THE FIFTH INTERNATIONAL RADIOCARBON INTERCOMPARISON (VIRI): AN ASSESSMENT OF LABORATORY PERFORMANCE IN STAGE 3

E Marian Scott[1] • Gordon T Cook[2] • Philip Naysmith[2]

ABSTRACT. Proficiency testing is a widely used, international procedure common within the analytical chemistry community. A proficiency trial (which VIRI is) often follows a standard protocol, including analysis that is typically based on *z*-scores, with one key quantity, σ_p. From a laboratory intercomparison (sometimes called a proficiency trial), we hope to gain an assessment of accuracy (in this case, from dendro-dated samples), laboratory precision (from any duplicate samples), and generally, an overall measure of performance, including measurement variability and hence realistic estimates of uncertainty. In addition, given our stated aim of creating an archive of reference materials, we also gain a determination of consensus values for new reference materials.

VIRI samples have been chosen to deliver these objectives and the sample ages included in the different stages, by design, spanned modern to background. With regard to pretreatment, some samples required intensive pretreatment (e.g. bone), while others required none (e.g. cellulose and humic acid). Sample size was not optimized, and indeed some samples were provided solely for accelerator mass spectrometry (AMS) measurement. In this sense, VIRI presented a more challenging exercise than previous intercomparisons, since by its design in stages, one can explore improvements (or deteriorations) over time in laboratory performance. At each stage, more than 50 laboratories have participated, with an increasing demographic shift towards more AMS and fewer radiometric laboratories.

INTRODUCTION

The Fifth International Radiocarbon Intercomparison (VIRI) continued the tradition of the TIRI (third) and FIRI (fourth) intercomparisons providing an independent check on laboratory procedures. VIRI was designed to have 3 stages, spread over several years, involving 2 sets of specific sample types (grain and bone), and then a final stage involving a wide variety of common sample materials. Separate papers have summarized the results for first 2 stages (Scott et al. 2007, 2010).

In Stage 3 of VIRI, 7 samples were provided to all laboratories, comprising 3 wood samples, in addition to cellulose, shell, barley mash, and humic acid samples (samples K, L, M, O, R, S, U). Radiometric laboratories received a further charcoal sample (Sample P), while a further 4 samples were provided for AMS laboratories, comprising 1 further wood sample, 1 charcoal, and 2 humic acid samples (samples N, Q, T, J). More than 50 laboratories (of which 19 were radiometric) participated. In addition, many AMS facilities reported replicate results.

Analysis of the results reported here focuses on the definition of the consensus values as well as exploring overall performance using *z*-scores.

SAMPLE DESCRIPTIONS FOR VIRI, STAGE 3

In summary, 12 samples were distributed (1 for radiometric only, 4 for AMS only). Materials included known-age wood, cellulose, shell, barley mash, humic acid, and charcoal. Samples ranged in age from modern, to a few thousand years, to more than 40,000 yr. A brief description of each group of materials is given below.

Humic Acid: Samples J (AMS only), T (AMS only), and U

Sample T: humic acid (<1 half life) from a peat deposit in Scotland.

[1]Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland. Corresponding author.
Email: marian@stats.gla.ac.uk.
[2]SUERC, Scottish Enterprise Technology Park, East Kilbride G75 0QF, Scotland.

Proceedings of the 20th International Radiocarbon Conference, edited by A J T Jull
RADIOCARBON, Vol 52, Nr 2–3, 2010, p 859–865

Sample J: humic acid from a peat deposit in Siberia, provided by Prof Kh Arslanov, St Petersburg. This sample is close to background.

Sample U: humic acid from a peat deposit at St Bees, Cumbria, (~2 half-lives) already used as FIRI E.

Wood: Samples K, L, M (AMS only), N, and O (Cellulose)

Sample K: wood (anticipated to be background), provided by Michael Friedrich, Hohenheim.

Sample L: wood (known age) provided by Mike Baillie, Belfast. The sample is identified as Corlea, Q5994.

Samples M, N: wood (<1 half-life), provided by Gordon Cook, SUERC. Oak (alder) samples from Loch Tay (a crannog site).

Sample O: cellulose (known age) from Cambridge, corresponding to 60 rings from a plateau period; previously used.

