[RADIOCARBON, VOL. 34, NO. 3, 1992, P. 506–519]

THE IAEA ^{14}C INTERCOMPARISON EXERCISE 1990

KAZIMIERZ ROZANSKI[1], WILLIBALD STICHLER[1], ROBERTO GONFIANTINI[1], E. M. SCOTT[2] R. P. BEUKENS[3], BERND KROMER[4] and JOHANNES VAN DER PLICHT[5]

ABSTRACT. As a follow-up to the meeting of experts convened at the International Atomic Energy Agency (IAEA) in February 1989, and the International ^{14}C Workshop held in Glasgow in September 1989, the ^{14}C Quality Assurance Program was formulated. In a joint effort of several radiocarbon teams and IAEA staff, we have prepared a set of five new intercomparison materials. These are natural materials frequently used by radiocarbon laboratories. The materials were distributed to 137 laboratories in May 1990. In February 1991, a meeting of experts was convened in Vienna to evaluate the results, to determine the radiocarbon activity of the five samples expressed in % Modern (pMC) terms and to define the $^{13}C/^{12}C$ ratio, and to make recommendations on further use of these materials. We present here the results of the exercise and the agreed consensus values for each of the five materials and discuss the different analyses that were undertaken.

INTRODUCTION

The radiocarbon community has participated in a number of interlaboratory checks during the last decade (Otlet *et al.* 1980; ISG 1982, 1983). The most ambitious project to date was launched by the Glasgow group and supported by over 50 radiocarbon laboratories. This three-stage study was recently completed and the results published (Aitchison *et al.* 1990; Cook *et al.* 1990; Scott *et al.* 1990). The latter two studies have highlighted difficulties in the comparability of ^{14}C laboratories, and have quantified excess variability in the results.

At a meeting during the 13th International Radiocarbon Conference held in Dubrovnik in June 1988, several laboratories expressed the need for ^{14}C reference materials in addition to the recent oxalic acid standard of NBS (now NIST). Accordingly, experts convened at the International Atomic Energy Agency (IAEA) in Vienna in February 1989, where an outline of the ^{14}C Quality Assurance Program was formulated (Rozanski 1989). The issue was further discussed during the International Workshop on Intercomparison of ^{14}C Laboratories, held in Glasgow in September 1989. The Agency's offer to provide and distribute intercomparison materials as part of the ^{14}C Quality Assurance Program, was thoroughly discussed there and accepted (Mook 1990). The new intercomparison exercise forms a part of the Analytical Quality Control Service (AQCS) (Gonfiantini *et al.* 1990) initiated by the IAEA, to assist laboratories engaged in various fields of scientific research to check the quality of their work.

As a follow-up to the Vienna and Glasgow meetings, a joint effort of several colleagues from the radiocarbon community and Agency staff resulted in the preparation of five new intercomparison materials. These are natural materials frequently used in radiocarbon laboratories.

THE EXERCISE

In May 1990, after passing homogeneity tests, the new intercomparison materials (Table 1) were distributed to *137 laboratories* worldwide. The ANU Sucrose Secondary Standard, internationally calibrated against the NBS Oxalic Acid Standard (Currie & Polach 1980) was added to the set of distributed materials. The laboratories were asked to report technical details of the preparation and

[1]International Atomic Energy Agency, A-1400 Vienna, Austria
[2]Department of Statistics, Glasgow University, Glasgow G12 8QW Scotland
[3]Isotrace Laboratory, University of Toronto, 60 St. George Street, Toronto, Ontario M5S 1A7 Canada
[4]Institut für Umweltphysik, University of Heidelberg, Im Neuenheimer Feld 366, D-6900 Heidelberg, Germany
[5]Centre for Isotope Research, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

TABLE 1. IAEA ^{14}C Quality Assurance Materials

IAEA code	Material	Prepared by	Amount stored
C-1	*Carbonate* Slab of freshly cut Carrara marble milled down to a dust-free fraction of 1.6–5.0 mm by IAEA	IAEA	~70 kg
C-2	*Carbonate* Fresh water travertine deposit collected near Munich, Germany, and homogenized by IAEA	IAEA	~70 kg
C-3	*Cellulose* Batch of cellulose produced in 1989 from one season's harvest of *ca.* 40-year-old trees.	W. G. Mook J. van der Plicht	In bulk
C-4	*Subfossil wood* Subfossil wood excavated from peat bogs in the north island of New Zealand, near Waikato	W. G. Hogg H. A. Polach	~80 kg
C-5	*Subfossil wood* Subfossil wood originating from buried bed forest in eastern Wisconsin, USA, near the western shore of Lake Michigan	R. M. Kalin A. Long IAEA	~50 kg
C-6*	*Sucrose*	H. A. Polach	in bulk

*ANU Sucrose Secondary Standard was internationally calibrated against the NBS Oxalic Acid Standard and made available to the ^{14}C community in the early 1980s.

measurement procedures adopted in the analysis. Following the recommendations of the Vienna and Glasgow meetings, permission was requested from participants to disclose the name of their laboratories, in association with their results. *Only results submitted by laboratories that agreed to disclose their identity were further evaluated and published in the summary report.*

Experts then convened in Vienna to evaluate the results submitted by the laboratories participating in the exercise, and to provide guidelines for further distribution of the available set of intercomparison materials. An IAEA report (Rozanski 1991) summarizes the meeting, and provides the primary reference to the materials and the submitted results.

