INTERLABORATORY COMPARISONS: LESSONS LEARNED

E. M. SCOTT,[1] D. D. HARKNESS[2] and G. T. COOK[3]

ABSTRACT. Interlaboratory comparisons have been widely used in analytical chemistry and radiochemistry as an important part of ongoing quality assurance programs. The ^{14}C community has been no exception in this respect, and in just under 20 years, there have been a number of significant and very extensive interlaboratory trials organized by individual laboratories and the International Atomic Energy Agency (IAEA) to the benefit of the ^{14}C community (both labs and users) (Otlet *et al.* 1980; ISG 1982; Scott *et al.* 1990; Rozanski *et al.* 1992; Scott *et al.* 1992; Gulliksen and Scott 1995). The comparisons have varied widely in terms of sample type and preparation, but all have had as their primary goal the investigation of the comparability of results produced under possibly quite different laboratory protocols. As techniques have been developed and new labs formed, the reference materials created as a result of the intercomparisons have presented an opportunity for checking procedures and results. Users have been reassured by the existence of regular comparisons as one sign of the concern that laboratories have to ensure highest quality results, but also confused about how to make use of the results from past exercises in the interpretation of sets of dates from many laboratories. The laboratories have also learned valuable lessons from participation in such studies. These have included identification of systematic offsets and additional sources of variation and in studies which have used realistic samples requiring pretreatment, chemical synthesis and counting, it has been possible to identify the stage at which such problems have arisen and to quantify the relative contributions to the overall variation. In this paper, we will briefly review the comparisons so far, draw some conclusions from their findings, and make proposals for the future organization of intercomparisons.

INTRODUCTION

The two questions of reliability and reproducibility of routinely acquired ^{14}C dates have been and continue to be of interest to both providers and users. One of the most direct means of assessing these properties has been through organized interlaboratory comparisons, of which there are a number of significant examples. We will discuss these in detail later. Such intercomparisons form an important part of a laboratory quality assurance program, the other components of which include documented in-house laboratory procedures and the provision of suitable and well-referenced standards or reference materials.

Participation in interlaboratory comparisons has a number of advantages: for an individual laboratory, an opportunity to verify analytical performance, to identify any problems, their source and magnitude; for new laboratories in particular, such organized intercomparisons provide an invaluable opportunity to test procedures and equipment and for the user, an opportunity to be assured of the reliability and traceability of the ^{14}C results and to have confidence in the quality of the laboratory. In addition, it provides an independent assessment of interlaboratory variation, which may be important within any given research project which uses dates from different laboratories.

General Objectives and Design

There are a number of objectives of an interlaboratory comparison. In the first instance, and for the user and laboratory, it provides direct evidence of the comparability or otherwise of the results from different laboratories. It requires that a series of test samples be provided to each participating laboratory, and that it should be possible to demonstrate the homogeneity of these samples in terms of the analyte of interest. One of the main concerns is to describe the pattern of variation and to identify lab-

[1]Department of Statistics, University of Glasgow, University Gardens, Glasgow, G12 8QW, United Kingdom
[2]Natural Environment Research Council (NERC) Radiocarbon Laboratory, Scottish Enterprise Technology Park, East Kilbride, G75 0QF, United Kingdom
[3]Scottish Universities Research and Reactor Centre (SURRC) Radiocarbon Laboratory, Rankine Avenue, Scottish Enterprise Technology Park, East Kilbride, G75 0QF, United Kingdom

