[RADIOCARBON, VOL 32, NO. 3, 1990, P 259-265]

# AN OVERVIEW OF SOME INTERLABORATORY STUDIES

E M SCOTT[1], M S BAXTER[2], T C AITCHISON[1], D D HARKNESS[3] and G T COOK[2]

ABSTRACT. Many interlaboratory studies have been made in the $^{14}C$ community at irregular intervals over the past ten years. At times, the results from these studies have been contentious, mostly because of the lack of consistency in their findings. The importance of regular exercises has become particularly acute due to the large number of operating laboratories and the diversity of their methodologies. Hence, we briefly review the studies that have been made in the 1980s, focusing on those in which our laboratories participated. These include the $^{14}C$ Interlaboratory Comparison in the UK (Otlet *et al* 1980), the International Comparison (ISG 1982, 1983) and the first two parts of the current International Collaborative Program (Scott *et al* 1989a, b). The development of each study, its findings and shortcomings, are highlighted in order to assess the concordance of the conclusions.

## INTRODUCTION

A number of collaborative studies involving subgroups of the $^{14}C$ community have been made in the last decade. These studies can be distinguished from the familiar intercalibration of standards (eg, Currie & Polach 1980) by the type of sample used and, generally, the number of participating laboratories.

Large collaborative trials serve several very important functions in the scientific community. First, they provide laboratories with an opportunity to perform a large number of cross checks on routine samples where the results are not known beforehand; second, they provide a rigorous, predetermined protocol with clearly defined aims; third, they demonstrate to the user community the commitment and care taken in providing accurate and precise $^{14}C$ age determinations.

Currently, one of the largest collaborative studies undertaken by the $^{14}C$ community is reaching completion and it seems particularly relevant at this stage to review the previous studies with a view to extracting key pointers for future collaborative work.

With this in mind, a reconsideration of the British Museum/Harwell study (Otlet *et al* 1980), the International Comparison Study (ISG 1982, 1983) and Stages 1 and 2 of the current study (Scott *et al* 1989, in press) seems appropriate.

## BRITISH MUSEUM/HARWELL STUDY

### *Study Organization and Design*

The study (Otlet *et al* 1980) was organized jointly by the British Museum and Harwell $^{14}C$ laboratories and involved a total of six UK laboratories. It was designed to investigate the problems of sample preparation, the comparability of results (over a wide range of sample ages) and their presentation.

The sample material selected was benzene and five levels of $^{14}C$ activity were chosen, equivalent to 20,000, 10,000, 5000 and 2000 BP, and 200% modern. The benzene was prepared in the Harwell laboratory and rigorous tests of purity were made. The laboratories were not informed of the sample activities, but were given an approximate guide to the highest level. One further sample was included which had been used to dilute all the prepared samples.

[1]Department of Statistics, Glasgow University, Glasgow G12 8QQ, Scotland
[2]Scottish Universities Research and Reactor Centre, East Kilbride, Glasgow G75 OQU, Scotland
[3]NERC Radiocarbon Laboratory, East Kilbride, Glasgow G75 OQU, Scotland

*Results*

Only the per cent modern or equivalent age results were published initially (partly to preserve the anonymity of participating laboratories). All results were in close agreement. The between laboratory (or external) variability, as measured by the standard deviation of all results for a single sample agreed well with the error commonly quoted on a $^{14}C$ date of similar age. No data points were discordant and no results were rejected. This seemed a further indication that the quoted errors were good estimates of the true errors.

The authors concluded on the basis of the results that interlaboratory alignment could be maintained over a wide range of ages and that the distribution of results supported the commonly quoted errors. The authors, however, do clearly state that, "in most cases, these estimates refer to only part and not the full processing....".

This was one of the first collaborative studies that included samples, the ages of which were not known beforehand and it was intended as a forerunner to further work. It involved only one sample material, benzene, but a wide range of sample ages. It was completed over a relatively short period, and considered only one stage in the dating procedure, the counting process.

In 1981, a larger study followed, involving 20 laboratories from around the world, which used typical sample material, in this case, wood, requiring full processing, including pretreatment.

## INTERNATIONAL TREE-RING STUDY

*Study Organization and Design*

This second study (ISG 1982, 1983) was designed to quantitatively assess the experimental variability in routine $^{14}C$ dating. A total of 20 $^{14}C$ laboratories received a set of eight tree-ring samples taken from a short floating chronology spanning 200 years provided by Dr A Heyworth, each sample being identified on a tree-ring width plot. Participants were asked to treat the samples routinely, general (non-specific) instructions concerning pretreatment were issued and questions concerning the calculation of the commonly quoted errors were included. Results were to be returned within eight months of the sample dispatch.