Charcoal: Samples P and Q (AMS only)

Sample Q: charcoal from Iceland, provided by Dr Mike Church, University of Durham.

Sample P: charcoal from Mexico, provided by Dr L Manzanilla from the Teotihuacán archaeological site.

Barley Mash

Sample S: barley mash from Glengoyne distillery, 2001.

Shell

Sample R: murex shell from the Tel Dor archaeological site, provided by Elisabetta Boaretto, Weizmann Institute.

METHODS

Consensus Values

Following the procedure outlined in Scott (2003), preliminary consensus values were calculated using the median, which also leads naturally to the identification of a number of outlying values, which are then screened out. The final consensus or assigned values are calculated using a weighted average, where the weights are defined by laboratory quoted uncertainty. The exception to this is for samples whose results are censored (i.e quoted as greater than), such as Sample K, where an alternative non-parametric procedure is used.

PROFICIENCY TESTING AND *Z*-SCORES

Proficiency testing is widely used in the analytical chemistry communities. VIRI and its predecessors are examples of proficiency tests and, as is common in the analytical chemistry community, they follow standard protocols. Analysis of the results also followed fairly standard procedures, evaluation of the assigned value (e.g. ^{14}C age) and measures of performance, typically based on *z*-scores derived using one key quantity, σ_p. Interpretation of *z*-scores includes accuracy, precision, and "fitness for purpose."

For the analysis, we have reported *z*-scores, calculated as

$$z = (X_M - X_A) / \sigma_p$$

https://doi.org/10.1017/S003382220004594X Published online by Cambridge University Press

where X_M is the reported result, X_A is the assigned or true value for the material, and σ_p is the target value for the standard deviation for values of X. The value for σ_p is determined by fitness for purpose and represents the amount of uncertainty in the results that is tolerable in relation to the purpose of the analysis, although more commonly the laboratory quoted error is used. X_A may be known or assessed as the consensus value. Interpretation of the z-score reflects the accuracy achieved and provides a means of making a judgement concerning fitness for purpose.

It is commonly assumed that z should be Normally distributed with zero mean and variance 1, where

- A z-score of 0 implies a *perfect* result.
- A z-score between –2 and +2 is generally considered as complying with fitness for purpose.
- A z-score outwith –3 or +3 would be very unusual, with further investigation needed.

RESULTS

Consensus Values

Some 52 laboratories reported results (32 AMS, 2 GPC [gas proportional counting]) in Stage 3, but there are many more sets of results (>60 sets) due to the multiple reporting particularly by AMS facilities. Table 1 presents the summary statistics for the samples, showing the mean, median, and standard deviation as well as the interquartile range (broken down by laboratory type: AMS or radiometric (R)) in pMC. (Note: due to the small number of GPC laboratories, it is not possible for statistical summaries to use the usual convention of AMS, LSC, and GPC.) Figure 1a and b show the mean and standard deviation (in pMC) for all samples.

Table 1 Summary statistics for each sample by lab type.[a]

Sample	Lab type	*n*	Mean	Median	Std dev	Q_1	Q_3
J	AMS	40	0.521	0.477	0.160	0.43	0.555
	R	—	—	—	—	—	—
K	AMS	36	0.097	0.045	0.246	0	0.129
	R	10	0.644	0.155	1.082	0.068	0.903
L	AMS	35	75.774	75.800	0.757	75.480	75.960
	R	14	75.773	75.840	1.767	75.040	77.265
M	AMS	37	73.938	73.840	0.358	73.695	74.115
	R	14	73.396	73.650	1.792	72.985	74.142
N	AMS	38	73.874	73.829	0.470	73.580	74.280
	R	—	—	—	—	—	—
O	AMS	49	98.355	98.490	0.771	98.225	98.695
	R	14	97.864	98.400	2.738	96.955	99.262
P	AMS	4	78.68	80.05	4.70	73.85	82.15
	R	15	80.457	80.520	1.609	79.770	81.78
Q	AMS	32	92.426	92.502	0.512	92.032	92.685
	R	—	—	—	—	—	—
R	AMS	37	73.321	73.25	0.767	73.015	73.520
	R	13	73.652	73.600	2.348	73.170	74.705
S	AMS	51	109.61	109.91	2.46	109.61	110.33
	R	15	105.27	108.59	10.63	105.60	110.20
T	AMS	36	65.871	65.865	0.345	65.703	66.045
	R	—	—	—	—	—	—
U	AMS	45	23.090	23.090	0.219	22.915	23.170
	R	14	23.276	23.335	0.869	22.990	23.500