Proceedings of the 16th International ^{14}C Conference, edited by W. G. Mook and J. van der Plicht
RADIOCARBON, Vol. 40, No. 1, 1998, P. 331–340

oratories producing discrepant results. It is possible to quantify the extent and possible causes of interlaboratory variation, which is often broken down into a systematic component, the *bias* and the random component, the *precision*, which will include components of within-test sample and between-laboratories variation. It may, if designed appropriately, give an insight into the contributions of the various dating processes to the overall dating error. Commonly, interlaboratory trials are summarized by the properties of repeatability and reproducibility. They are defined as follows: *repeatability* refers to the variability of results performed in a single laboratory, under as near identical conditions as possible while *reproducibility* refers to the variation in results under widely varying conditions in different laboratories. In effect, they represent two extremes of variation. Typically, the focus of the trial is the laboratory and laboratory performance, but in the case of characterization of reference materials, the trial can also be used to define the qualities of the test specimens. The quality of performance of an individual laboratory can be assessed and compared with that of other laboratories (evaluation of relative bias and precision); where a laboratory falls outside performance requirements, remedial action can be taken. If the laboratories can be divided into two or more (*e.g.*, gas proportional (GPC), liquid scintillation (LSC) and accelerator mass spectrometry (AMS)) categories, the results can be compared on a method basis. If the true ages or activities of the test specimens are known, then an assessment of overall accuracy can be obtained, otherwise the results may be used to produce a consensus value for the material.

Design Issues

There are a number of design issues of a collaborative trial; many relate to the sample material, but there are also issues concerning the conduct of the trial. These are discussed briefly below.

Sample Material

There are two options in the selection of material. In the first case, all samples are of a single class of material (*e.g.*, only shell or peat or wood). This limits the generalizability of the results, and so more commonly for ^{14}C dating at least, the materials used have been representative of routinely dated material. The activity or age of the test samples should cover the ^{14}C time scale. A key question when using natural samples particularly is the homogeneity of the material, which should be tested. Obviously, as sample requirements in terms of weight may vary quite widely (through differences in pretreatment procedure, counting and technique), it is necessary that the sample should be demonstrably homogeneous at the finest level required. This is an important issue as there is an ever-growing demand for dates from smaller and smaller samples.

The number of samples is balanced between the needs of the statistical analysis of the data and of course the practical commitments of the participating laboratories. Preferably, numbers of test samples should be greater than four, and there should be replication (with the identity of duplicate pairs withheld from the participating laboratories). The presence of duplicate samples allows a direct assessment of a laboratory's repeatability, or the within-lab variation.

Other Issues

Other issues include the anonymity of participating labs, the detail of instruction concerning treatment of the samples and the reporting of results. It is hoped that the test samples from the trial will be treated routinely by the laboratory, but it is not generally feasible to introduce the samples blindly to the laboratory. Also, laboratories have typically been given no detailed instructions concerning method of pretreatment, thus increasing the variation observed, but the results are therefore more typical. Reporting of results is another important feature of the design of the trial; it must be clearly

stipulated exactly what is required for each test sample in terms of any corrections applied, the error quoted, and for old samples, it is particularly important that laboratories should be encouraged to give the exact results, rather than the more common practice of giving "censored results" (*i.e.*, in the form of greater than ages). Anonymity is an issue that concerns many users, but it must be recognized that the trial is for the participating laboratories, it represents a snapshot in time and it is likely that should a lab identify a problem as a result of participation, they will immediately take steps to remedy it. It is becoming increasingly common for laboratories to refer to their participation in such intercomparisons in publications of their work. Nevertheless, users may wish to know more and to consider how the results from an intercomparison affect their dates. The relationship between submitter and laboratory is an important one, founded on trust, but it should be remembered that the quality of results is not purely determined by the laboratory, but also by the skill and experience of the submitter who has collected the sample. Therefore the implications of intercomparison performance for a specific project should be *jointly* assessed by lab and user.

Statistical Analysis

The statistical analysis of the results from an interlaboratory trial can be carried out in a number of different ways, but always it is driven by the objectives to identify any anomalous observations, to describe the variation in the results and to characterize the test material. Evaluation of consensus values is usually done by identification of a set of homogeneous results (*i.e.*, determinations that are all in agreement given the quoted uncertainties (Wilson and Ward 1981) on which to base the calculations. Summary statistics such as the mean and median are used to define age/activity and then the overall variation around the mean is broken down into variation between replicates (within lab) and variation between labs. This latter term will typically also include a systematic component, namely the bias. The laboratory quoted uncertainties complicates the analysis, and in many cases, the laboratory variability has been expressed as a multiplier of the quoted uncertainties.