RESULTS AND ANALYSIS

Results were received from 69 laboratories (38 of them representing liquid scintillation counting (LSC), 25 gas proportional counting (GPC), and 6 accelerator mass spectrometry (AMS)). Altogether, 441 ^{14}C analyses were reported to the IAEA by the participating laboratories. In the following analyses, multiple results from a single laboratory are treated independently. Eight of the 69 laboratories did not grant permission to publish their results in the IAEA summary report. Figures 1 and 2 show the results for each of the samples in the form of a boxplot, the central box of which shows the middle 50% of the data; other features include the extremes and outlying observations. The diagrams clearly demonstrate the general agreement of the results, but also the existence of outlying observations.

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

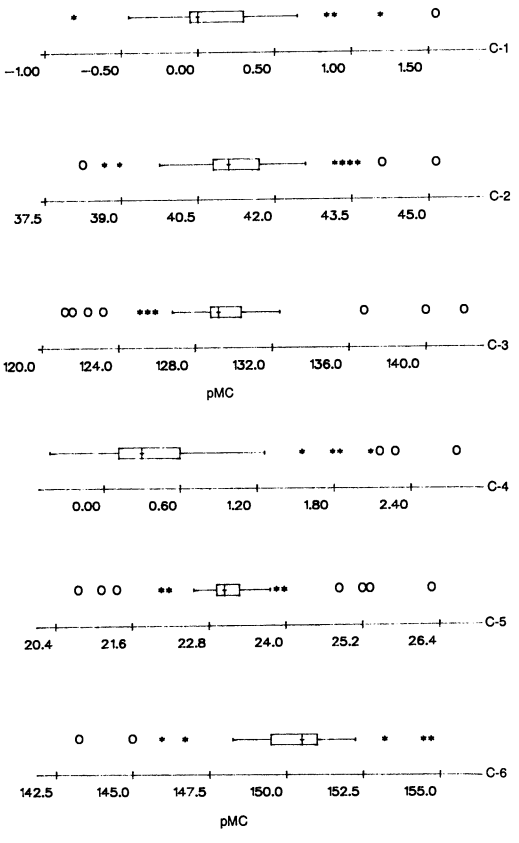
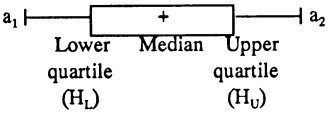


Fig. 1. Boxplots of the results for samples C-1, C-2 and C-3 with the following construction



*, O – location of unusual observations

a_i most extreme observation, such that (i = 1,2)

$a_1 > H_L - 1.5\ (H_U\text{-}H_L)$

and $a_2 < H_U + 1.5\ (H_U\text{-}H_L)$

Fig. 2. Boxplots of the results for samples C-4, C-5 and C-6

The analysis of the results reported here is designed to: 1) characterize the reference samples; 2) investigate the level of variability in the results, *i.e.*, considering the influence of operational factors, such as lab type (LSC, GPC or AMS), counter technology, modern standard and $\delta^{13}C$. We present only part of the full analysis, excluding: 3) estimation of indicators of laboratory performance; this will be the focus of a future publication.

The Reference Samples

The overall aim of the characterization procedure is to evaluate an unbiased estimate of the percent modern carbon (pMC) for each sample, and to evaluate the precision of the estimate. The process by which the consensus value for each reference sample has been achieved follows:

Stage 1. Outlying observations were omitted (taking no account of error). For each sample, a number of observations identified as outliers were omitted. If we define H_L to be the lower quartile of the data, H_U to be the upper quartile, then values exceeding either $H_L-3.0*(H_U-H_L)$ or $H_U+3.0*(H_U-H_L)$ were excluded. This resulted in 7 values on C-1, 4 on C-2, 3 on each of C-3 and C-4, and 2 on C-5 and C-6 being excluded.

We obtained an overall preliminary consensus value from the remaining results, again taking no account of the quoted errors. This preliminary consensus value is the median of all the results, which should be robust to any remaining outlying observations and denoted, m. Table 2 provides a basic summary of the preliminary consensus value and the data remaining after outlier exclusion. We can demonstrate that the results remaining for each sample do not comprise a homogeneous

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

TABLE 2. Preliminary Summarization of Results

Sample	Total no. of analyses	No. of analyses after outlier removal	Median	Interquartile range	Quartiles H_L	Quartiles H_U
C-1	73	66	0.06	0.36	−0.016	0.34
C-2	92	88	41.18	0.81	40.92	41.73
C-3	84	81	129.46	1.69	128.76	130.45
C-4	79	76	0.32	0.52	0.12	0.64
C-5	75	73	23.05	0.37	22.93	23.30
C-6	39	37	150.57	1.56	149.50	151.06

group within the errors claimed. This requires further data manipulation before final consensus values can be obtained.

Stage 2. In order to achieve a more precise measure of the pMC, we identify a subgroup of results by accepting the result x,s if $|(x-m)/s| < 2$, where x is the pMC, s the quoted error and m, the preliminary consensus value found in Stage 1 and described in Table 2.

Stage 3. We then evaluated the final consensus value as a weighted average. Appendix 1 shows the model and mathematical details of the calculations for this stage. The procedure adopted in Stage 2 identifies a homogeneous group of labs in terms of result and quoted error, and allows the use of the weighted average in characterizing the samples. An additional error term has been included in the formula for the estimated standard error (ESE) for the weighted average, which effectively inflates the ESE by a degree related to the level of homogeneity of the results. The final result achieved in this way should be *accurate* and *precise*, and has been *calculated without undue influence being given to results from laboratories with unrealistically small quoted errors*. Table 3 provides the consensus values achieved using this approach. Sample C-1, Carrara marble, should be considered as a background sample, having no measurable ^{14}C activity.

Figures 3 to 8 present the results remaining after Stages 1 and 2. These results were used to derive the consensus values of Table 3. Each figure shows the laboratory results, marked by a symbol denoting laboratory type and a horizontal line indicating twice the quoted error. The consensus value and two vertical lines indicating a 95% confidence interval for the consensus are also shown.

