[RADIOCARBON, VOL 28, NO. 2A, 1986, P 571–577]

AN INTERCOMPARISON OF SOME AMS AND SMALL GAS COUNTER LABORATORIES

RICHARD BURLEIGH, MORVEN LEESE, and MICHAEL TITE
Research Laboratory, The British Museum, London WC1B 3DG, England

ABSTRACT. The performance of six laboratories with the capacity to date small samples (4 AMS and 2 small gas-counter laboratories) has been compared using 100mg samples of textiles from Ancient Egypt and Peru, with the British Museum laboratory acting as independent coordinator. This intercomparison was one of normal practices and has demonstrated that a coherent series of results can be obtained when several laboratories undertake blindfold measurements, although the occurrence of outliers emphasizes the continuing need for the dating of unusually important or controversial samples to be undertaken by a group of laboratories.

INTRODUCTION

The advent of successful techniques of ^{14}C dating using small samples (*ie*, accelerator and small-counter techniques) has made possible, among many other applications, the direct dating of highly valuable or unique objects for which the use of conventional ^{14}C techniques would be too destructive. In particular, the dating of the Shroud of Turin would now be possible in principle although it is generally agreed that any such measurement ought not to be undertaken by a single laboratory, or even by the use of one technique alone. Such an objective apart, there is an intrinsic scientific need to establish, in a controlled way, the variation among laboratories using small sample techniques, when the same, known-age, samples are measured blindfold.

With this in view (and with particular relevance to any proposal for dating the Turin Shroud) an intercomparison was planned in which two samples of textile of different age would be sent to 4 accelerator (AMS) and 2 small-counter laboratories by an independent laboratory whose role would also be to collate and report on the results, anonymity of the individual results being maintained. The British Museum was chosen to perform this task on the basis of impartiality, experience in ^{14}C dating, and ready access to suitable materials. The six ^{14}C laboratories taking part in the exercise were Arizona, Bern (using the Zurich AMS facility), Brookhaven, Harwell, Oxford, and Rochester, of which Brookhaven and Harwell were the two small-counter laboratories. Two samples, each weighing ca 100mg, 1 from Ancient Egypt (linen, 1st Dynasty, ca 3000 BC) and 1 from Peru (cotton, Chimu style, ca AD 1200), labeled, respectively, Sample 1 and Sample 2, were sent to each of these laboratories in May 1983. The provenance of each sample was stated, but their historical ages were not disclosed.

First results received for Sample 2 suggested that the material was of much more recent date than expected. This was probably erroneous as it turned out, but by agreement with all the participating laboratories, a third sample (Sample 3—cotton, Peruvian, Late Intermediate period, ca AD 1000–1400) was issued in February 1984 under the same conditions as previously, to replace Sample 2. In the event most laboratories taking part had measured Sample 2 and the analysis of these results is also included in this report.

The Egyptian sample, originally from Tarkhan, came from the Petrie

Collection at University College, London and the Peruvian samples came from the collection of the Museum of Mankind (Department of Ethnography, The British Museum). The Egyptian textile came from a tomb of the First Dynasty and could be dated archaeologically to ca 3000 BC. Contemporaneous linen from the same bulk of material from Tarkhan had previously been informally exchanged with a number of ^{14}C laboratories and the mean of the results obtained (4250 bp) was close to the mean value obtained for the Tarkhan sample in this intercomparison. The Peruvian textiles, though datable stylistically, were unprovenanced. Their ascription to particular periods is certain, but their precise dating within these periods is less secure. These materials were chosen for their homogeneity, and typical state of preservation, as well as their respective historical ages, and the individual samples were cut from the same area of each textile, away from selvedges or designs.

RESULTS

To preserve anonymity, the dates and their errors as reported by the laboratories are listed in Table 1 in order of increasing age for each sample and not by laboratory. The dates are given in years BP based on the 5570-year half-life and have been corrected for fractionation relative to PDB from measurements made either by conventional isotope mass spectrometry or during the course of the AMS determinations, even where no actual $\delta^{13}C$ values are given (errors of delta values were ± 0.5 to ± 3.0‰, the conventional measurements being the more precise; values in brackets are estimated). Errors are assumed to be equivalent to 1 standard deviation ($\pm 1\sigma$). One laboratory reported two measurements of Samples 1 and 2; one laboratory measured only Sample 1, reporting two measurements. All the reported measurements, including duplicates from the same laboratory, have been treated as separate, independent age estimates.