*Results*

The results returned by the participating laboratories all overlapped in the age range 4800 - 5200 BP. Figure 1 shows the age determinations with the $2\sigma$ error estimates quoted. Evidence of considerable variability is apparent in this diagram, eg, for individual samples, results differ by 310 to 730 years.

The analysis of the results addressed three main questions:

1. an assessment of systematic bias for individual laboratories
2. an assessment of the observed variability and its relationship to the claimed variability (*ie*, quoted errors)
3. a consideration of the implications of the study findings for users.

The study found that many laboratories showed systematic biases of up to 200 years relative to the "known age." Figure 2 shows the point estimates for each laboratory and a 95% confidence interval (*ie*, range of plausible values) for the bias. As a measure of variability (in this case interlaboratory), an external error multiplier (EEM) and its confidence interval were calculated for each laboratory. (Note: the EEM relates the quoted laboratory error to the observed variability, and if the interval estimate for the multiplier does not contain the value of 1, this indicates that the quoted errors inadequately describe the observed overall variability). Figure 3 shows the estimated error multipliers and corresponding interval estimates with the reference value of 1 indicated. There

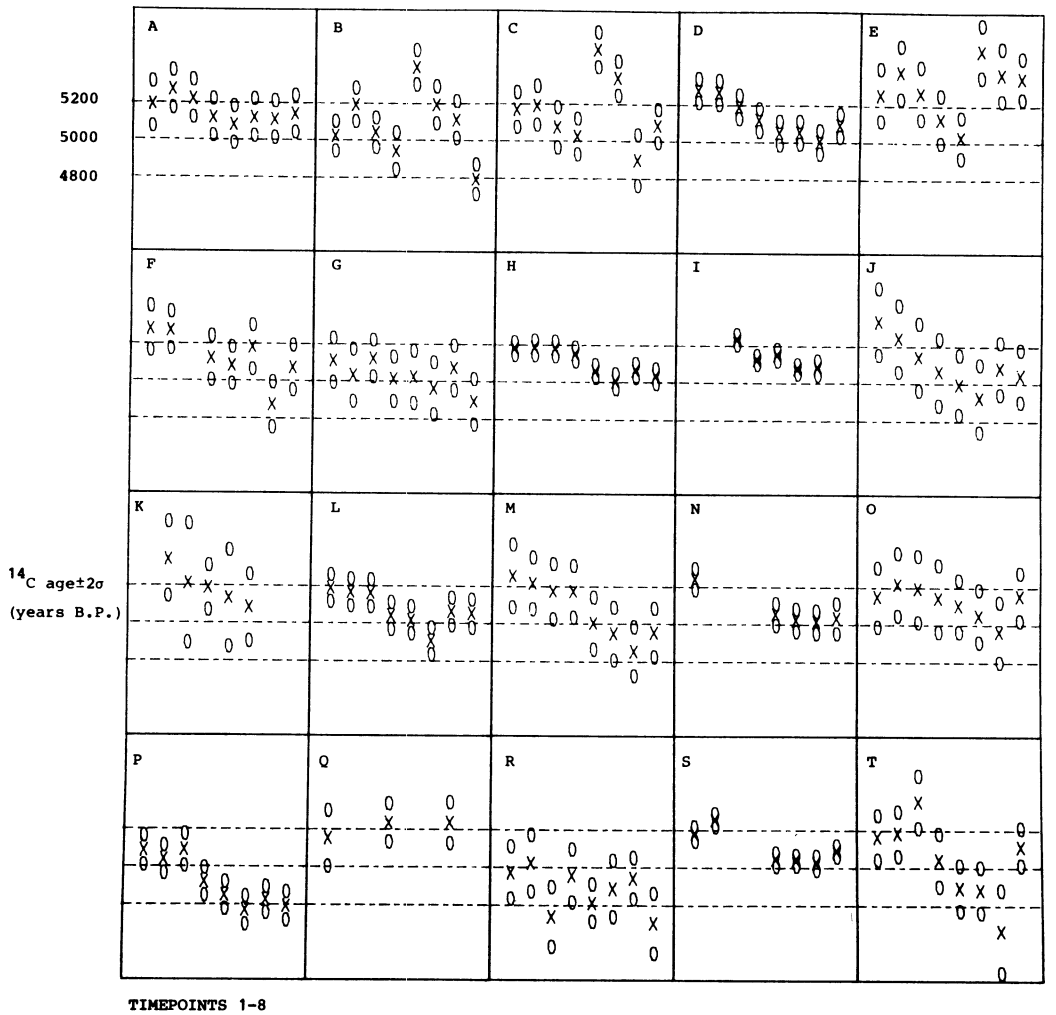


Fig 1. Results from first international collaborative study (ISG) (individual lab results are shown in each box, labeled A-T)