Comments. Results remaining after Stages 1 and 2 were tested under the homogeneity hypothesis and satisfied the criterion. However, attention must be directed to C-1 and C-4, which previous diagrams (Figs. 1, 2) showed had a skewed distribution of results, and C-6, which shows much lower precision than the other samples (a combination of larger errors and fewer labs). It is also worth noting that creation of *these subgroups of homogeneous results* radically reduces the available pool of measurements by 30 to 60%. This is a *very worrisome statistic.*

For sample C-4, additional care had to be taken in evaluating an appropriate consensus value as considerable variation in the results skewed them to the right. For this reason, the median value is given, along with its 95% confidence range and interquartile range, in preference to the weighted average. This sample continues to be evaluated for inhomogeneity.

Consensus values were also evaluated under two other criteria: 1) stricter – $|(x-m)|/s < 1$; 2) less strict – $|(x-m)|/s < 3$. For 1), we made an even larger reduction in the number of allowable results. Again, all the remaining results are homogeneous for each of the materials, and the changes in the consensus values from those in Table 3 are small. For the less strict criterion of 2), considerably

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

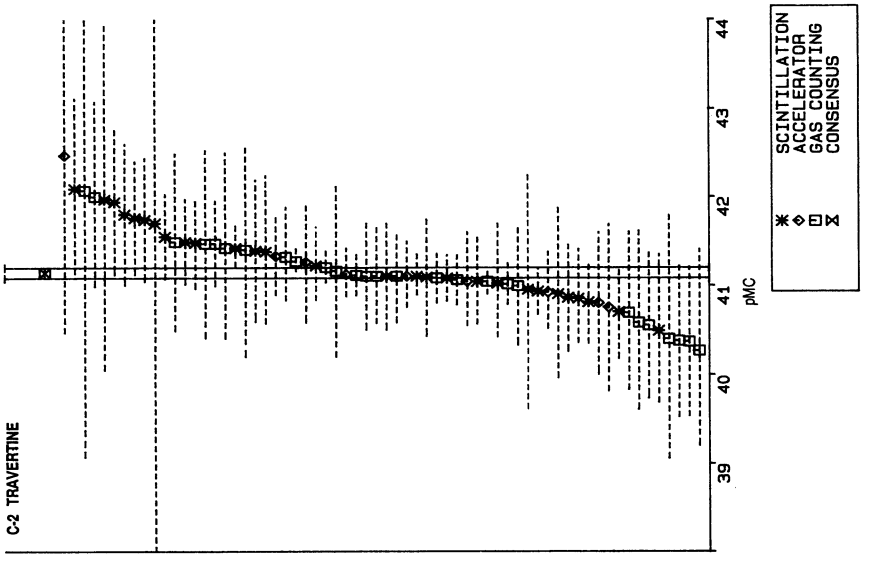


Fig. 4. Plot of pMC ± 2 quoted errors for all results on C-2 used to define consensus value

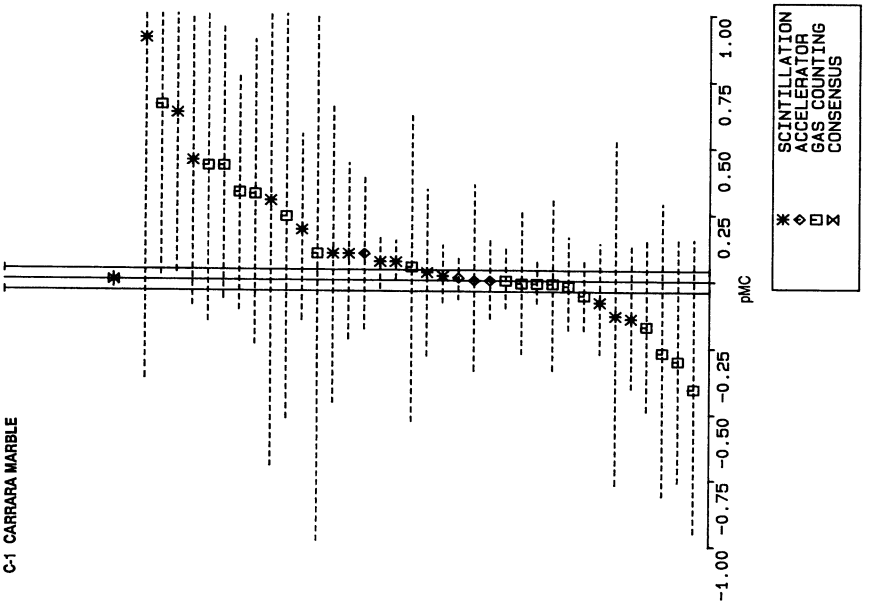


Fig. 3. Plot of pMC ± 2 quoted errors for all results on C-1 used to define consensus value

