[RADIOCARBON, VOL 32, NO. 3, 1990, P 325-326]

DISCUSSION - SESSIONS II & III

Principal discussants: Mebus Geyh, Hannover, Austin Long, University of Arizona, Wim Mook, University of Groningen, Henry Polach, Australian National University, Doug Harkness, NERC, Barbara Ottaway, University of Bradford, Roberto Gonfiantini, IAEA, Jacques Evin, University of Lyon, Jon Pilcher, Queen's University of Belfast

In the overview of all three stages, participants expressed particular interest in the breakdown into the components of variation. Notably, in the case of gas counting and accelerator labs, Stage 1 did not represent only the counting process as the samples required synthesis to produce the necessary counting medium. The separation into components of variation due to pretreatment, synthesis and counting could then be more correctly ascribed to the Stages. Thus, the interpretation of the sources of variation might differ depending on the lab type.

Further discussion followed on the errors quoted by laboratories and their meaning. Discussants suggested that, for the counting error alone, the concept of error multiplier was valid as an indicator of how well labs operate since counting statistics give a minimum error.

The organizers stated that they had requested labs to quote their errors as "typically provided to a user." The controversy concerning the use of an error multiplier explicitly shows that, regardless of how the total error was calculated, it was still inadequate in most cases. However, the breakdown into individual components of variation might then be invalidated.

Continuing on this theme, the discussion turned to the meaning of error multipliers. The errors being discussed were 'statistical'; the reasons for the variation, however, are more likely to be 'blunders', which are non-random. In this situation, it may be difficult to assign a physical cause to the 'blunders', thus justifying their treatment as random. The representative nature of the subgroup of all labs that participated also provided further justification. The use of a blanket error multiplier was disadvantageous to some labs, but labs could, of course, produce an individual error multiplier.

Two additional factors that might influence the observed variation were: 1) $\delta^{13}C$, and 2) the standards used in primary, secondary and tertiary modes. These could account for some of the outliers. The organizers undertook to investigate these points.

A further question concerned the size of the standard deviations. Discussants suggested that subsequent analysis might group laboratories depending on the typical error quoted. The organizers agreed that this should be considered.

Participants noted that the more modern samples in the study seemed to demonstrate the largest variation. They interpreted this as indicative of problems with the modern standards used and lent support to the view that the calibration of in-house standards should be checked.

Disagreement arose over laboratory anonymity. The organizers stated that each laboratory was at liberty to publish its own results and to identify itself. In addition, however, the question of anonymity of the group arose as a general issue. If all laboratories identified themselves, then individual labs would benefit from the contact in helping solve mutual difficulties. This issue was deferred for further discussion.

Discussants felt, in light of the results of the study, improvements in technique were necessary. They also made clear that, although no improvement was apparent from Stage 1 to Stage 3, discrimination against labs because of bad results should be avoided. All agreed to the need for quality assurance. How best to achieve this for the benefit of all had to be a primary consideration. The participants endorsed this view wholeheartedly and added that, from a user's perspective, the *means* of an improvement is less important than the end result: improvement is

imperative. Ultimately, users want to be able to compare, without bias, results from different laboratories; hence, the request for quality control.

The general opinion was that the two Glasgow studies were significant in showing the effects of problems occurring in a complex technique. Each lab has experienced fluctuations in quality which, in the past, might have been addressed by exchanging information to identify and solve problems. The Glasgow studies provided a formal framework within which such collaboration could take place. The relationship between user and laboratory is equally important, and should be honest and collaborative. Problems arise not only with the dating technique, but also with the interpretation of results, usually with archaeological samples. In these cases, both the dater and the submitter should work together to resolve mutual problems.

Discussants pointed out that results of the study showed that seven labs met all the performance requirements and hoped that these labs would continue to do so over the next ten years. However, for labs that did not meet the approved quality criteria, it is important not only to improve, but also to maintain that improvement. If all ^{14}C labs could achieve this status, this would epitomize the true meaning of quality assurance.

The study results showed that the error quoted should be a long-term error. If standards are measured regularly, then the same variations should be apparent, and any errors quoted should be based on the long-term variation. However, discussants noted that the same variation was not apparent in standards, partly because pre-processing removes outliers. Thus, all results should be included in the calculation of the error.

Again, the question of laboratory identification arose with a plea for identification of a group of labs that made very accurate and precise measurements, so that these labs could act as consultants. The organizers agreed that this mode of operation could be arranged and that future work would stress openness and cooperation amongst labs.

Some commented that the study has been detrimental to the overall reputation of the ^{14}C community and how users perceive it. Labs that have had problems and have solved them will still be perceived as being in difficulty, which argues for intercomparisons on a regular basis, to allow laboratories to demonstrate their improvements. However, users should view participation in such intercomparisons as a positive step. Intercomparison results enable users to gain information on laboratory performance before submitting samples.

The organizers stressed that the purpose of the study was to effect accountability, improvement and reduction of errors. These certainly benefit both the individual laboratory and the users. Most participating labs specifically requested anonymity. Accordingly, the organizers were bound by the consensus view. We should also remember that any participating lab that did not reveal its code was in the same position, with respect to the user, as a non-participating lab.

Regarding practical applications of the study results, particularly in cases of comparing pre-existing dates from different labs, one suggestion is to increase the error term by some multiplicative factor. The difficulty with this approach is that, if an error is present, we do not know how it affects the results. An alternative possibility was to retain the statistical error on the counting statistics and evaluate individual multiplicative error factors, if desired.

To conclude, participants agreed to retain the statistical error, that labs should make available a clear outline of their procedures, and that archaeologists and other users must educate themselves in handling and interpreting ^{14}C dating errors.