[RADIOCARBON, VOL. 34, NO. 3, 1992, P. 520–527]

FURTHER ANALYSIS OF THE INTERNATIONAL INTERCOMPARISON STUDY (ICS)

E. M. SCOTT[1], G. T. COOK[2], D. D. HARKNESS[3], B. F. MILLER[3] and M. S. BAXTER[4]

ABSTRACT. The major findings of the Intercomparison Study (ICS) have already been published (Scott, Long & Kra 1990), but a number of questions remain unresolved. We address here some issues of user and technical relevance, which include: 1) further investigation of the quoted errors and their relation to the perceived precision and accuracy, which is of interest to users of ^{14}C dates; 2) the analysis of the known-age wood samples provided in Stages 2 and 3 of the ICS; 3) an investigation of the corresponding $\delta^{13}C$ data base, of more technical relevance to laboratories.

INTRODUCTION

The original analysis of the data generated during the Intercomparison Study (ICS) was intended to address a number of key topics, in particular: 1) the role of the quoted error as a measure of internal consistency as indicated by the duplicate analyses; 2) the existence, or otherwise, of systematic biases and the role of the quoted error in adequately explaining any such interlaboratory variation; 3) a comparison of the performance of each laboratory type: liquid scintillation counting (LSC), gas proportional counting (GPC) and accelerator mass spectrometry (AMS).

To answer these questions, we evaluated three measures of laboratory performance, one that assessed systematic bias and two (error multipliers) that attempted to quantify the inter- and intralaboratory variation in a simple manner (Scott *et al.* 1990). The available data permit further investigation that may be useful to users of ^{14}C data.

First, we consider further the question of error, and describe some new evaluations, now completed, to explore more thoroughly the relationship of quoted error, error multiplier and absolute error. We also examine the results by grouping of quoted error.

Second, although ICS had three stages, only the 2nd and 3rd stages involved natural samples – Stage 2 provided homogenized, pretreated samples of cellulose and humic acid, both of which could be directly related to whole samples of wood and peat provided in Stage 3. In addition, the cellulose and wood in Stages 2 and 3 were provided by the Belfast Palaeoecology Laboratory and had been tree-ring dated. In the previous analysis, little use was made of the tree-ring dates; here we consider the spread of ^{14}C measurements in these samples and its relation to the "true" age.

Finally, we consider some of the additional information provided by laboratories, and their relation to the results, in particular, $\delta^{13}C$.

Further Investigation of the Level of Variability

In this section, we concentrate on results from Stage 3 and on the Internal Error Multiplier (IEM). This is calculated from the differences between the duplicate samples, and quantifies the reproducibility of results. (Appendix 1 defines the model and estimation of the IEM). In this sense, the IEM is an analog to the Level-3 error, which is "based on the statistical analysis of count rates of samples repeatedly reprocessed through the entire procedure in the lab" (Long & Kalin 1990: 330).

[1]Department of Statistics, University of Glasgow, Glasgow G12 8QW Scotland
[2]Radiocarbon Laboratory, SURRC, East Kilbride, Glasgow G75 OQU Scotland
[3]NERC ^{14}C Laboratory, SURRC, East Kilbride, Glasgow G75 OQU Scotland
[4]IAEA International Marine Laboratory, MC 98000 Monaco