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

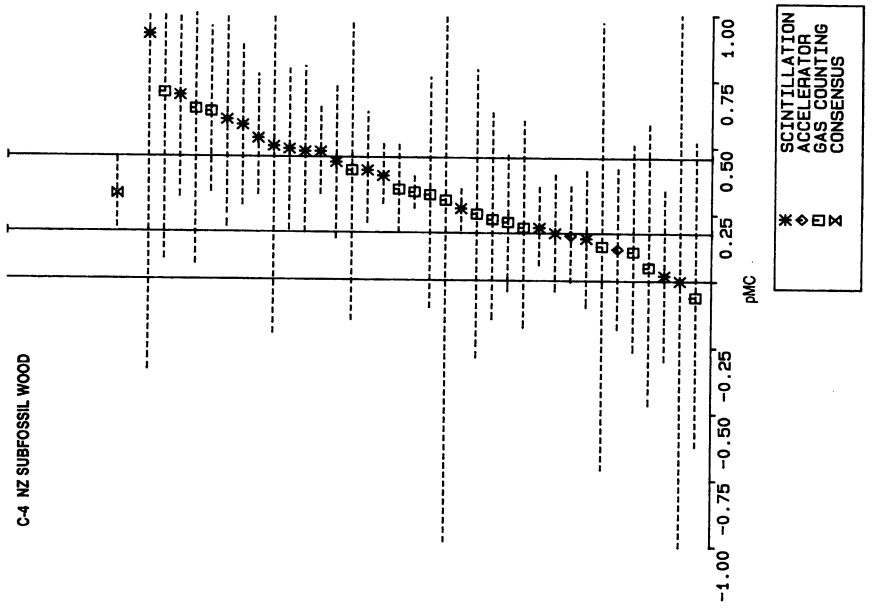


Fig. 6. Plot of pMC ± 2 quoted errors for all results on C-4 used to define consensus value

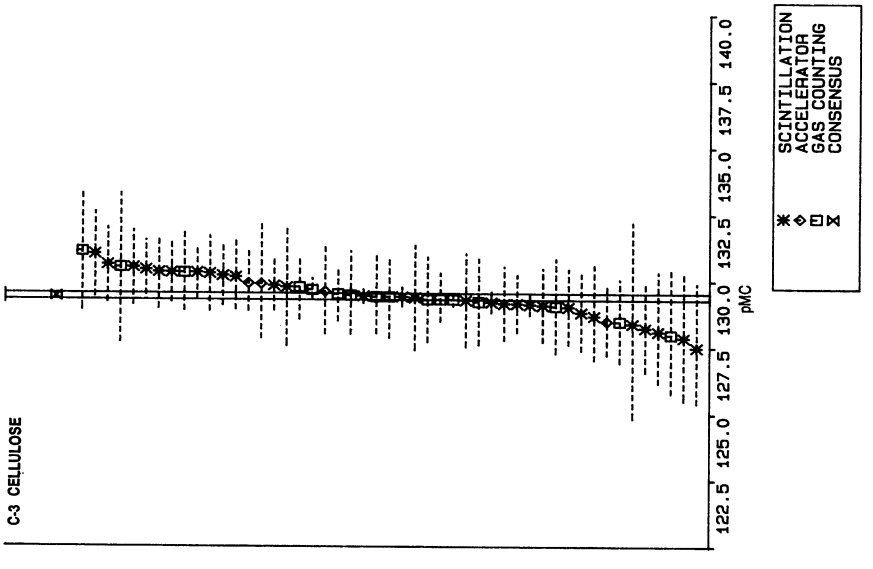


Fig. 5. Plot of pMC ± 2 quoted errors for all results on C-3 used to define consensus value

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

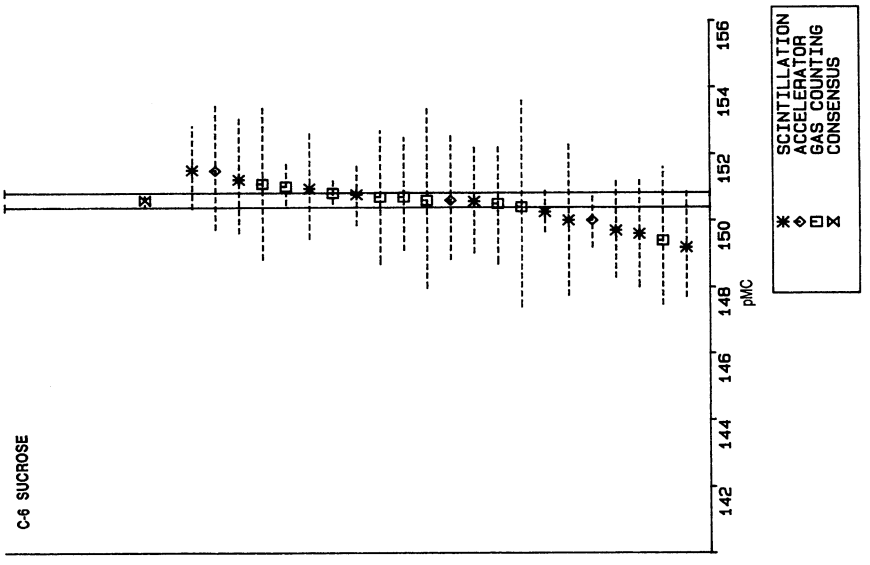


Fig. 8. Plot of pMC ± 2 quoted errors for all results on C-6 used to define consensus value

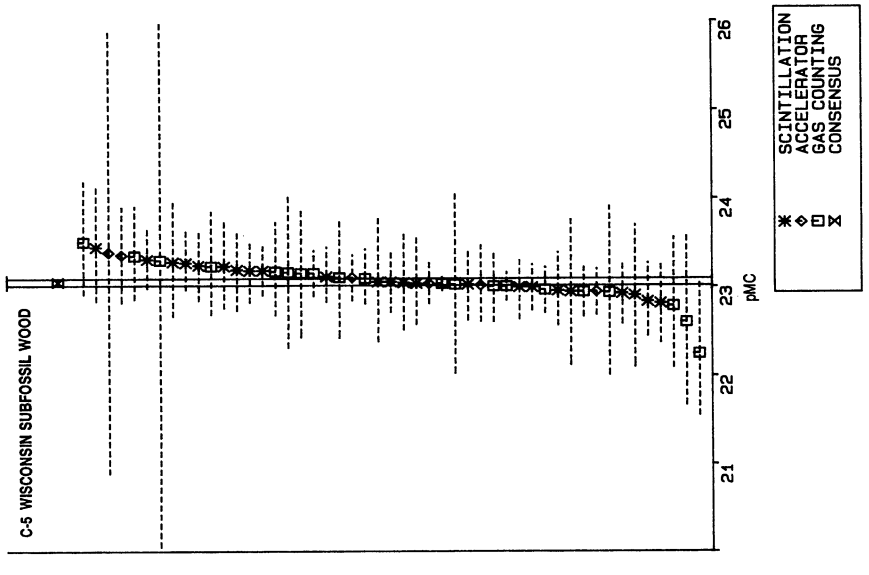


Fig. 7. Plot of pMC ± 2 quoted errors for all results on C-5 used to define consensus value

