[RADIOCARBON, VOL 32, NO. 3, 1990, P 271-278]

REPORT ON STAGE 3 OF THE INTERNATIONAL COLLABORATIVE PROGRAM

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

ABSTRACT. This report on the third and final stage of the International Collaborative Program concentrates on the analysis of internal and external variability of ^{14}C dates obtained from samples involved in the full ^{14}C dating process. Thirty-eight laboratories took part in this stage with most producing 8 ^{14}C dates from 3 sets of duplicate material (eg, wood, shell and peat) and 2 single samples of wood of known ages 190 yr BP apart. From the 3 sets of duplicates for each laboratory, the internal precision of most laboratories was adequate; 6 labs grossly underestimated their internal reproducibility. From the ^{14}C determinations from the 5 distinct samples for each laboratory, we discovered significant systematic biases, often greater than 100 years, in 15 laboratories and even accounting for bias, 12 laboratories had significantly greater external variability than explained by their quoted errors. In total, 23 out of the 38 laboratories in this stage of the study, FAILED to meet these 3 basic criteria for an adequate performance in the production of ^{14}C dates.

INTRODUCTION

The ^{14}C dating community has acknowledged the importance of interlaboratory checks through its willingness to participate in a number of collaborative studies (Otlet *et al* 1980; ISG 1982, 1983). Recently, a third and considerably more ambitious project (Scott *et al* 1986) has been completed with the support of over 50 ^{14}C laboratories. We present here a report on the third and final stage of this study.

OVERALL AIMS

The major aims of the study have been to:

1. gain insight into the contribution of the various dating processes to the overall dating error
2. provide experimental comparison and validation of the diverse laboratory techniques used in dating
3. quantify uncertainties on routine results obtained by the modern generation of ^{14}C laboratories.

THE STUDY

Scott *et al* (1986) give full details of the completed study organization. Briefly, the study has three sequential stages, each introducing a further laboratory process. Stage 1 primarily involved the counting process. Stage 2 introduced sample synthesis and results were reported in September 1988 (Scott *et al* 1989). Finally, Stage 3 included full pretreatment as well as counting process and sample synthesis.

Study Sample Structure

The hierarchical sample structure is an important element of the study design. Table 1 illustrates the various sample materials offered throughout the program. Duplicate samples were included at each stage as well as four known-age samples in Stages 2 and 3. In total, each laboratory participating in all three stages will complete a minimum of 16 ^{14}C dates. Harkness *et al* (1989) give a full description of the sample preparations.

[1]Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland
[2]NERC Radiocarbon Laboratory, East Kilbride, Glasgow G75 0QU, Scotland
[3]Scottish Universities Research and Reactor Centre, East Kilbride, Glasgow G75 0QU, Scotland

TABLE 1
Sample materials

Stage 1 - benzene, calcium carbonate
Stage 2 - cellulose, algal lithothamnium, humic acid
Stage 3 - wood, shell, peat

Initially, 80 laboratories were invited to participate, of which 58 full participants were registered. A total of 52 laboratories returned results for Stage 1, 37 returned results for Stage 2 and 38 completed most of the Stage 3 samples. Table 2 shows the number of each type of laboratory, *ie*, gas counting (GC), accelerator mass spectrometric (AMS) and liquid scintillation counting (LSC), and the composition of the study group at all stages, in terms of counting technique. The fact that 26 of the laboratories in Stage 3 completed all 8 of the samples and another 12 missed, at most 2, is testimony to the considerable effort of the participating group.

TABLE 2
Laboratory types

Type	Stage		
	1	2	3
GC	23	18	20
AMS	8	5	5
LSC	20	14	13

We report here a summary of conclusions only from Stage 3 of this study; an overview of the complete study also appears in this issue (Scott *et al*). The latter paper deals with the relative importance of the possible sources of error through the three stages in the ^{14}C dating process (*ie*, pretreatment, sample synthesis and counting).