The following aspects of the data are discussed: 1) Are there any outliers, *ie*, individual values which are very unlike other measurements on the same sample? 2) Is the observed variation between measurements consistent with the quoted errors, after outliers have been removed? 3) Are the calibrated date ranges concordant with the historical dates ascribed to the samples?

TABLE 1
Results reported by laboratories

1 Egyptian (ca 3000 BC)			2 Peruvian (ca AD 1200)			3 Peruvian (AD 1000–1400)		
Date	Error	$\delta^{13}C$	Date	Error	$\delta^{13}C$	Date	Error	$\delta^{13}C$
3440	145	−26.6	80	110	−23.3	460	190	—
4100	110	−25.5	300	100	(−22.0)	600	100	−22.0
4170	90	−24.2	450	80	−25.6	620	100	−26.1
4230	100	−27.0	450	90	−26.2	670	130	−22.0
4340	170	—	530	110	−22.8	1550	90	−24.4
4350	110	−22.0						
4380	100	−24.1						
4517	140	(−24.1)						

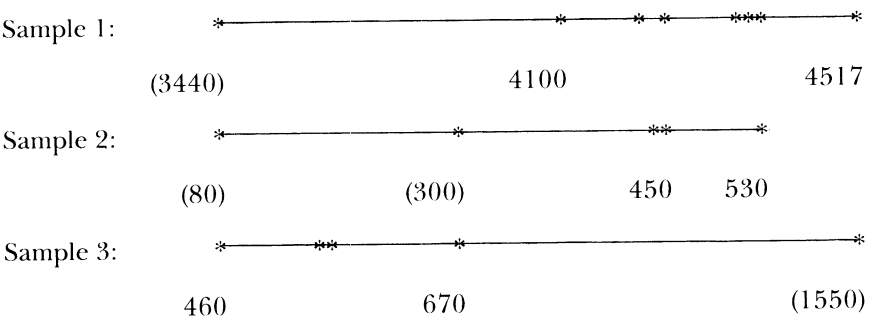
https://doi.org/10.1017/S0033822200007748 Published online by Cambridge University Press

The statistical techniques used are standard and are largely those used by Clark (1975) and Otlet *et al* (1980), except for the outlier tests described in the next section. There were insufficient data to compare individual laboratories or to compare the accelerator (AMS) method with the small-counter method, so this has not been attempted.

The means used to find jointly-determined calibrated date ranges are all unweighted. It was not felt that the individual dates should be weighted according to the inverse squared errors (as commonly recommended) because not all laboratories provided complete information about how the errors were computed and they may include different sources of error. For example, the errors of the AMS measurements may be estimated from variations within a series of multiple measurements, thereby including several sources of error, whereas the errors of the gas-counter measurements may depend more directly on counting statistics. If it could be ascertained that the errors reported by all the participating laboratories were compatible then weighting would be justified. Clark's 1975 calibration curve was used to find the calibrated dates; that of Klein *et al* (1982) gives similar results.

Outliers

The relative spreads of the dates for each sample are shown in the diagrams below (each sample has its own scale). Values suspected as being possible outliers are indicated in brackets and were tested for significance. The values closest to the suspected outliers and the end-points of the ranges are also shown.



The tests for outliers were based on the 'excess/range' statistics of Dixon, described by Barnett and Lewis (1978) who give tables of critical values. The underlying assumption in such tests is that the distribution of values is Normal with unknown variance. The variance is assumed to be unknown because, as noted earlier, the errors are not necessarily comparable with one another and may not include all possible sources of variation. It should also be noted that the numbers of dates are small so that only very high deviations can be rejected as definitely discordant.