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

TABLE 3. Consensus Values of the IAEA ^{14}C Quality Assurance Materials

		^{14}C			^{13}C	
Material	No. of analyses*	Consensus value** (pMC)	Estimated standard error† (pMC)	Number of analyses*	Consensus value‡ (‰) PDB	Standard deviation‡ (‰) PDB
C-1	36 (73)	0.00(0.02)§	0.02	59 (63)	2.42	0.33
C-2	64 (92)	41.14	0.03	73 (79)	−8.25	0.31
C-3	49 (84)	129.41	0.06	70 (75)	−24.91	0.49
C-4	36 (79)	0.20-0.44‖	–	67 (68)	−23.96	0.62
C-5	49 (75)	23.05	0.02	66 (68)	−25.49	0.72
C-6	22 (39)	150.61	0.11	35 (36)	−10.80	0.47

*Number of accepted analyses; total number of analyses submitted to IAEA is indicated in parentheses.

**Calculated as weighted average: $\bar{x}_w = (\Sigma x_i/w_i^2)/(\Sigma 1.0/w_i^2)$

†Estimated standard error, calculated according to:

$$ese(\bar{x}_w) = \hat{\sigma}_w\sqrt{(\sum 1.0/w_i^2)} \quad \text{where} \quad \hat{\sigma}_w^2 = \frac{\sum((x_i - \bar{x}_w)^2/w_i^2)}{n}$$

(see Appendix 1A)

‡Calculated according to the 3-σ criterion (see text for details)

§This material is considered as a background sample, having no measurable ^{14}C activity. The weighted mean is indicated in parentheses.

‖95% confidence interval for the median

fewer results are excluded, but the remaining ones are no longer homogeneous within the quoted errors. Thus, this would be an inappropriate grouping to select for characterization of the materials.

In the final calculation, for each material, we grouped results according to the quoted error and pMC estimated by the weighted average. The weighted average and its estimated standard error at each level of quoted error show variation, and again, in general, inhomogeneity within each subgroup in terms of the quoted error. These difficulties justify the introduction of external criteria upon which to judge a result, in this case, in terms of both its accuracy (difference from the consensus value) and its precision (ratio of the difference from the consensus value and the lab quoted error).

Influence of Operational Factors

As Figures 3–8 demonstrate, results vary considerably. Consequently, we have omitted up to 60% of the submitted values in arriving at the final consensus values. We seek to explain some of this variability in terms of operational factors, provided by the laboratories in their reports. Table 4 lists these factors and the classifications used in the analysis. The original results are now transformed to give deviations of the form $(x_{ij} - m_i)/s_{ij}$, where x_{ij} is pMC quoted by lab j for sample i, s_{ij} is the corresponding quoted error, and m_i is the agreed consensus value for sample i. This transformation combines both submitted result and its quoted error in subsequent analyses. Where no error was quoted, results are not included in these analyses.

Analysis 1: Laboratory Type

Figures 9A–9C show the scatter of deviations for each of the three laboratory types. Some differences can be seen for the samples across the laboratory types, *e.g.*, there is an indication that

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

TABLE 4. Operational Factors

Laboratory Type		Gas proportional counting (GPC) Liquid scintillation counting (LSC) Accelerator mass spectrometry (AMS)
Counting Technology		
LSC	Counters	Quantulus (Q) Packard (P) Tricarb (T) LKB (L) Other (O)
GPC	Counting gas	CO_2 other (CH_4 and C_2H_2)
Modern Standards		ANU Sucrose NBS Oxalic I NBS Oxalic II Other
$\delta^{13}C$		