Summary of Past Intercomparisons

Otlet et al. (1980)

This study involved a small group of UK laboratories, and made use of benzene as the sole test material, with activities spanning 200% modern to 20 ka BP. The benzene had been prepared by the organizers and undergone considerable pretesting. All results were in close agreement, with no discordant results reported. The organizers thus concluded that results were comparable and there was no evidence of greater than expected variation. However, the study involved only one material and one stage of the dating process. The samples are not representative of the routinely dated material. Nonetheless, they could be considered as giving an indication of the minimum level of variation achievable in laboratory results.

International Study Group (1982, 1983)

Twenty laboratories each received a set of 8 tree-ring samples from a short floating chronology spanning 200-yr growth. The laboratories identified the samples from a tree-ring width. Since the material was provided without any preparation by the organizers, the samples were hoped to be representative of routinely submitted samples, requiring pretreatment, synthesis and counting. The results returned all lay in the range 4800–5200 BP, but there was evidence of considerable between-laboratory variation, with the span of results for an individual sample being as much as 700 yr. Each individual sample was summarized by the consensus age (necessary since the true age was not known), and the analysis proceeded by estimating the bias of a laboratory relative to the consensus

age. At the same time, the precision of results was also investigated using an error multiplier. Further, it was possible to compare GPC and LSC laboratories, with the conclusion that there appeared to be a difference in performance (generally there appeared to be an improved performance by GPC labs).

Considering the components of this trial, there were sufficient test samples (8) and sufficient participation (20 laboratories) to consider that its results could be generalized to the general laboratory population. However, the test samples were all of a single type, there were no duplicates, and the timespan was narrow. Nonetheless, it indicated that, under working conditions, comparability of results was not always achieved and that there was substantial between-laboratory variation at a scale more than anticipated. The ^{14}C community reacted positively to the conclusions of this study and undertook a further, more complex study as well as the development of a "code of practice" (Long and Kalin 1990), which recommended further interlaboratory comparisons.

International Collaborative Study (1990)

This study extended the work undertaken in earlier trials, and introduced a more complex design, to allow the quantitative assessment of the between-laboratory variation previously reported. The study had three stages, with different test materials in each stage, but also included known-age material. The study was sequential, since at each stage, an additional procedure was introduced, bringing an additional contribution to the overall variation. Each stage also included duplicate samples to allow assessment of within-lab variation and its relation to the quoted uncertainties. The study ran for 4 years, with over 50 participating laboratories. Results were summarized at an international workshop (Scott, Long and Kra 1990). Due to its design, the possible analyses of the results were much more complex and powerful. Three performance indices were defined and used to describe laboratory performance relating respectively to within- and between-laboratory variation and bias. Two of the indices were multipliers of the quoted error (internal and external error multiplier), while the third was the laboratory bias. From the duplicate results, it was concluded that the within-laboratory variation was adequately described by the quoted uncertainties, but that the between-laboratory variation (both systematic and random) was, in many cases, larger than anticipated. One conclusion was that some of the variation observed reflected the difficulties in maintaining suitable and sufficient laboratory standards and reference materials for calibration, and following this study, international efforts were made to extend the suite of reference materials available.