Duplicates and Statistical Methods

We sent one wood sample and both shell and peat samples as duplicate sets to each laboratory. We gain information on the *internal* consistency of a lab (*ie*, *intralaboratory* variability) from the *differences* between the duplicates for a particular laboratory. The main interest here was to ascertain whether the three differences agreed with the quoted errors of the lab itself. To this end, we defined the *disparity* of a single pair of duplicates as the unsigned difference divided by the square root of the sum of squared quoted errors. From these disparities, an *internal error multiplier* for the particular lab can be estimated as well as an appropriate 95% confidence interval for this multiplier (for details, see ISG 1982). If the estimate of this internal error multiplier for a lab considerably exceeds one (eg, >2) and, more importantly, if the *whole* of the 95% confidence interval exceeds one, then the lab is grossly underestimating its *internal* consistency, not to speak of interlaboratory comparability.

We then calculated the weighted averages (Ward & Wilson 1978) with appropriate standard errors (based on the possibly erroneous assumption that the quoted errors of the lab are correct) for each of the 3 pairs of duplicates and then *combined* with the 2 single wood samples to give a set of (normally) 5 dates and associated errors. These can then be used to assess the *external* consistency of the laboratory (*ie*, *interlaboratory* variability) as well as any systematic bias for a

particular lab. From these, we can calculate estimates of a possible systematic bias and an external error multiplier for the lab as well as associated 95% confidence intervals for these separately and jointly (ISG 1982).

Any laboratory that had a large bias estimate and the resulting 95% confidence interval *excluded* zero has a very definite problem, which may be traced to a specific cause (eg, contamination or a misaligned modern standard). This emphasizes the necessity of laboratory checking with either, or preferably both, external standards or other laboratories. Further, if the external error multiplier grossly exceeds 1 or, more particularly, the *whole* of its confidence interval is larger than 1 that laboratory is grossly underestimating its external consistency with respect to the other laboratories or known age, as appropriate.

Since we made no attempt to correct for the potential inaccuracy of quoted errors with respect to the internal consistency of a laboratory, the internal and external error multiplier estimates are likely to be positively correlated and should not be combined, except as a conservative estimate of the overall variability of the laboratory.

Internal Consistency Results

Figure 1 plots raw unsigned differences for each of the three pairs of duplicates and summarizes statistics of data partitioned into the three main categories of ^{14}C laboratories, *viz*, LSC, GC and AMS. The overall message is that, typically, 60 to 70-year differences exist, although differences of >120 years occurred in 1 sample in 4, and differences of >600 years were recorded for 2 samples! There appears to be little difference in the true internal variability of the three types of ^{14}C laboratories but, wood samples seem more variable than shell and, particularly, peat.

SUMMARY STATISTICS of UNSIGNED DIFFERENCES of DUPLICATES

MATERIAL	N	MEDIAN	MIN.	MAX.	UPPER QUARTILE
WOOD	36	65	1	610	120
SHELL	34	59.5	0	697	114
PEAT	31	70	6	300	120

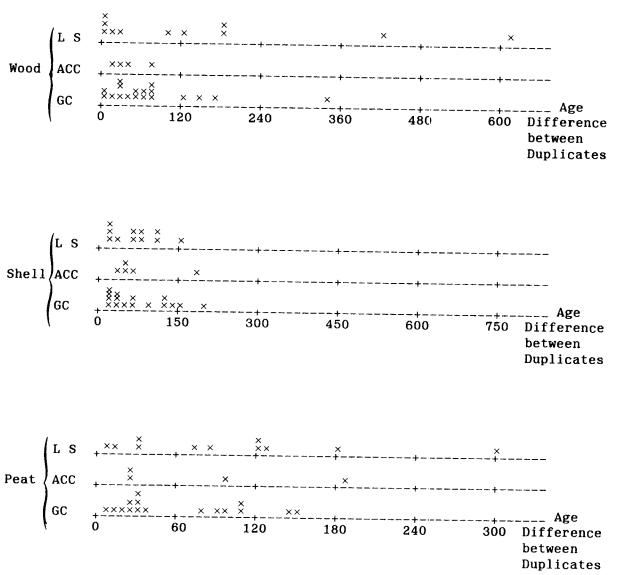

Fig 1. Summary statistics and dot plots of the differences between duplicate samples

When we correct these raw differences for their quoted errors, we obtain the disparities defined above. When these are combined over all available samples from a laboratory, we can estimate the internal error multiplier for that laboratory. Figure 2 shows these and their resulting 95% confidence intervals. The interval estimates are often wide simply because each is based on, at most, three differences.