[a]AMS = accelerator mass spectrometry; R = radiometric.

https://doi.org/10.1017/S003382220004594X Published online by Cambridge University Press

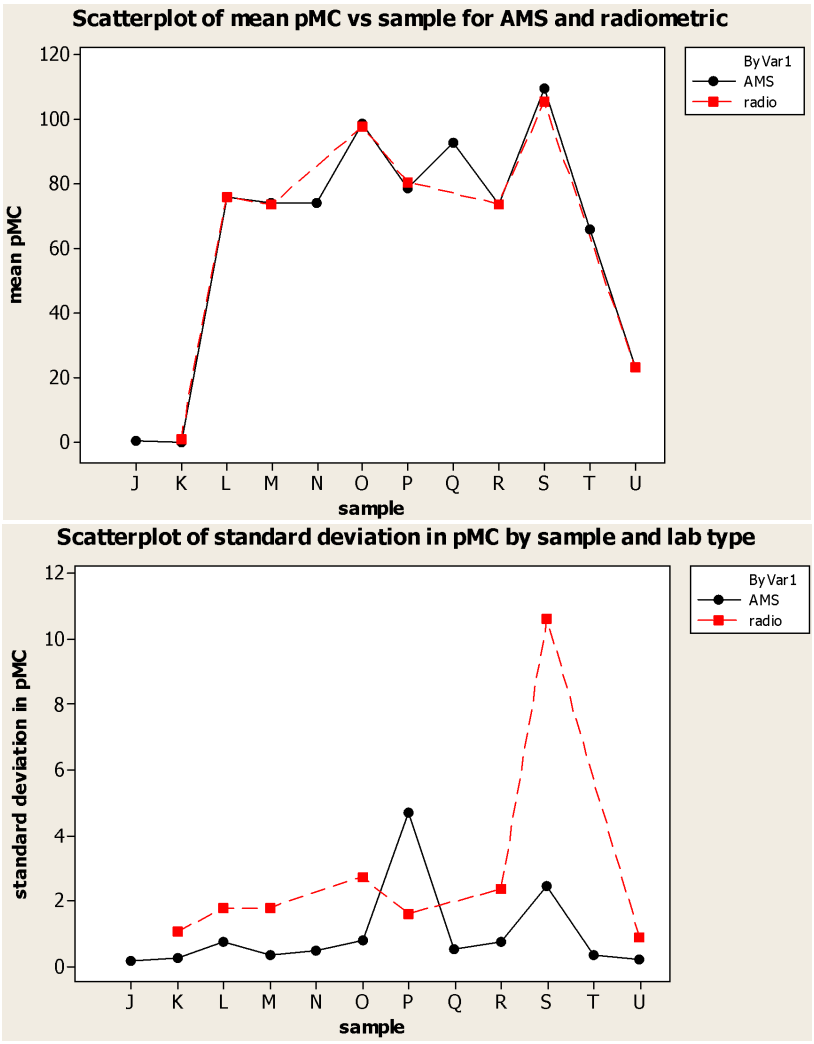

Figure 1 a) Mean pMC for each sample by laboratory type; b) standard deviation for each sample by laboratory type.

Table 2 gives the consensus values (in pMC) for all the samples in Stage 3.

Table 2 Consensus values for VIRI Stage 3.

Sample	Consensus value (pMC)	1 σ	Sample	Consensus value (pMC)	1 σ
J	0.4603	0.008	P	80.457	0.0862
K	0.0576	0.0062	Q	92.383	0.0512
L	75.719	0.0395	R	73.338	0.0368
M	73.900	0.0322	S	109.96	0.0417
N	73.839	0.0392	T	65.821	0.0333
O	98.457	0.0385	U	23.079	0.0155

https://doi.org/10.1017/S003382220004594X Published online by Cambridge University Press