IAEA—Reference Materials (Rozanski et al. 1992)

Six new reference materials were distributed in 1990 to over 130 laboratories for characterization. This study was less concerned with laboratory performance and more with the suitability of the test materials and their future use. The materials had already undergone homogeneity testing before distribution; they ranged in age from modern to background and included a number of different sample classes (wood, cellulose, sucrose and carbonate). Results from 69 laboratories were reported. Overall there was generally good agreement in the results, but a number of difficulties were subsequently identified. Analysis proceeded by identifying a set of results which satisfied a homogeneity criterion (key issue when using natural samples, and one which must be fully addressed since it may contribute substantially to the overall variation in results) which would then be used to estimate the consensus values. This analysis highlighted problems with some of the reference samples (C-1 (Carrera marble) which showed difficulties with background samples and the problems of contamination) and C-4 (Kauri wood, where some contamination occurred as a result of the milling process), and in some cases up to 60% of the original results were excluded from the statistical evaluation. Finally, the influence of operational factors was explored; these included laboratory type, which was found

to be insignificant (*i.e.*, there were no significant differences between the lab types), whereas the effect of modern standard used was found to be of significance.

TIRI (Scott et al. 1992; Gulliksen and Scott 1995)

TIRI (the Third International Radiocarbon Intercomparison) began in 1991, and again involved a large number of labs (more than 70). One of the features of the design of TIRI was its two-stage nature (in the second stage, an optional set of materials was available) and the fact that again, all the test materials were natural. TIRI was designed to provide an independent assessment of laboratory performance, following the recently completed IAEA study and hence the materials were designed to test the full laboratory procedure. In the first stage, a series of core samples (6 in total) were distributed to all laboratories. The samples had been broadly classified into age ranges: modern; <1 half life; between 1 and 2 half-lives; between 2 and 3 half-lives; and >3 half-lives. It included grain (modern); wood (dendrochronologically dated); cellulose from the IAEA study (providing a link to the IAEA study); peat; humic acid; and calcite (background). Thus, the samples covered the ^{14}C age spectrum. Table 1 shows the information about the samples used. The preprocessing of the samples was strictly limited; in some cases, the samples were homogenized by grinding and mixing, but with no chemical pretreatment, in others (*e.g.*, humic acid) chemical pretreatment was applied before dispatch. All laboratories received all the core samples. In the second stage, laboratories were able to select test samples from a list of materials which were of a more specialized nature and which might be seen as less routine. The optional samples included whalebone, whole peat, wood and travertine. Some of the samples in the second stage were related to those used in the first stage (peat and humic samples). A substantial proportion of labs opted to take at least some of the optional samples.

Analysis of TIRI Results

The first step in the analysis of TIRI was to identify anomalous observations, and define consensus values for the samples. The approach taken here was similar to that in the IAEA study (Rozanski *et al.* 1992), *i.e.*, first a robust measure of activity (or age) is evaluated, then results are omitted from the final calculation if they are more than 2 quoted errors away from the robust measure. Finally, a weighted average of the results remaining is used as the consensus value.

Table 1 summarizes the consensus values evaluated using this approach for both core and optional samples. It should be noted that for sample F, the calculations were made more difficult by the fact that many results were given as greater than values. The consensus values are then used in the next stage of the analysis, which involves exploring any laboratory bias and evaluating laboratory precision, relative to the consensus values. At this stage, it is also possible to explore whether there are any differences in the different lab groups (GPC, AMS, LSC).

Results

The pattern of variation (both systematic and random) can be studied by exploring the *deviations* which are defined as

$$\text{deviation} = (\text{age} - \text{consensus value})/\text{quoted uncertainty on age} \quad . \qquad (1)$$

Figures 1A–D show plots of the deviations for some participating labs for samples from both stages. The horizontal lines at ±2 aid the interpretation of such deviations: in the ideal case (no systematic bias and no variation in excess of quoted uncertainties), the results for a lab should all lie between these lines. The pattern of observed behavior over the participating laboratories is well represented in these figures and can be classified into four groups: Group 1, laboratories whose results lie within

TABLE 1A: Consensus Values for Stage 1 TIRI Samples

Sample	Consensus value	Estimated precision (1σ)
A: barley mash	116.35 pMC	0.0084
B: Belfast pine	4503	6
C: IAEA cellulose	129.7 pMC	0.08
D: Hekla peat	3810	7
E: Ellanmore humic	11129	12
F: Icelandic doublespar	46750	208
	0.18 pMC	0.006