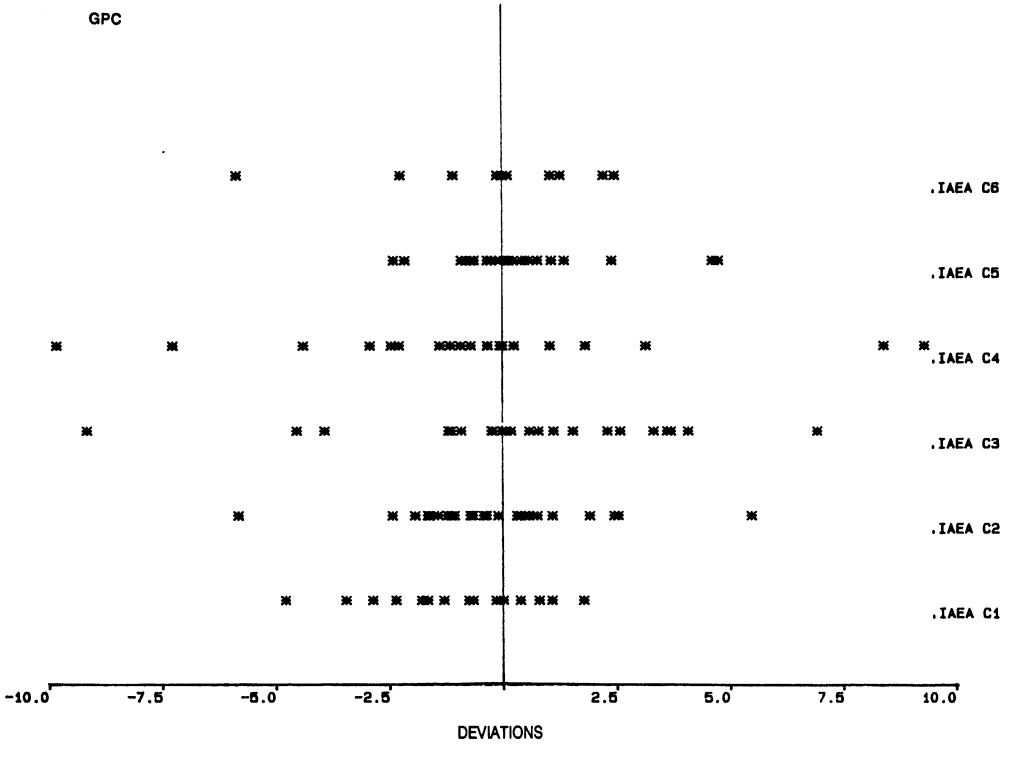


Fig. 9A. Plot of deviations from the consensus for each of the quality assurance samples for GPC laboratories

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

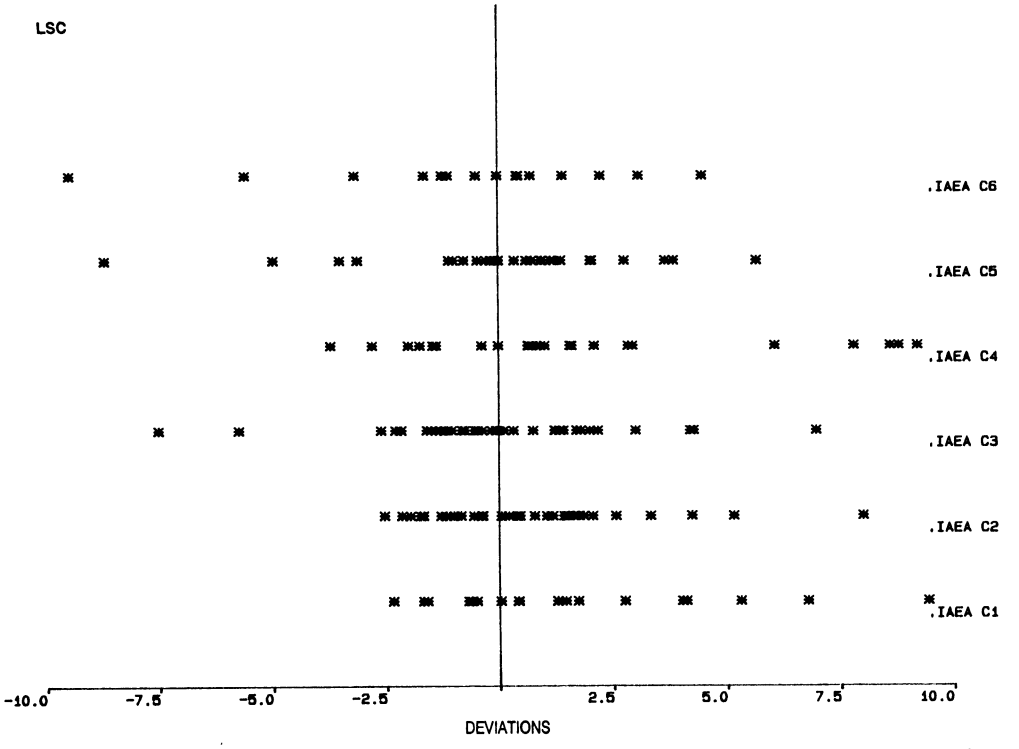


Fig. 9B. Plot of deviations from the consensus for each of the quality assurance samples for LSC laboratories

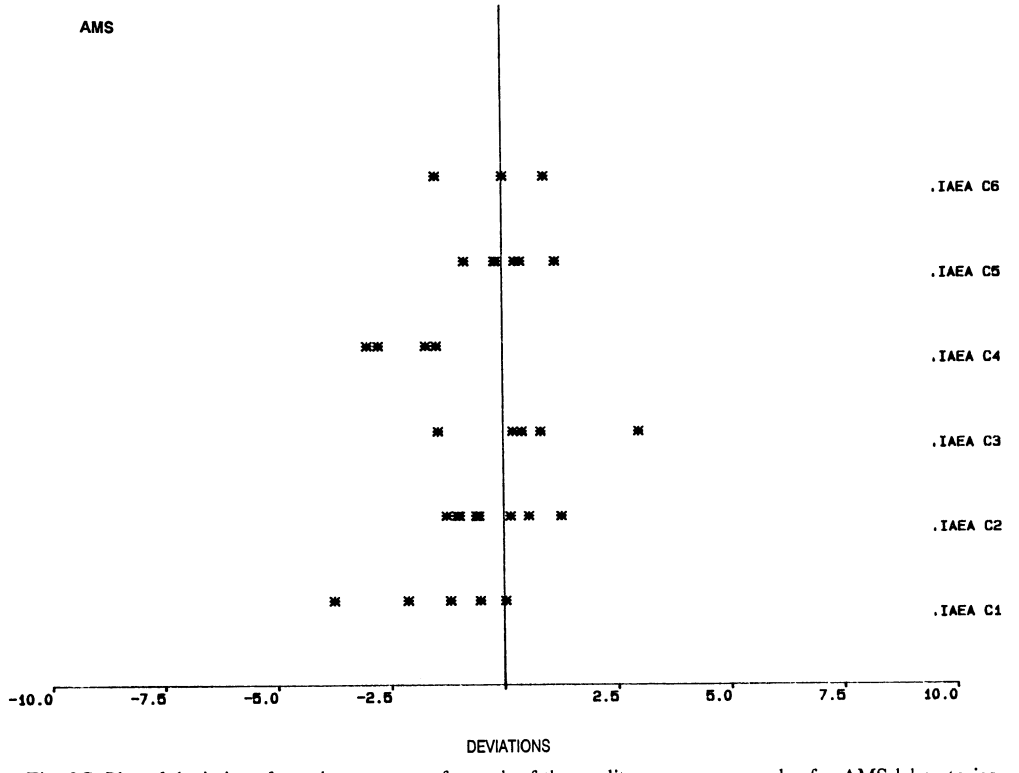


Fig. 9C. Plot of deviations from the consensus for each of the quality assurance samples for AMS laboratories