Table 2 shows the results of outlier tests. For Samples 1 and 3, the test statistic is $O_1 = (x_n - x_{n-1})/\text{range}$, where x_n is the candidate outlier and

https://doi.org/10.1017/S0033822200007748 Published online by Cambridge University Press

TABLE 2
Results of outlier tests

Sample	Candidate outlier	Test statistic	Probability of higher value of O
1	3440	$O_1 = 0.6$	<1% (significant)
2	80	$O_2 = 0.8$	>5% (not significant)
3	1550	$O_1 = 0.8$	<1% (significant)

x_{n-1} the value next in order of magnitude. For Sample 2, it is

$$O_2 = (x_n - x_{n-2})/\text{range},$$

where x_{n-2} is the next but one in order of magnitude; O_2 thus avoids the other possible outlier, 300. (In the notation of Barnett and Lewis, O_1 and O_2 are equivalent to N_7 and N_{11}).

The outlier tests do not show overwhelming evidence that the value of 80 lies outside the expected range of the other Sample 2 measurements. However, 3440 and 1550 are highly discordant and have been excluded from the subsequent analysis. According to those concerned these aberrant results can be explained as being due to the use of non-standard pretreatment procedures and the measurements have been repeated using fresh samples of the same materials. A summary of this explanation and the results of the further measurements are given in APPENDIX 2.

The overall variation of the measurements of Sample 2 is high compared to the measurement errors and this makes the detection of individual outliers much less certain; if it were not for the value of 300 in this set, 80 would clearly be an outlier and the conclusion that this textile is much more recent than its ascribed historical date would no longer hold (see INTRODUCTION).

Variation between Measurements on the Same Sample

If the quoted errors could be accepted as compatible with one another, it would be reasonable to compare deviations of the individual measurements from the appropriate weighted mean with the corresponding quoted errors. Table 3 shows the result of such a comparison using the test statistic recommended by Ward and Wilson (1978), $X^2 = \Sigma\ (D_i/S_i)^2$ where D_i is the deviation from the weighted mean, S_i the error quoted for the measurement, and n is the number of measurements. The value of X^2 has a chi-squared distribution, with $n-1$ degrees of freedom, when the deviations are consistent with the quoted errors.

TABLE 3
Deviations from the mean compared to quoted errors (outliers excluded)

Sample	X^2	Degrees of freedom	Probability of higher value of X^2
1	9	6	>10% (not significant)
2	11	4	<5% (significant)
3	1	3	>>10% (not significant)

https://doi.org/10.1017/S0033822200007748 Published online by Cambridge University Press

It is concluded from this test that there is no evidence that the measurements of Samples 1 and 3 are significantly more variable than expected on the basis of the quoted errors. Measurements of Sample 2, on the other hand, are more variable than their quoted errors would suggest.

Agreement between Calibrated Date Ranges and Ascribed Dates

The means, measurement errors, calibration uncertainties and calibrated date ranges are shown in Table 4, outliers having been excluded. The means are unweighted. The measurement error is taken to be the observed standard deviation between measurements on the same sample, *ie*, it is *not* based on the quoted errors. It therefore includes all sources of error contributing to the uncalibrated date.

The measurement errors have been divided by $\sqrt{n}$ to give the standard error *on the means*. Calibration uncertainties are those suggested by Clark; they have been combined with the standard errors on the means using the usual sum-of-squares rule to give the total errors; ca 95% confidence limits are obtained for each sample by adding and subtracting twice the total error from the mean date. These limits have been converted to calibrated date ranges.

TABLE 4
Calibrated date ranges (based on all measurements except 3440 for Sample 1, 1550 for Sample 3)

Sample	1	2	3
Mean age (BP)	4298	362	588
Number	7	5	4
Measurement error (standard deviation)	141	178	90
Standard error on mean	53	80	45
Calibration uncertainty	60	50	50
Total uncertainty	80	94	67
Date range	4138–4458 (BP)	174–550 (BP)	454–722 (BP)
Calibrated date range (95% confidence limits)	3255–2827 (cal BC)	1400–1668 (cal AD)	1289–1438 (cal AD)
Ascribed date	3000 (BC)	1200 (AD)	1000–1400 (AD)

SUMMARY

Sample 1

There is one outlier, but the other measurements are in statistical agreement with each other. The 95% confidence interval based on all the acceptable data is ca 500 years long and includes the ascribed date.

Sample 2

The variation between samples is higher than expected on the basis of quoted measurement error. The 95% confidence interval is too late by 300 years, centering on ca AD 1500.

