Inter-comparisons (1998-2001) J. Quaternary Science, 2002, 17(7), 633-639.
28. Scott E M, Cook G T, Naysmith P (2010). A report on phase 2 of the 5th international radiocarbon inter-comparison. Radiocarbon 52 (2) 846-859.
29. Scott E M, Cook G T, Naysmith P (2010). The 5th international radiocarbon inter-comparison (VIRI): an assessment of laboratory performance in stage 3. Radiocarbon 52(2), 859-866.
30. Scott, E.M., Naysmith, P. and Cook, G.T. (2010) VIRI – summary results and overall assessment. Radiocarbon 52, 859-865.
31. Scott, E.M., Cook, G.T., Naysmith, P. and Bryant, C., O'Donnell, D. (2007) A report on phase 1 of the 5th international radiocarbon intercomparison (VIRI). Radiocarbon 49, 409-426.
32. Currie L A (1968) Limits for qualitative detection and quantitative determination. Application to Radiochemistry. *Anal. Chem.*, **1968**, *40* (3), pp 586–593
33. Armbruster D A and Pry T (2008). Limit of Blank, Limit of Detection and Limit of Quantitation. Clin Biochem Rev, V29(1).