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

AMS results are less scattered. Outlying results in each of samples C-1 to C-6 can be seen, and they may occur in any of the three groups. This is confirmed by a formal analysis of variance where the 'average' differences do not differ significantly in each of the three laboratory groups. Of the six tests carried out, only C-1 and C-4 results approach statistical significance.

Analysis 2: Counter Technology

GPC. Nineteen sets of results came from labs using CO_2 as their counting gas, the remaining 18 using (CH_4 or C_2H_2), thus defining two groups of roughly equal size. We compared the mean deviation in the two groups for each of the six quality assurance samples and found no significant differences.

LSC. We found no evidence of statistically significant differences among different counter technologies used by LSC labs in the survey.

We did not investigate AMS technology in this analysis.

Analysis 3: Modern Standard

In this analysis, results were classified initially into four groups depending on the modern standard used. Of the four classifications proposed in Table 4, the 'other' category is excluded, as it occurs for only two of the samples, and only one lab uses 'other'. This analysis concentrates on ANU sucrose, Oxalic 1 and Oxalic 2.

For all of the reference samples, with the exception of C-2 and C-4, we carried out an analysis of variance and found that the modern standard is a significant factor, *i.e.*, the average 'deviation' is statistically significantly different across the groups. We interpret this as indicating that modern standard used is significant in explaining the variability in results. We feel this is an important finding, one that was postulated previously as a potential cause of excess variability (Scott *et al.* 1990), but for which there was no conclusive proof.

We further investigated this topic using samples C-3 and C-6, where the effects of any 'difficulties' with modern standard should be most easily observed, by considering the correlation between results. Figures 10A, B show the scatterplots for the results on these two samples. This analysis included *all* results, as we are no longer concerned with characterizing the ^{14}C activity of the reference materials, but rather in searching for clues that might indicate some of the sources of the variability in the results.

The overall correlation between the results on C-3 and C-6 is 0.376. However, if we select those labs using Oxalic 1, the correlation decreases to 0.039, but increases to 0.812 for the Oxalic 2 group. This result is highly significant for this latter group; if results on C-3 are high, associated results on C-6 also tend to be high. These findings require further investigation, perhaps through reconsideration of published data on the calibrations of Oxalic 2 and ANU sucrose.

Analysis 4: $\delta^{13}C$

Consensus values for $\delta^{13}C$ were also calculated and are referenced in the IAEA report (Rozanski 1991) and in Table 3. We evaluated $\delta^{13}C$ consensus values after consecutive rejection of outliers; the remaining data values lie within three standard deviations. However, in this analysis, we consider the relationship between the $\delta^{13}C$ where quoted and the deviation for the reference samples. We find no evidence of a significant relation between $\delta^{13}C$ and the deviation for any of the reference samples (correlations ranged from −0.287 to 0.042).

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

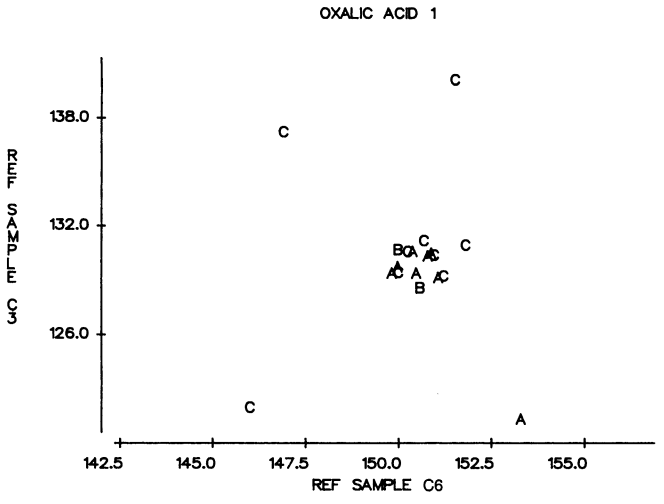


Fig. 10A. Scatterplot of results for reference samples C-3 and C-6 for laboratories using Ox I, coded by laboratory type A – GPC; B – AMS; C – LSC

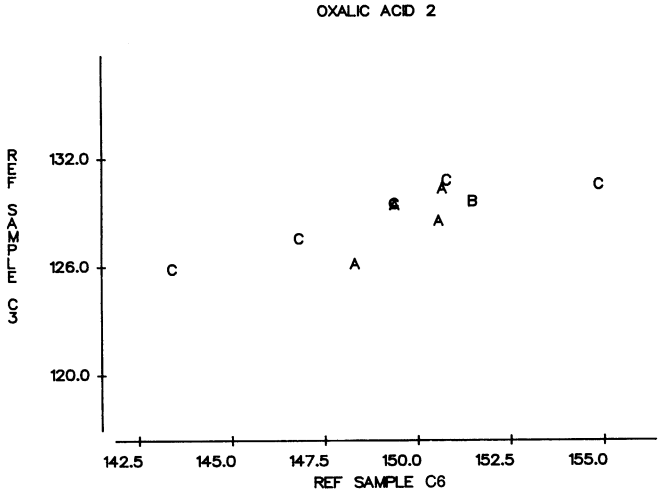


Fig. 10B. Scatterplot of results for reference samples C-3 and C-6 for laboratories using Ox II. See Fig. 10A.