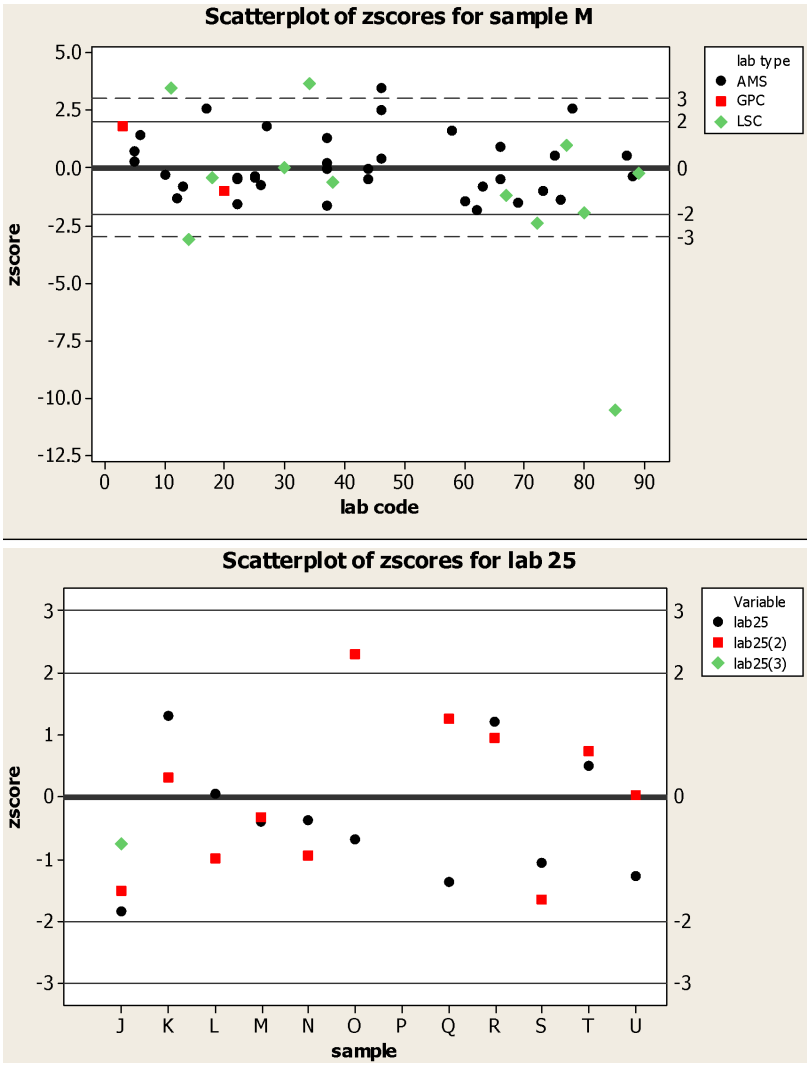

Figure 2 a) *z*-scores for Sample M; b) *z*-scores for an individual laboratory.

Discussion

Figure 1a shows little evidence of difference on average between radiometric and AMS facilities, which is supported by findings in the earlier stages of VIRI. Figure 1b does show some evidence of differences over the suite of samples in terms of variability (note: Sample P should be discounted since although predominantly a radiometric sample, a few AMS results were also returned). Again, this is consistent with previous findings in terms of variability, specifically for sample P and sample S. Any outliers have not been discounted for this preliminary graphical representation, and they would have a strong effect on the quoted standard deviation.

https://doi.org/10.1017/S003382220004594X Published online by Cambridge University Press

z-scores

Figures 2a and b show illustrative *z*-score plots for an individual sample and for an individual laboratory, while Figure 3 shows the boxplots of all *z*-scores for each sample. In these calculations, the laboratory's 1-σ quoted error has been used, so that we would expect that most values should lie within the ±2 band, indicating acceptable performance. Values beyond ±2 are observed, which may be a function of a large difference between measurement and consensus value, and/or a small error. Use of such plots allows a) identification whether a specific material has proved problematic (e.g. perhaps being inhomogeneous) and b) for an individual laboratory, whether any specific sample has an extreme *z*-score, indicating an unusual result.