TABLE 1B: Consensus Values for Stage 2 TIRI Samples

Sample	Consensus value	Estimated precision (1σ)
G: Fuglaness wood	39784	620
H: Ellanmore whole peat	11152	23
I: Travertine	11060	17
J: Crannog wood	1605	8
K: Turbidite carbonate	18155	34
L: Whalebone	12788	30
M: Icelandic peat	1682	15

or close to these limits; Group 2, laboratories as in Group 1, but with a single anomalous value; Group 3, laboratories whose results lie systematically outside the lines; and Group 4, laboratories whose results are widely scattered.

Figures 2A and 2B show typical graphical summaries of the deviations for two of the samples by laboratory type. In Figures 1 and 2, it is clear that anomalous values occur; in Figure 1, the anomalous value refers to a single result by a laboratory, while in Figure 2 (anomalous value denoted by *) it refers to an individual laboratory. These simple diagrams again graphically provide evidence of variation in results exceeding the quoted uncertainties. Finally, we can summarize the overall performance using a very simple model based on the deviations and which allows us to make use of an error multiplier and laboratory bias term.

In 14 cases, laboratories were found to have a significant bias; in all other cases (55), no such systematic bias was found. For these 55 laboratories, an error multiplier was then evaluated and Figure 3 shows a histogram of the results. Of the 55 laboratories, 28 had an error multiplier <2, and a significant number of these had a multiplier <1.

The error multiplier is a rather simple tool, which has advantages and disadvantages in its use. Its main advantage is that it is very simple to use, and relates the observed variation in a direct way to the quoted uncertainties, but it is difficult to meaningfully interpret, at least from the analyst's perspective, and it is highly sensitive to anomalous observations. It refers to the results as reported and thus may not be directly generalizable beyond the study to which it refers.