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

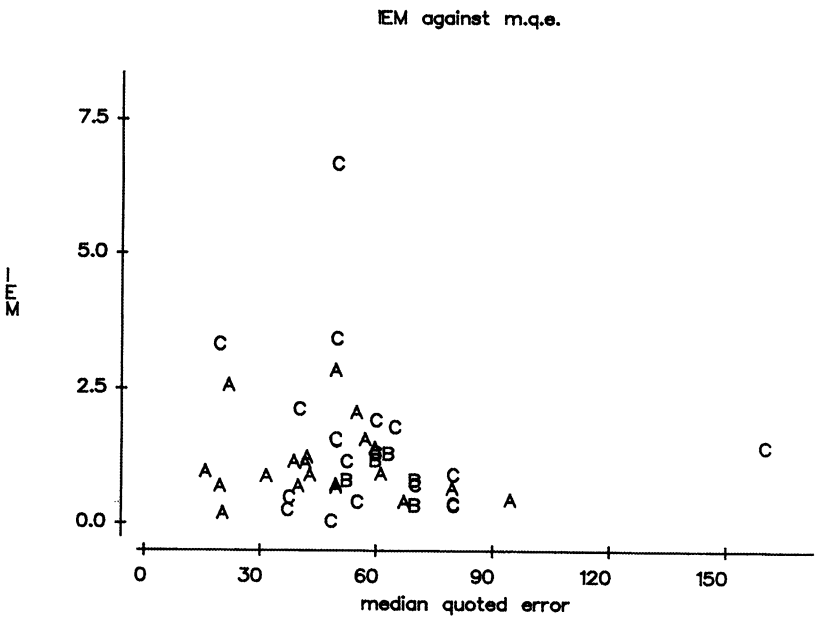

Fig. 1. Internal error multiplier plotted against median quoted error. Points are labeled as follows: A – GPC; B – AMS; C – LSC.

A total of 44 labs submitted results for Stage 3. Figure 1 shows the individual laboratory IEM plotted against the median quoted error (mqe). The correlation co-efficient, estimated as –0.130, is not statistically significant, so there is no evidence of a relation between mqe and IEM. We find no evidence of differing performances in IEM among the three lab types, nor do we find any significant difference in IEM if we group laboratories according to their mqe. For this work, we have chosen the following categorization to ensure adequate numbers of laboratories in each group:

$mqe \leq 30$
$30 \leq mqe \leq 50$
$50 < mqe \leq 70$
$mqe > 70$.

Tables 1 A and B summarize the IEM calculations.

TABLE 1A. Summary of the Distribution of IEM in Stage 3, by Lab Type

	Lab type		
	GPC	AMS	LSC
Average IEM	1.08	0.9	1.55
No. of labs	20	6	18

TABLE 1B. Summary of the Distribution of IEM in Stage 3, by mqe

	Median quoted error (mqe)			
	$mqe \leq 30$	$30 < mqe \leq 50$	$50 < mqe \leq 70$	$mqe > 70$
Average IEM	1.52	1.53	1.10	0.66

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

TABLE 1C. Index of Homogeneity on Duplicate Samples

Sample		
Wood A	Shell	Peat
1.65	2.04	
1.64*	1.07*	1.37

*After removal of a single outlier

In general, we conclude that the quoted error for most laboratories adequately described the reproducibility of their results. However, this is not the only component of variability of interest, since we must also assess the comparability of results, and hence, the component due to different laboratories dating the same material. One measure of comparability that can be simply quantified is that of an *index of homogeneity* σ_w (Ward & Wilson 1978; Wilson & Ward 1981) (see Appendix). The index was originally defined to compare simultaneity and combine a group of ^{14}C dates typically from the same laboratory. We apply this technique to groups of dates of the same material, but from different laboratories. The index defines an overall level of variation, and is based on a model that assumes, on average, laboratories are measuring the same ^{14}C activity, but also permits the variability around the true mean level for lab i, to be modeled as $\sigma_w^2 s_i^2$, where s_i represents the quoted error. Thus, the index is a sample rather than laboratory-specific measure, and includes a component of variation due to the natural variability within the *sample* material. The IEM, based on duplicates, provides the laboratory equivalent of the index.

Initially, the index had been calculated on the duplicate differences for the wood, shell and peat samples. Table 1C shows the results. Since the index generally exceeds 1, we see some evidence of overdispersion in the results.

The index was then calculated for all the samples in Stage 3, with duplicate results no longer combined. Table 2A shows the index and the estimate of the 'true' ^{14}C age. Again, we see evidence of overdispersion. All the wood samples have an index of *ca.* 2, the peat sample gives the lowest value at 1.80 and the shell sample, the highest, at 3.04. This ordering is further supported by the known provenience of each sample. The peat samples were milled before dispatch, the wood samples consisted of either 20-year or 30-year sections; the shells were generally whole, and of the same species taken from a large deposit (Cook *et al.* 1990), but were believed to have come from a well-defined archaeological context. This analysis was repeated for two subclassifications of the data – by lab type and by median quoted error. Table 2B shows that GPC and LSC labs generally measure the same ^{14}C activity, but that the index for LSC labs tends to be higher than those for GPC labs. AMS lab results have noticeably lower indices, indicative of 1) a more consistent set of results, and 2) more appropriate quoted errors. Interestingly, we see some differences in the average ^{14}C activity among the different lab types, this being most pronounced in the two younger wood samples (B and C).