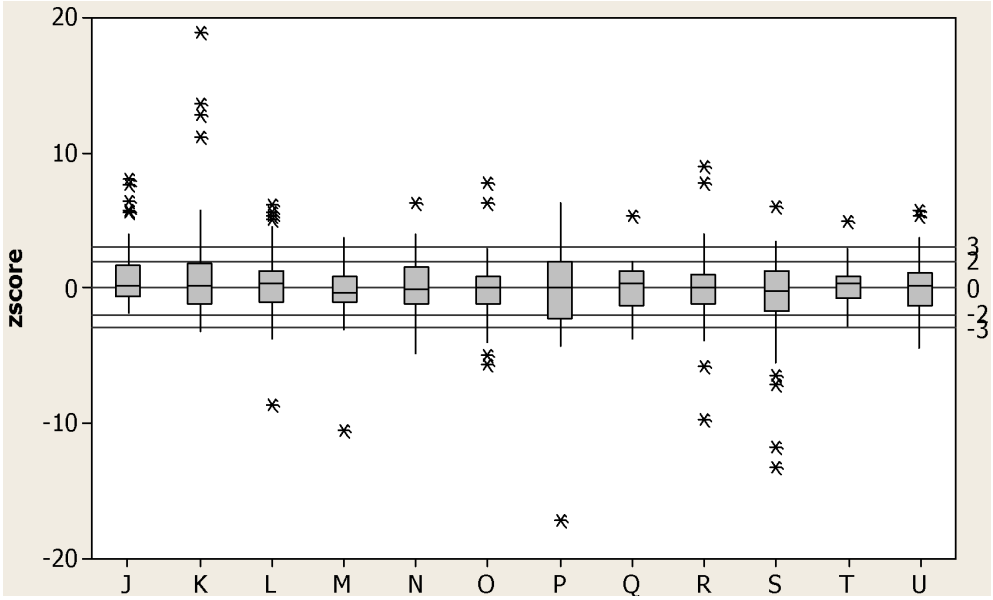

Figure 3 Boxplots of *z*-scores for all samples

Figure 3 shows how the distribution of *z*-scores varies over the samples. Individual *z*-scores can be seen that lie well beyond the ±3 range, and these are particularly apparent for samples K (background) and S (modern). In such a large data set, we would expect to find some results that lie outwith the acceptable bounds of ±2; however, extreme values should be the subject of further investigation. The results are, however, broadly consistent over the pMC range of 0.05 to 110 pMC, with more variation at the endpoints of this range, in that the majority of results for each sample lie within ±2 and ±3.

CONCLUSIONS

Participation in a proficiency test helps ensure the results from a laboratory are meaningful, contributes to and enhances a laboratory's reputation. What do participating laboratories want? They expect relevant test material (samples), confidence in homogeneity of test material, and confidence in the assigned value. The VIRI program has striven to deliver this using wherever possible sample material that is in routine use. Stage 3 covered a wide range of sample materials and spanned the ^{14}C activity range.

https://doi.org/10.1017/S003382220004594X Published online by Cambridge University Press

Good, practical quality assurance (QA) procedures are in place in laboratories, but participation in proficiency trials is a necessary part of routine activity. Their value depends on the quality of the test materials and on the willingness of so many labs to participate. Within the VIRI program, overall more than 60 laboratories worldwide have participated, with now many more AMS than radiometric facilities participating.

The results have shown that for "old" samples, we see more between-laboratory variability, but there is little evidence of differences on average between the laboratory types. For this round, individual laboratory performance has been assessed using *z*-scores, and the majority of results have fallen in the ±2 satisfactory (or "fit for purpose") band. Some unusual values are also observed that would merit further investigation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the many sample providers: Kh A Arslanov, A Bayliss, M Baillie, E Boaretto, M Church, N Dixon, M Friedrich, S Gulliksen, T Higham, P Reimer, G Zaitseva, L Beramendi-Orosco; all participating laboratories, English Heritage, and Historic Scotland.

REFERENCES

Scott EM 2003. The Third International Radiocarbon Intercomparison (TIRI) and the Fourth International Radiocarbon Intercomparison (FIRI). *Radiocarbon* 45(2):135–408.

Scott EM, Cook GT, Naysmith P, Bryant C, O'Donnell D 2007. A report on Phase 1 of the 5th International Radiocarbon Intercomparison (VIRI). *Radiocarbon* 49(2):409–26.

Scott EM, Cook GT, Naysmith P. 2010. A report on phase 2 of the Fifth International Radiocarbon Intercomparison (VIRI). *Radiocarbon* 52(2–3):846–58.