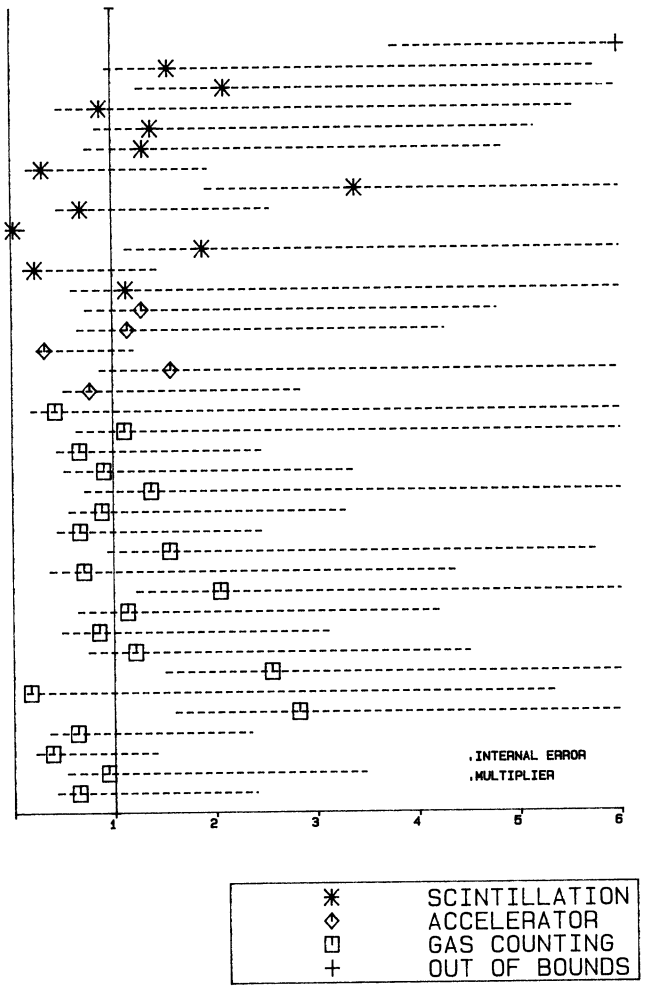

Fig 2. 95% confidence intervals for the estimates of the internal error multipliers for each laboratory.

Of the 20 gas counting labs, 12 have point estimates <1 and seem to offer conservative estimates of internal precision. However, 3 out of the 20 have internal error multipliers significantly >1 (*ie*, 3 labs have confidence intervals >1). Thus, these 3 laboratories are greatly underestimating their internal precision.

All five accelerator labs seem to estimate their internal consistency adequately.

SUMMARY STATISTICS for the 5 DISTINCT ^{14}C AGE DETERMINATIONS

(i.e. 3 weighted averages and 2 individual dates)

Material	Type of Laboratory: Liquid Scintillation	Accelerator	Gas Counting
WOOD (2185 BP)	2232 (2010, 2410)	2180 (2080, 2340)	2222 (2027, 2410)
WOOD (290 BP)	290 (-141, 670)	289 (117, 440)	312 (120, 730)
WOOD (100 BP)	163 (-326, 400)	5 (-130, 260)	140 (-3, 380)
SHELL	639 (-24, 1040)	642 (555, 705)	680 (470, 790)
PEAT	3400 (3217, 3630)	3420 (3320, 3475)	3406 (3215, 3465)

MEDIAN
(Values given are (MINIMUM, MAXIMUM) Age (years BP)