Table 2C shows the indices calculated for the different subgroups of labs, classified according to mqe. No overall pattern emerges from this table, other than evidence of overdispersion in the results. The analytical approach described here assumes that the overdispersion can be modeled as additional random variation. Previously, we estimated systematic components of intralaboratory variation, namely bias. The latter analysis also showed evidence of significant differences among laboratories. More recently, the IAEA 1990 intercomparison (Rozanski *et al.* 1992) also revealed overdispersion of results, which may be linked to difficulties involved in calibration to modern standards.

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

TABLE 2. Index of Homogeneity for All Stage 3 Samples

		Sample			
	Wood A	Wood B	Wood C	Shell	Peat
A. Overall					
Weighted average (yr BP)	2209	313	127	660	3377
Index	2.06	2.52	2.18	3.04	1.80
B. By lab type					
GPC wt. average	2212	317	139	659	3383
GPC index	1.97	2.64	1.86	2.63	1.43
AMS	2170	285	45	641	3395
	1.55	2.00	1.78	0.93	1.01
LSC	2215	306	113	663	3365
	2.25	2.51	2.43	3.86	2.23
C. By mqe					
≤ 30 wt. average	2221	317	135	676	3392
≤ 30 index	1.77	0.82	1.75	1.28	2.1
30 < mqe ≤ 50	2191	308	116	657	3336
	2.58	3.6	2.02	3.83	1.6
50 < mqe ≤ 70	2221	305	109	617	3415
	1.27	1.36	2.5	3.24	1.87
mqe > 70	2212	284	138	643	3373
	1.84	1.44	2.18	1.36	1.32

The Analysis of Known-Age Wood Samples from Stages 2 and 3

Table 3 summarizes the known-age wood samples and their corresponding tree-ring dates, as well as the consensus ^{14}C ages used in the evaluation of the results. Table 4 shows an extract from the calibrations of Pearson and Stuiver (1986), with the appropriate high-precision ^{14}C dates corresponding to tree-ring-dated samples in the same time span as those available in ICS. If we first consider Samples B and C, we see that the consensus ^{14}C values are close to the high-precision values, and that the 'true' difference of 200 years agrees well with the observed average difference of 184 years. Figure 2A shows a histogram of the differences, with a clear mode at 200 years. Figure 2B shows a scatterplot of the results, the theoretical line, Wood C = 200 + Wood B, indicating that the fit is good. This analysis, as did the duplicate analysis, demonstrates the ability of laboratories to achieve internally consistent results. If we now consider the duplicate samples

TABLE 3. Known-Age Wood Samples

Wood Stage 2 – cellulose, extracted from tree rings dated to 241–260 BC
Stage 3 – wood, provided in duplicate, tree-ring date of 221–240 BC
– wood, single sample, tree-ring date of AD 1521–1550
– wood, single sample, tree-ring date of AD 1841–1870

Consensus ^{14}C age values (yr BP)

Stage 2		Stage 3	
Cellulose	Wood A	Wood B*	Wood C*
2250	2218	300	120

*Wood B and C were not provided in duplicate.

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

TABLE 4A. High-Precision Calibrated Results for Known-Age Samples

Sample	Tree-ring age	Corresponding ^{14}C ages (Pearson & Stuiver 1986)		
		Mid-	^{14}C age	Average
Cellulose	241–260 BC	250 BC	2195 ± 16	2223
	mid-250 BC	270 BC	2251 ± 16	
Sample A	221–240 BC	210 BC	2206 ± 13	2195
	mid-230 BC	230 BC	2183 ± 17	
Sample B	AD 1521–1550	AD 1535	314 ± 16	293
	AD mid-1535	AD 1525	273 ± 14	
Sample C	AD 1841–1870	AD 1840	95 ± 10	
	AD mid-1855			