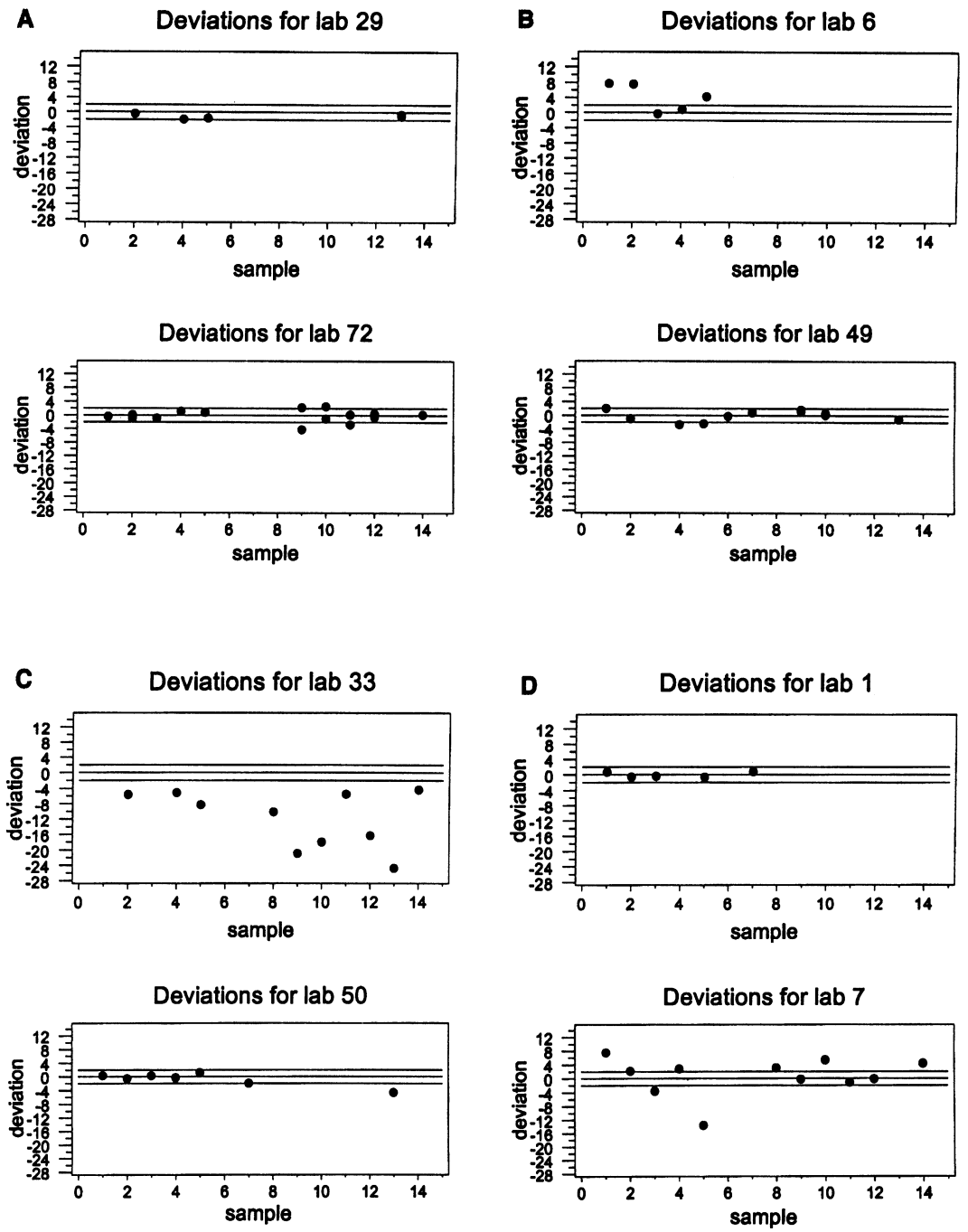

Fig. 1A–D. Deviations for individual laboratories within TIRI

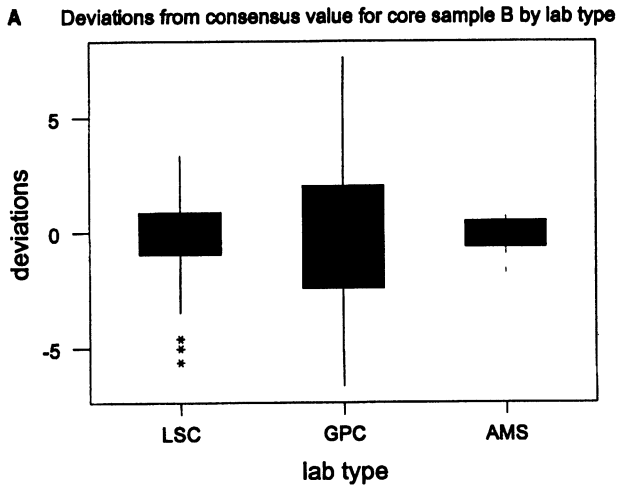

Fig. 2A. Deviations for sample B by lab type

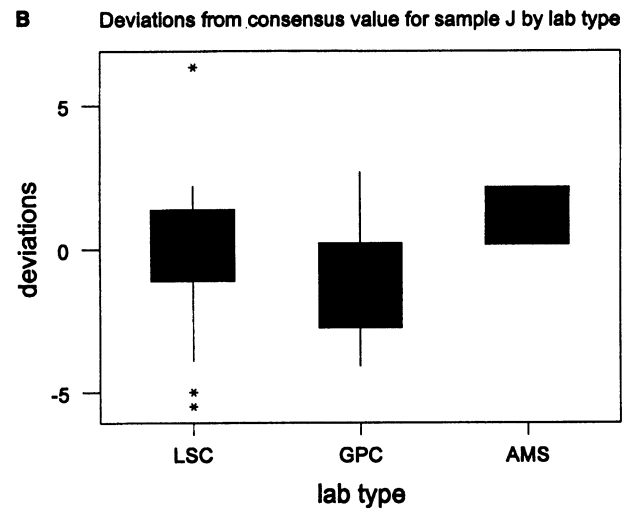

Fig. 2B. Deviations for sample J by lab type

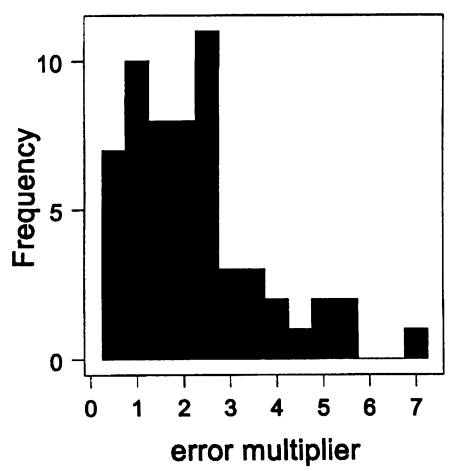

Fig. 3. Histogram of error multipliers

Nevertheless, in TIRI as in the other studies, it points to variation in the results beyond that described by the quoted uncertainties. TIRI was not intended to explore the sources of the variation in the results, but it should be noted that at the TIRI workshop (Gulliksen and Scott 1995), there had been discussion concerning the homogeneity of the test samples, the issues of selection of small samples for AMS dating and the question of differing measured ^{14}C contents depending on the chemical fraction dated. It is clear that in any study using natural samples, some part of the extra variation must be due to the sampling of the bulk material. These issues are ones which will become increasingly important in future dating exercises.

Conclusion

All of the studies cited above have provided valuable information to laboratories and hence to users. In all cases where natural samples have been used, there has been evidence of additional variation in the results; only in one case (Otlet *et al.* 1980), which used artificially prepared samples, was there no evidence of increased variation. In all studies, anomalous observations have been found, although there is no evidence that they occur on a frequent basis. The studies have all highlighted additional variation, with no clear evidence of substantial improvement over the years; in each study since 1981 significant between-laboratory variation has been identified. Does this mean that ^{14}C laboratories have learned no lessons from their participation? Resoundingly, the answer must be no: by the nature of the technique (random decay process), by the natural variation of ^{14}C in the environment, it is clear that there will always be variation in the determinations. This cannot be reduced to zero, but what can be done, however, is to eliminate systematic biases and to ensure that the uncertainties quoted by the laboratories are realistic. As