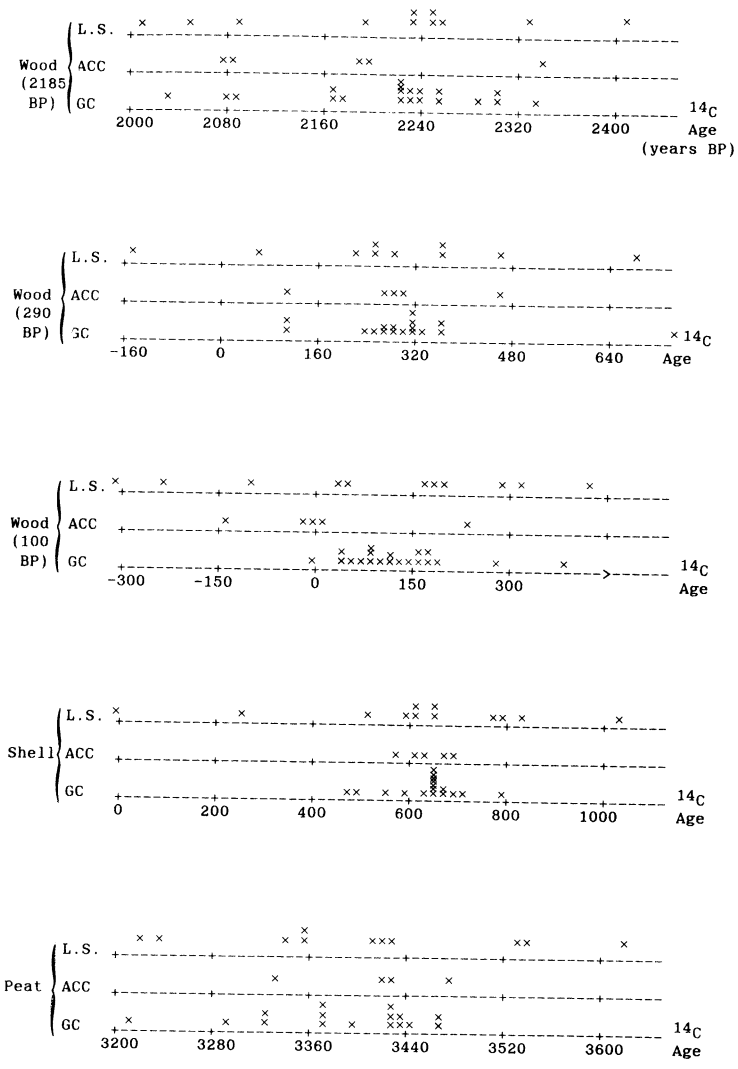

Fig 3. Summary statistics and dot plots of the 14C age determinations for all 5 distinct samples

Of the 13 liquid scintillation labs, 2 grossly overestimate their true internal reproducibility, whereas 3 have internal error multipliers significantly >1 and, thus, underestimate their internal reproducibility.

We can conclude that internal consistency is being adequately expressed in a majority of ^{14}C laboratories but there are quite a few that grossly underestimate their internal precision.

External Consistency Results

Figure 3 gives dot plots of the estimates of ^{14}C age for each of the (up to) five samples (*ie*, using the weighted average of duplicates for one of the wood samples, shell and peat in combination with the other two wood samples). Figure 3 also gives some summary statistics partitioned by laboratory type.

The main question, then, is, just how much more variability are we likely to see in results from liquid scintillation laboratories compared to the other two laboratory types? Note, however, that for all three laboratory types, there is no evidence of differing variabilities across the sample materials (*ie*, wood, shell and peat).

From Figure 3 it is clear that all three types of laboratories show more variability for more modern samples than for older samples (*ie*, 2000 - 3000 yr BP). This is particularly so for liquid scintillation laboratories. It is also worrisome to see the considerable systematic bias in accelerator laboratories for the most modern wood sample (*ie*, a difference of 135 years relative to the gas counting laboratories).

When the results from all the samples for a particular laboratory are combined, we can estimate any systematic bias and an external error multiplier for the laboratory. Figure 4 presents these labeled by laboratory type. Quite a few laboratories, mostly gas counting, perform very well with negligible bias and an error multiplier estimated around one. However, again, a number of laboratories exhibit gross deficiencies in either and/or both bias and error multiplier.