is some indication that the quoted errors inadequately describe the true variability of many laboratories. Fourteen of the laboratories have EEMs exceeding 1, of which seven have interval estimates which do not include 1.

The effects of the type of pretreatment and method of counting on the results were also investigated. Differences due to method of pretreatment were found to be small, whereas method of counting did appear influential. (Both gas counting and liquid scintillation laboratories were found equally likely to be biased, but gas laboratories were, in general, less variable than liquid scintillation laboratories (fewer gas laboratories had interval estimates for EEM which did not include 1).

The study provided clear indications for some laboratories, at least, of the existence of a systematic laboratory bias and of a level of variability not entirely explained by the quoted errors. We concluded that further investigation of errors should be undertaken and that more collaborative research involving different sample types and age ranges should be done.

https://doi.org/10.1017/S0033822200012856 Published online by Cambridge University Press

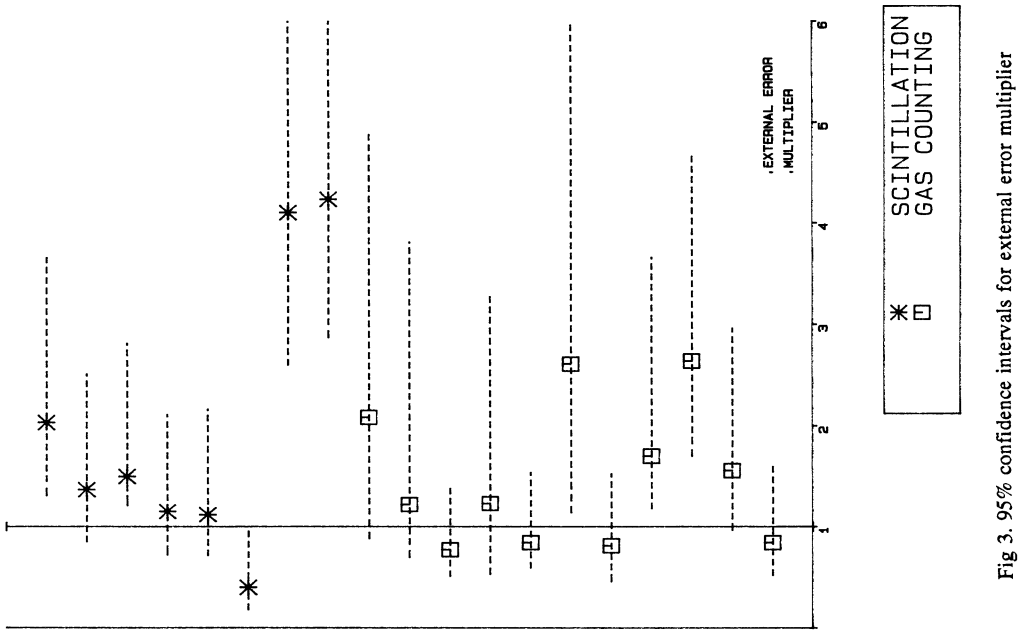


Fig 3. 95% confidence intervals for external error multiplier

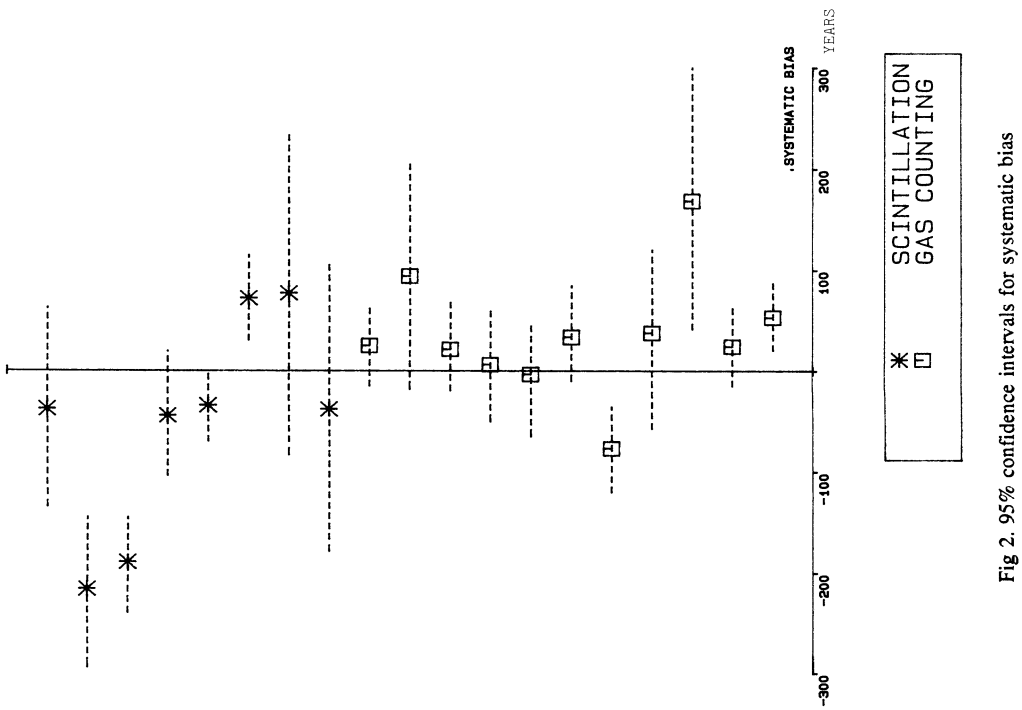


Fig 2. 95% confidence intervals for systematic bias

https://doi.org/10.1017/S0033822200012856 Published online by Cambridge University Press

Again, this study involved a single sample material, wood which required the full dating process. The sample age range was extremely restricted (200 years) and it too was completed over a relatively short time period.

Thus far, the two studies have shown considerable disagreement on the two major points of the relationship between interlaboratory variability and the quoted error and the existence of systematic biases. However, the studies have dealt with different processes in the dating procedure. Neither study is ideal as they suffer from problems of limited sample materials, large idealized sample sizes and short study periods. Both do agree, however, on the need for further work.

The challenge of designing and implementing a further study was taken up in late 1985 after discussion at the 12th $^{14}C$ conference at Trondheim. The new study finally got underway in 1986 with circulation of a detailed design protocol and clear statement of aims. In recognition of the size and importance of the undertaking, the Science and Engineering Research Council provided funding for the project.