a result, it is clear that such checks as TIRI and others are and will continue to be necessary and that they must operate in addition to any within-laboratory procedures. Nor is it clear in these studies that the increased availability of an extensive range of reference materials has presented an immediate solution to the problem of laboratory comparability as might have been hoped. Increasing the scope of reference materials and standards is important, since by their inclusion, the dating determinations can be better constrained but only if laboratories make regular use of them in routine operation. Since the 1980s when these large-scale studies began, there have been significant changes in the mode of operation of many laboratories. More and more requests are being made for ^{14}C determinations which cannot be classed as strictly routine. There is still a need for routine dating, where intermittent checks are necessary and which can be satisfied by materials such as the IAEA reference materials and by programs such as TIRI which were directed more at large sample dating, but there is clearly also a need for further exploration of comparability and variation at the limits of the technique (very small or very old samples).

There is increasing pressure to date smaller (even to the molecular level) and older samples, and more conventional laboratories are forming close collaborations with accelerator labs, which has meant developing in-house techniques for target preparation. Thus, an accelerator lab may have a number of target preparation labs providing it with targets and presenting new issues of comparability. Perhaps, however, the most significant factor is that as we strive to measure smaller and smaller samples, the issue of sample homogeneity becomes more and more important—indeed the definition of a sample becomes critical. In some of the studies already completed in which AMS labs have participated, some evidence of sample inhomogeneity has been reported, which the conventional laboratories were not able to detect. There are difficulties in taking a representative subsample from the bulk of material—indeed how do we know it is representative? We do not fully know the potential scale of natural ^{14}C variation in sample matrices.