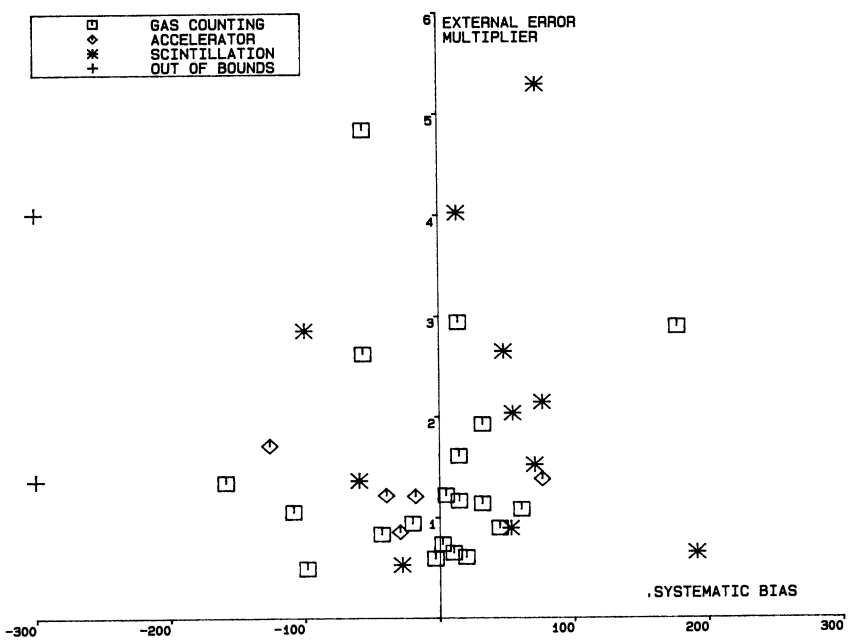

Fig 4. Estimated systematic bias plotted *vs* estimated external error multiplier for each laboratory

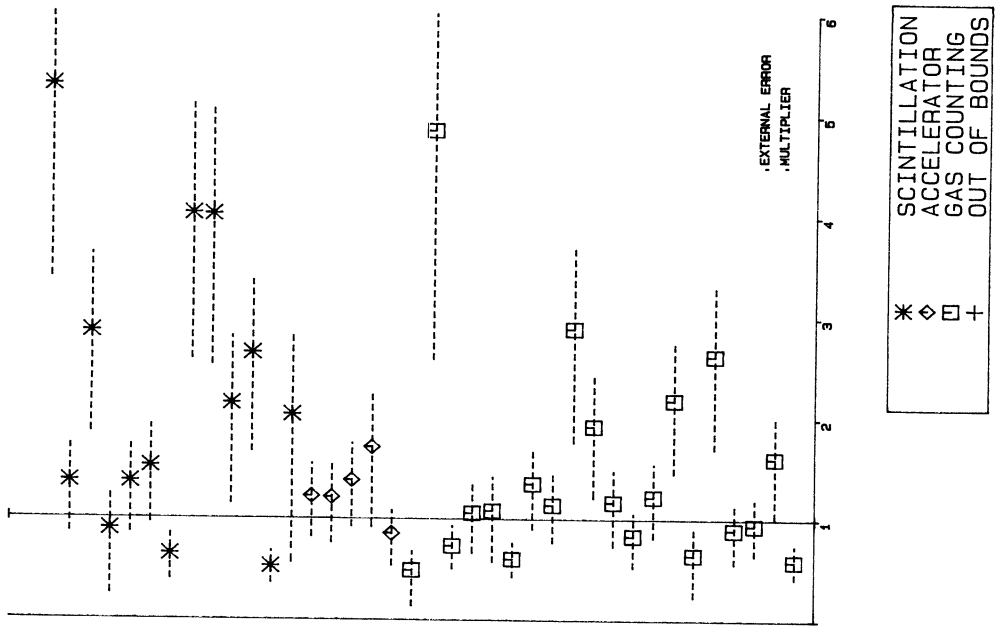

Fig 6. 95% confidence intervals for the estimates of the external error multipliers for each laboratory.