The design of this study, the sample materials, their number and ages form the subject of a further paper (Cook *et al*, this issue). At this point, it is perhaps relevant to indicate that the new study involved a variety of sample materials requiring differing degrees of laboratory processing, and that the study was hierarchical, the processes of counting, synthesis and pretreatment being introduced in a sequential manner.

## INTERNATIONAL COLLABORATIVE STUDY

The aims of the study (Scott *et al* 1989; in press) were the quantitative assessment of variability and its attribution to the processes of counting, synthesis and pretreatment. Duplicate samples were introduced at each stage to allow assessment of internal reproducibility (*ie*, analytical precision). The study involved a wide range of sample materials of varying ages and was conducted over a four-year period.

Stage 1 used calcium carbonate and benzene samples to investigate the variability due to the counting process. In Stage 2, each laboratory was provided with homogenized, pretreated samples of shell , peat and cellulose to investigate the variability due to sample synthesis. Both stages have been completed and reported (Scott *et al*, 1989; in press).

### *Results*

To briefly summarize, for Stage 1, we concluded that the participating laboratories were internally consistent (the quoted errors adequately described the observed variability in the duplicate samples) but there was considerable variation amongst the laboratories. Figure 4A shows the disparity[4] data evaluated for the duplicate samples. The Stage 1 samples of benzene and carbonate are clearly indicated. The vast majority of values are < 1, but several outlying results are evident. Results for the first benzene sample show a considerable scatter. Figure 4B gives an indication of the level of interlaboratory variability, showing the offset data (the difference between the observed results and the 'true' value, here taken to be the consensus value of all the results). There is more variation in results for the two benzene samples than for the carbonate samples, indicating that interlaboratory variation is more pronounced for the liquid scintillation laboratories.