Sample 3

There is one outlier. The remainder agree well with each other and the 95% confidence interval (ca 200 yr long) overlaps the ascribed date range.

CONCLUSIONS

The limited number of laboratories taking part in this intercomparison and the relatively small number of measurements have not allowed a very detailed statistical analysis of the results to be made. In particular, comparison between laboratories was not possible because of the small number of repeat measurements by individual laboratories. Nevertheless, some useful general conclusions can be drawn from the exercise:

1) Overall, there is good agreement between the results obtained and the expected historical dating of the samples, in particular as far as Samples 1 and 3 are concerned.

2) There do not appear to be differences between the AMS and small-counter techniques although, as stated in the RESULTS section above, it was not possible to test this fully (there is no *prima facie* reason for supposing there would be differences, however).

3) A coherent series of results can be obtained when several laboratories undertake separate blindfold measurements of the same sample.

4) As expected, there are no special difficulties in dating textiles by ^{14}C using small sample techniques, as the concordance of the calibrated ^{14}C and historical dates for two textiles separated in time by nearly 4000 years clearly shows.

5) The distribution of the results, containing as it does two outliers, lends added emphasis to the need for the dating of any important relic such as the Shroud of Turin to be shared by several laboratories simultaneously if the results are to have maximum credibility. Possibly also, as a further check, exchange of pretreated samples by these laboratories might be desirable.

ACKNOWLEDGMENTS

The paper was presented for us at the 12th International Radiocarbon Conference by Sheridan Bowman.

REFERENCES

Barnett, V and Lewis, T, 1978, Outliers in statistical data: New York, John Wiley & Sons.

Clark, R M, 1975, A calibration curve for radiocarbon dates: Antiquity, v 49, p 251–266.

Klein, J, Lerman, J C, Damon, P E and Ralph, E K, 1982, Calibration of radiocarbon dates: tables based on the consensus data of the Workshop on Calibrating the Radiocarbon Time Scale: Radiocarbon, v 24, p 103–150.

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, Internatl ^{14}C conf, 10th, Proc: Radiocarbon, v 22, no. 3, p 936–946.

Ward, G K and Wilson, S R, 1978, Procedures for comparing and combining radiocarbon age determinations: a critique: Archaeometry, v 20, p 19–31.

APPENDIX 1: PRETREATMENT OF SAMPLES

Procedures used by the laboratories taking part in the exercise for pretreatment of the three textile samples varied in detail, but consisted broadly of washing in hot dilute acid and alkali, with intermediate and final washing in distilled water. The samples were then dried and either combusted directly or pyrolized to carbon dioxide, followed by 1) purification by standard procedures for gas counting, or 2) reduction by various routes to elemental carbon for preparation of targets for AMS measurements.

APPENDIX 2: INVESTIGATION OF OUTLYING RESULTS

Two definite outliers were detected on purely statistical grounds in our initial analysis, both obtained by the same laboratory. These results were so evidently incompatible with all the others that an explanation for the discrepancy was called for. One possibility examined was that the pretreatment of the samples in question could have affected the results and indeed this proved to be so. These samples had been subdivided and subjected to two different methods of pretreatment, one well tested over a number of years and the other new and untried. The use of this second method was unknown to those making the age measurements. The results obtained when the proven, normal method of pretreatment was used were consistent with those of other laboratories, whilst those obtained from the samples pretreated by the new method were grossly discordant. Further tests by the laboratory concerned and repeat measurements using fresh samples and standard pretreatments (HCl/NaOH/HCl, first pair of results for each sample below; HCl alone, second pair) confirm that this must have been due to contamination introduced by the new method of pretreatment. As stated in the RESULTS section above, the outlying results were not included in the final statistical analysis and can now be discounted altogether. The repeat measurements given below are well within the range of the other laboratories' results and leave our general conclusions unchanged.

1 Egyptian (ca 3000 BC)			3 Peruvian (AD 1000–1400)		
Date	Error	$\delta^{13}C$	Date	Error	$\delta^{13}C$
4150	80	−24.0	540	60	−23.5
4080	80	−24.0	430	60	−23.5

https://doi.org/10.1017/S0033822200007748 Published online by Cambridge University Press