TABLE 4B. Calibrated Results

	Cellulose	Wood A	Wood B
^{14}C age	2250	2218	300
Assumed σ	10	10	10
95% interval	386-365 cal BC	369-352 cal BC	cal AD 1523-1565
	279-262 cal BC	311-271 cal BC	cal AD 1634-1642
		269-238 cal BC	
		230-210 cal BC	

Note: Wood C could not be calibrated.

of cellulose and whole wood, the corresponding high-precision ^{14}C ages are 2223 and 2195 BP. The consensus values agree well, supporting the previous use of these consensus values in estimating bias. Further, the ^{14}C results for Sample A are in broad agreement with the ^{14}C values corresponding to tree-ring age 221–240 BC in the high-precision calibration work of Pearson and Stuiver (1986).

Table 4B shows the results of calibrating the consensus ^{14}C values using the probabilistic approach (van der Plicht & Mook 1989). The calibrated results for Wood A and B overlap the known tree-ring date, although for Sample A, they cover a broad range, and include multiple solutions. The calibrated results for the cellulose do not, in fact, include the tree-ring dates.

^{13}C and Its Influence

The final factor considered here is the $\delta^{13}C$ values quoted for each sample. Again, we concentrate on Stage 3, before looking at the related samples in Stage 2. We investigate the level of variability in $\delta^{13}C$, and how it relates to the overall variability in the ^{14}C results.

Table 5 summarizes the $\delta^{13}C$ values for each sample and the correlations between ^{14}C age and $\delta^{13}C$. We find no evidence of a significant linear relationship between the two, thus, $\delta^{13}C$ provides little clue to the source of overdispersion in the results.

If we now compare the $\delta^{13}C$ values of the cellulose and humic acid in Stage 2, with those for whole wood A and peat in Stage 3, we find a small significant difference, (the whole wood $\delta^{13}C$ is lighter than that for cellulose, and the humic acid $\delta^{13}C$ is lighter than that for the whole peat).

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

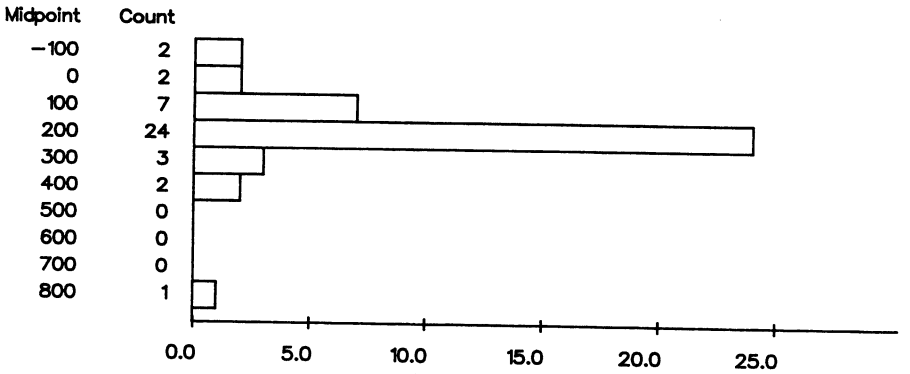

Fig. 2A. Histogram of the differences between wood samples B and C

Fig. 2B. Scatterplot of results for wood samples B and C showing theoretical ^{14}C relation

TABLE 5. $\delta^{13}C$ Summaries

	Stage 2		Stage 3				
Sample	Cellulose	Humic acid	Wood A	Wood B	Wood C	Shell	Peat
Mean $\delta^{13}C$ value	−24.1	−28.60	−25.15	−25.10	−25.90	1.30	−28.30
Correlation with ^{14}C age	−0.102	0.165	0.212	0.122	0.255	0.254	0.098

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

CONCLUSIONS

We have further investigated the data base of results from the ICS and in particular, the variability in the results for each sample. We find evidence of overdispersion of results, which is not purely sample-derived, and that the level of inhomogeneity apparent in the results is not dependent solely on laboratory type.

Analysis of the known-age material confirms the use of the previously defined consensus values, and demonstrates that, in the context of intercomparison, provided that a sufficient number of results are available, a consensus value defines an appropriate baseline.

The $\delta^{13}C$ values show a non-significant correlation with the ^{14}C ages. This finding might be expected, given the low level of variation among the $\delta^{13}C$ values themselves.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the Science and Engineering Research Council, and in particular, the contribution of all participant laboratories.