Proposals

The time has come for a further exercise, but in essence for the first time, we need to consider dealing with the issues of AMS and conventional dating separately. Continuation of this work is important. The linkage to previous work provides an invaluable continuity (*e.g.*, IAEA and other reference materials are still available and should be used), but in addition further, new materials should be sought, including known-age material. For the conventional laboratory, the typical sample requirement might be 5 g C with sample age ranges from 1 to 4 half-lives. However, for the AMS labs, and those conventional labs where small samples are dated, we need to explore the natural variation in reportedly single event samples (deposits of charcoal, grain from a single growing season, single insects from a well-defined stratum). This information is not just important for the laboratory, but is also of fundamental importance for the sample submitter who must select samples referring to the event of interest. There are new challenges for ^{14}C dating in continuing to ensure the quality of results. Such a study is planned for 1998. Its design plans for two distinct experimental arms for AMS and radiometric laboratories linked by common samples (notably known-age wood). Pretreated and non-pretreated, homogenized samples that are well constrained in age/activity will be distributed, and for some samples, duplicate analyses will be requested. Analysis of the results will concentrate on the comparability of results but will also attempt to estimate the components of variation in the results due to sampling, to natural variations in activity when selecting small samples, and to pretreatment procedures.

Acknowledgments

It is a pleasure to thank our many colleagues around the world who have contributed to much of the work discussed here and also those who have provided sample material used in the intercomparisons organized from Glasgow. The financial support of the EPSRC and NERC are also gratefully acknowledged.

References

Gulliksen, S. and Scott, E. M. 1995 Report of the TIRI workshop. *In* Cook, G. T., Harkness, D. D., Miller, B. F. and Scott, E. M., eds., Proceedings of the 15th International ^{14}C Conference. *Radiocarbon* 37(2): 820–822.

ISG 1982 An inter-laboratory comparison of radiocarbon measurements in tree-rings. *Nature* 198: 619–623.

____1983 An international tree-ring replicate study. *In* Waterbolk, H. T. and Mook, W. G., eds., ^{14}C and Archaeology. *PACT* 8: 123–233.

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

Otlet, R. L., Walker, A. J., Hewson, A. D. and Burleigh, R. 1980 ^{14}C interlaboratory comparison in the UK: Experiment design, preparation and preliminary results. *In* Stuiver, M. and Kra, R. S., eds., Proceedings of 10th International ^{14}C conference. *Radiocarbon* 22(3): 936–947.

Rozanski, K., Stichler, W., Gonfiantini, R., Scott, E. M., Beukens, R. P., Kromer, B. and van der Plicht, J. 1992 The IAEA ^{14}C intercomparison exercise 1990. *In* Long, A. and Kra, R. S., eds., Proceedings of the 14th International ^{14}C Conference. *Radiocarbon* 34(3): 506–519.

Scott, E. M., Aitchison, T. C., Harkness, D. D., Cook, G. T. and Baxter, M. S. 1990 An overview of all three stages of the international radiocarbon intercomparison. *Radiocarbon* 32(3): 309–319.

Scott, E. M., Harkness, D. D., Miller, B. F., Cook, G. T. and Baxter, M. S. 1992 Announcement of a further international intercomparison exercise. *In* Long, A. and Kra, R. S., eds., Proceedings of the 14th International ^{14}C Conference. *Radiocarbon* 34(3): 528–532.

Scott, E. M., Long, A. and Kra, R. S., eds. 1990 Proceedings of the International Workshop on Intercomparison of Radiocarbon Laboratories. *Radiocarbon* 32(3): 253–397.

Wilson, S. R. and Ward, G. K. 1981 Evaluation and clustering of radiocarbon age determinations: Procedures and paradigms. *Archaeometry* 23(1): 19–39.