A similar analysis for Stage 2 is also included in Figures 4A and B. Again, the internal consistency of laboratories is evident from the disparity data but there is considerable interlaboratory variation in the offset data. A number of outlying results can be found in each

[4]A disparity is defined as the unsigned difference between duplicate samples divided by the square root of the sum of the squared quoted errors.

https://doi.org/10.1017/S0033822200012856 Published online by Cambridge University Press

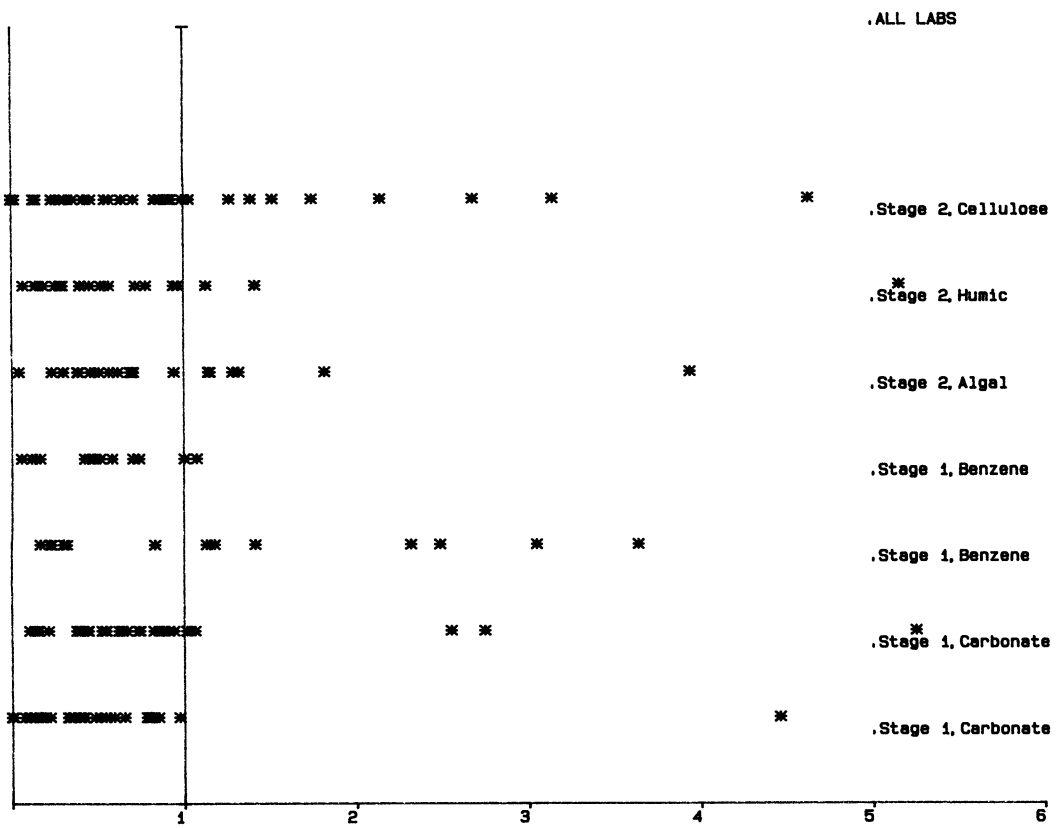


Fig 4A. Disparities at Stages 1 and 2 for all laboratories

stage and there is evidence of systematic biases. Considering the combined results for Stages 1 and 2, there is only a slight increase in the level of interlaboratory variation from Stages 1 and 2. Thus, it would appear that the major component of variability lies in Stage 1. Consideration of the disparity data also shows a slight increase from Stages 1 to 2; however, the internal precision of results is still adequately described by the quoted errors (the majority of disparities are < 1). We see little improvement in the results of the previous study (ISG).

## CONCLUSIONS

All the studies considered here have been concerned with the assessment of variability in routine $^{14}C$ dating. The BM/Harwell study primarily considered the counting process, the International Comparison Study looked at the dating procedure as a whole, whereas the current study has developed to consider the full process while allowing consideration of the individual processes.