REFERENCES

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 Radiocarbon Laboratories. *Radiocarbon* 32(3): 267–271.

Long, A. and Kalin, R. M. 1990 A suggested quality assurance protocol for radiocarbon dating laboratories. *In* Scott, E. M., Long, A. and Kra, R. S., eds., Proceedings of the International Workshop on Intercomparison of Radiocarbon Laboratories. *Radiocarbon* 32(3): 329–335.

Pearson, G. W. and Stuiver, M. 1986 High precision calibration of the radiocarbon time scale 500–2500 BC. *In* Stuiver, M. and Kra, R. S., eds., Proceedings of the 12th International ^{14}C Conference. *Radiocarbon* 28(2B): 839–862.

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. *Radiocarbon*, this issue.

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. *In* Scott, E. M., Long, A. and Kra, R. S., eds., Proceedings of the International Workshop on Intercomparison of Radiocarbon Laboratories. *Radiocarbon* 32(3): 309–321.

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.

van der Plicht, J. and Mook, W. G. 1989 Calibration of radiocarbon ages by computer. *In* Long, A. and Kra, R. S., eds., Proceedings of the 13th International ^{14}C Conference. *Radiocarbon* 31(3): 805–816.

Ward, G. K. and Wilson, S. R. 1978 Procedures for comparing and combining radiocarbon age determinations: A critique. *Archaeometry* 20(1): 19–31.

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

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press

APPENDIX 1.

(I) *Internal Error Multiplier* (IEM)

Notation: X_{1j} and X_{2j} are duplicate results on the j^{th} sample for an individual laboratory with corresponding quoted errors, S_{1j} and S_{2j}.

We assume $X_{1j} \sim N(\mu_j,\ \theta S_{1j^2})$ where μ_j is the 'true age'
$X_{2j} \sim N(\mu_j,\ \theta S_{2j^2})$ $\sqrt{\theta}$ is the IEM

i.e., X_{1j} and X_{2j} are Normally distributed
let $d_j = X_{1j} - X_{2j}$
then $d_j \sim N(0,\ \theta S_{j^2})$ $S_{j^2} = S_{1j} + S_{2j^2}$.

$$\text{Likelihood} = \frac{1}{(2\pi)^{\frac{n}{2}}}\,\theta^{-\frac{n}{2}}\prod_j S_{j^{-1}}\,\exp\left[-\tfrac{1}{2}\sum\frac{d_{j^2}}{\theta S_{j^2}}\right]$$

$$\text{log likelihood} = \tfrac{n}{2}\log\theta - \sum_j \log S_j - \tfrac{1}{2}\sum\frac{d_{j^2}}{\theta S_{j^2}}.$$

The value of θ, which maximizes the likelihood function is given by

$$\frac{\delta}{\delta\theta} = \frac{n}{2\theta} + \tfrac{1}{2}\sum\frac{d_{j^2}}{S_{j^2}}\cdot\frac{1}{\theta^2} = 0,$$

i.e., $$\frac{1}{\theta}\left[-\frac{n}{2} + \frac{1}{2\theta}\sum_j\frac{d_{j^2}}{S_{j^2}}\right] = 0,$$

i.e., $$\frac{1}{\theta}\sum_j\frac{d_{j^2}}{S_{j^2}} = n,$$

hence, IEM $= \sqrt{\dfrac{\sum\frac{d_{j^2}}{S_{j^2}}}{n}}$.

(II) *Index of Homogeneity* σ_W (Ward & Wilson, 1981)

Data: $X_1, \ldots, X_n$
$s_1, \ldots, s_n$ $\quad x_i$, ^{14}C date, and s_i 1σ error

$$\text{then} \quad \bar{x}_W = \frac{\sum_i\frac{x_i}{s_{i^2}}}{\sum\frac{1.0}{s_{i^2}}},$$

$$\text{hence} \quad \hat{\sigma}_W = \sqrt{\frac{1}{n}\frac{\sum(x_i - \bar{x}_W)^2}{s_{i^2}}}.$$

https://doi.org/10.1017/S0033822200063773 Published online by Cambridge University Press