Future Plans

These materials will be stored in the Agency as "IAEA ^{14}C Quality Assurance Materials," and will be available, upon request and free of charge, to all ^{14}C laboratories wishing to check the performance of their work. Following the recommendations of the meeting of experts held in February 1991, samples of these materials will be made available to any laboratory on a once-per-year basis only, and in quantities that will enable each laboratory to make at least two determinations. The quantity of the sample will be limited to a maximum of 25 g of elemental carbon, to preserve the stock and ensure continuity at least over the next ten years. We request that the recipients of these materials comment on their suitability and results.

Conclusions

The IAEA ^{14}C Intercomparison Exercise 1990, resulted in a precise evaluation of the ^{14}C concentration levels in five natural materials frequently used in ^{14}C laboratories. The analyses reported here were based on laboratory type, counting technology used, modern standard and $\delta^{13}C$ as factors

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press

in explaining the variation in the results. Our conclusions are: 1) no appreciable differences in the relative performances of different laboratory types on the individual reference samples; 2) no significant differences in performance due to the counter technology; 3) a significant indication that one important source of the variability in the results was the modern standard used; 4) no evidence that $\delta^{13}C$ was a significant factor. We feel Point 3 deserves further investigation.

The results of this work demonstrate the need for new reference materials within the international radiocarbon community. Its ability to capitalize on these developments in maintaining and improving its quality assurance will be further investigated in an international intercomparison to be organized in 1992 (TIRI, Scott *et al.* 1992).

ACKNOWLEDGMENTS

The assistance of IMEG, Vareggio, Italy, GSF, Institute of Hydrology, Neuherberg, Germany, a paper factory in Bergum, The Netherlands, in providing samples, is gratefully acknowledged.

REFERENCES

Aitchison, T. C., Scott, E. M., Harkness, D. D., Baxter, M. S. and Cook, G. T. 1990 Report on Stage 3 of the International Collaborative Program. *In* Scott, E. M., Long, A. and Kra, R. S., eds., Proceedings of the International Workshop on Intercomparison of ^{14}C Laboratories. *Radiocarbon* 32(3): 271–278.

Cook, G. T., Harkness, D. D., Miller, B. F., Scott, E. M., Baxter, M. S. and Aitchison, T. C. 1990 International Collaborative Study: Structuring and sample preparation. *In* Scott, E. M., Long, A. and Kra, R. S., eds., Proceedings of the International Workshop on Intercomparison of ^{14}C Laboratories. *Radiocarbon* 32(3): 267–270.

Currie, L. A. and Polach, H. A. 1980 Exploratory analysis of the international radiocarbon cross-calibration data: Consensus values and interlaboratory error. *In* Stuiver, M. and Kra, R. S., eds., Proceedings of the 10th International ^{14}C Conference. *Radiocarbon* 22 (3): 933–935.

Gonfiantini, R., Rozanski, K. and Stichler, W. 1990 Intercalibration of environmental isotope measurements: The program of the International Atomic Energy Agency. *In* Scott, E. M., Long, A. and Kra, R. S., eds., Proceedings of the International Workshop on Intercomparison of ^{14}C Laboratories. *Radiocarbon* 32(3): 369–375.

International Study Group 1982 An inter-laboratory comparison of radiocarbon measurements in tree-rings. *Nature* 298: 619–623.

_____1983 An international tree-ring replicate study. *In* Waterbolk, H. T. and Mook, W. G., eds., Proceedings of ^{14}C and Archaeology. *Pact* 8: 123–133.

Mook, W. G. 1990 Special report from the Glasgow Intercomparison Workshop on quality control and assurance. *Radiocarbon* 32(1): 107–108.

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 the 10th International ^{14}C Conference. *Radiocarbon* 22(3): 936–947.

Rozanski, K. 1989 Consultants' group meeting on the C-14 quality assurance programme. Vienna, IAEA: 7 p.

_____1991 Consultants' group meeting on C-14 reference materials for radiocarbon laboratories. Vienna, IAEA: 25 p.

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. *Radiocarbon*, this issue.

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

APPENDIX 1.

(A) *Weighted Estimation*

$x_i \sim N(\mu, \sigma_w^2 w_i^2)$

i.e., x_i is assumed to have a Normal distribution.

x_i = pMC measurement from lab i
w_i = quoted error (1σ)
μ = true reference sample activity

$$\hat{\mu} = \bar{x}_w = \frac{\sum x_i/w_i^2}{\sum 1.0/w_i^2}$$

^ indicates estimated value

$$\hat{\sigma}_w^2 = \frac{1}{n}\sum (x_i - \bar{x}_w)^2/w_i^2$$

$$\text{ese}(\hat{\mu}) = \hat{\sigma}_w \sqrt{\sum \frac{1.0}{w_i^2}}$$

all summations are from i = 1, ..., n.

In this model, an additional error term, σ_w has been incorporated, estimated on the basis of a weighted residual sum of squares. This term quantifies the lack of homogeneity in the group of results.

(B) *Test of Homogeneity* (Wilson & Ward 1981)

The test is based on $\hat{\sigma}_w^2$ evaluated above and takes the form:

Reject the hypothesis at the 5% level that the group of results (x_i, w_i) i = 1, ..., n is a homogeneous group if

$n\hat{\sigma}_w^2 > \chi^2(n\text{-}1, 0.95)$, *i.e.*, is greater than a value read from Chi-squared ($\chi 2$) tables.

https://doi.org/10.1017/S0033822200063761 Published online by Cambridge University Press