Both the Harwell/BM and ISG studies involved single sample materials without replication allowing only a limited comment on the nature and source of variability in the results. The current study (with duplicates at each stage), although expensive in effort, has proved beneficial in the assessment of internal consistency. The hierarchical nature of the study also enabled us to investigate the components of variation. The earlier studies were all conducted over limited time periods, restricting attention to short-term sources of variation, whereas the new study, having developed over a longer time period, gives us time to assess longer-term sources of variation. An

https://doi.org/10.1017/S0033822200012856 Published online by Cambridge University Press

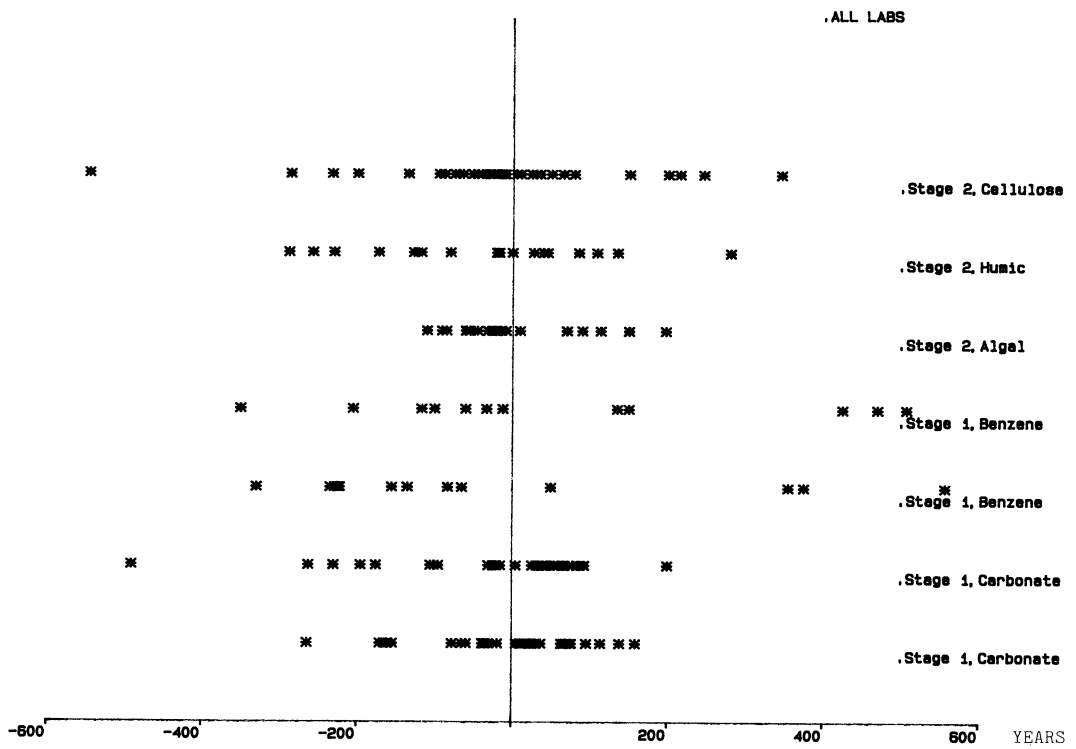


Fig 4B. Offsets at Stages 1 and 2 for all laboratories.

important aspect of all the studies has been the fact that the sample activities were not known in advance.

The general level of variability in the studies indicates that there is a requirement for an on-going experimental program, testing the full procedure as well as component processes in an exhaustive manner. This aim argues for a wide variety of sample materials in various states of preparation and covering as wide a sample age range as possible.

## References

Cook, GT, Harkness, DD, Miller, BF, Scott, EM, Baxter, MS and Aitchison, TC 1990 International Collaborative Study - Structuring and sample preparation. *Radiocarbon*, this issue.

Currie, LA and Polach, HA 1980 Exploratory analysis of the international radiocarbon cross-calibration data: Consensus values and interlaboratory error. Preliminary note. *In* Stuiver, M and Kra, RS, eds, Internatl $^{14}C$ conf, 10th, Proc. *Radiocarbon* 22(3): 933-935.

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, HT and Mook, WG, eds, $^{14}C$ and archaeology. *PACT* 8: 123-133.

Otlet, RL, Walker, AJ, Hewson, AD and Burleigh, R 1980 $^{14}C$ interlaboratory comparison in the UK: Experiment design, preparation and preliminary results. *In* Stuiver, M and Kra, RS, eds, Internatl $^{14}C$ conf, 10th, Proc. *Radiocarbon* 22(3): 936-946.

Scott, EM, Aitchison, TC, Harkness, DD, Baxter, MS and Cook, GT 1989a An interim progress report on Stages 1 and 2 of the International Collaborative Program. *In* Long, A and Kra, RS, eds, Internatl $^{14}C$ Conf, 13th, Proc. *Radiocarbon* 31(3): 414-421.

_____in press, Recent progress in the international calibration of radiocarbon labs. *In* Waterbolk, HT and Mook, WG, eds, Archaeology and $^{14}C$. *PACT*.