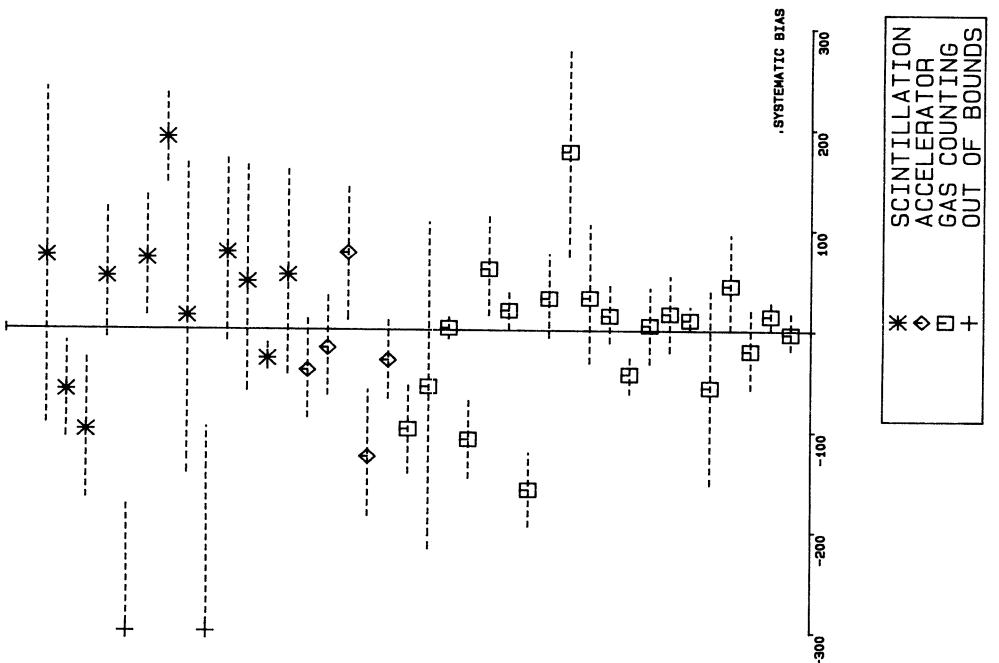

Fig 5. 95% confidence intervals for the estimates of systematic bias for each laboratory.

Figures 5 and 6 give, respectively, 95% confidence intervals for the systematic bias and external error multipliers. If the former *excludes* zero for a laboratory, then we can infer a significant systematic bias for that laboratory. If the latter interval is wholly greater than one, the laboratory is significantly underestimating its external precision.

The situation is extremely worrisome when 15 out of the 38 laboratories have a systematic bias significantly different from 0 and, of these, 9 appear to have a bias of well over 100 years - clear evidence that many laboratories are carrying out insufficient external checking.

Even allowing for any (significant or not) bias, the quoted errors for 12 laboratories are still not adequately explaining their external precision, as their external error multipliers are significantly greater than one (*ie*, these laboratories are grossly underestimating their external precision).

CONCLUSIONS

It seems reasonable to consider that a laboratory performs adequately if it has no significant systematic bias and assesses its internal and external variability adequately (*ie*, with error multipliers not significantly different from one). Accordingly, Table 3 presents the number of laboratories ***failing to meet*** each of these three requirements.

TABLE 3
Laboratory performance statistics
The numbers of laboratories FAILING to meet the designated requirement

Requirement	Internal error multiplier	Systematic bias	External error multiplier
LSC	3	7	6
AMS	0	2	0
GC	3	6	6

In total, only 15 of the 38 laboratories meet all 3 of these adequacy requirements, whereas 8 have 2 or more problems. Clearly, this is a cause for much concern in that **<50%** of the ^{14}C laboratories included in this study are meeting these basic requirements. It seems self-evident that the ^{14}C community will have to address the problems highlighted by this and other studies and commit to a continuous program of quality assurance monitored both internally and externally.

REFERENCES

Harkness, DD, Cook, GT, Miller, BF, Scott, EM and Baxter, MS 1989 Design and preparation of samples for the International Collaborative Study. *In* Long, A and Kra, RS, eds, Internatl ^{14}C conf, 13th, Proc. *Radiocarbon* 31(3): 407-413.

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

_____1983 An international tree-ring replicate study. *In* Waterbolk, HT and Mook, WG, eds, ^{14}C and Archaeology, Proc. *PACT* 8: 123-133.

Otlet, RL, Walker, AJ, Hewson, AD and Burleigh, R 1980 ^{14}C inter-laboratory comparisons 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, Baxter, MS, Aitchison, TC, Harkness, DD and Cook, GT 1986 Announcement of a new collaborative study for intercalibration of ^{14}C dating laboratories. *Radiocarbon* (28)1: 167-169.

Scott, EM, Baxter, MS, Harkness, DD, Aitchison, TC and Cook, GT 1989 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.

Ward, GK and Wilson, SR 1978 Procedures for comparing and combining radiocarbon age determinations: A critique. *Archaeology* 20: 19-31.