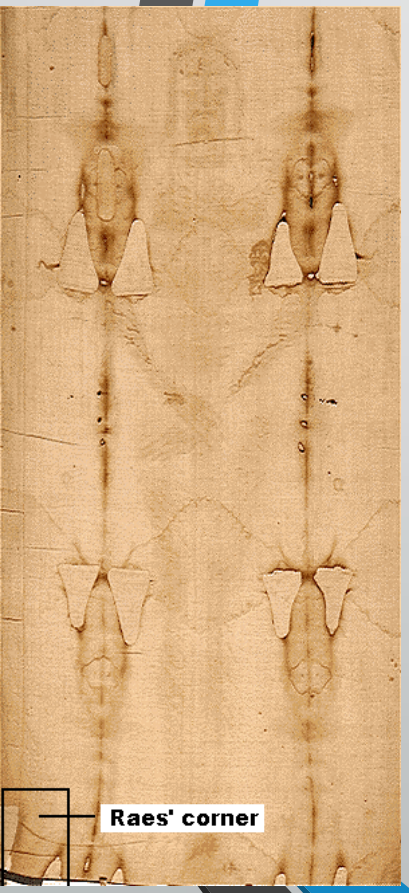
Tristan Casabianca

Radiocarbon Dating of the Turin Shroud: Lessons from Failure

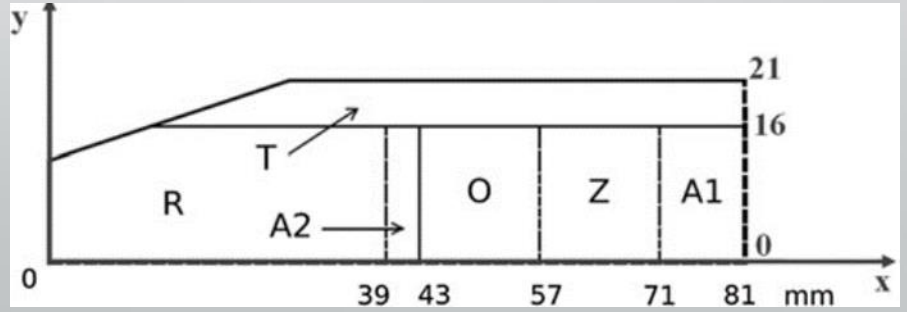
Ancaster

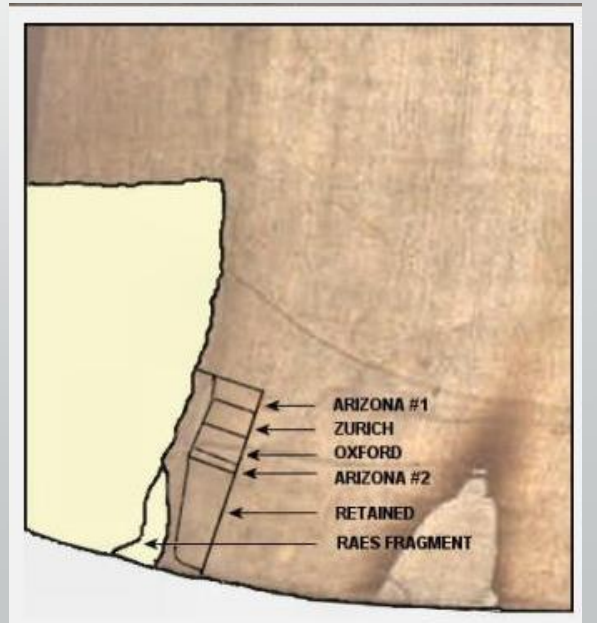
15 August 2019

Radiocarbon Dating of the Turin Shroud



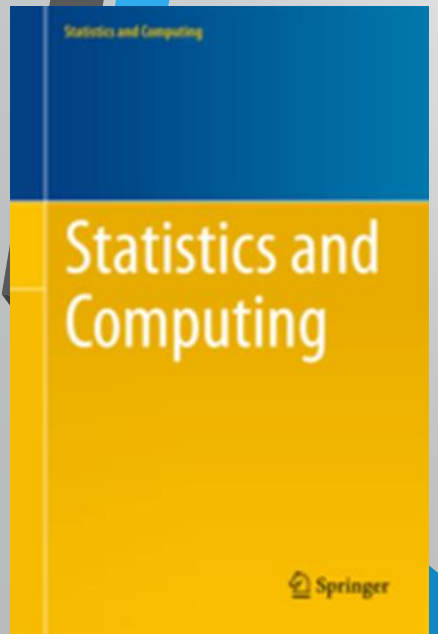
- One sample taken from one corner: problem of representativity of the sample?
- Sample cut into pieces and given to 3 laboratories: Oxford, Zürich, Arizona (Tucson)
- One method (Accelerator Mass Spectrometry)
- 3 control samples

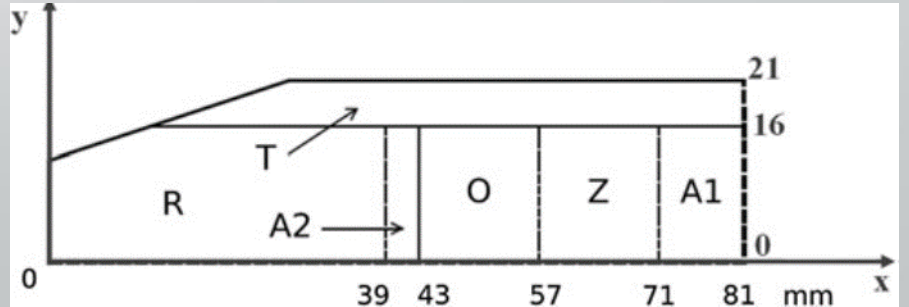




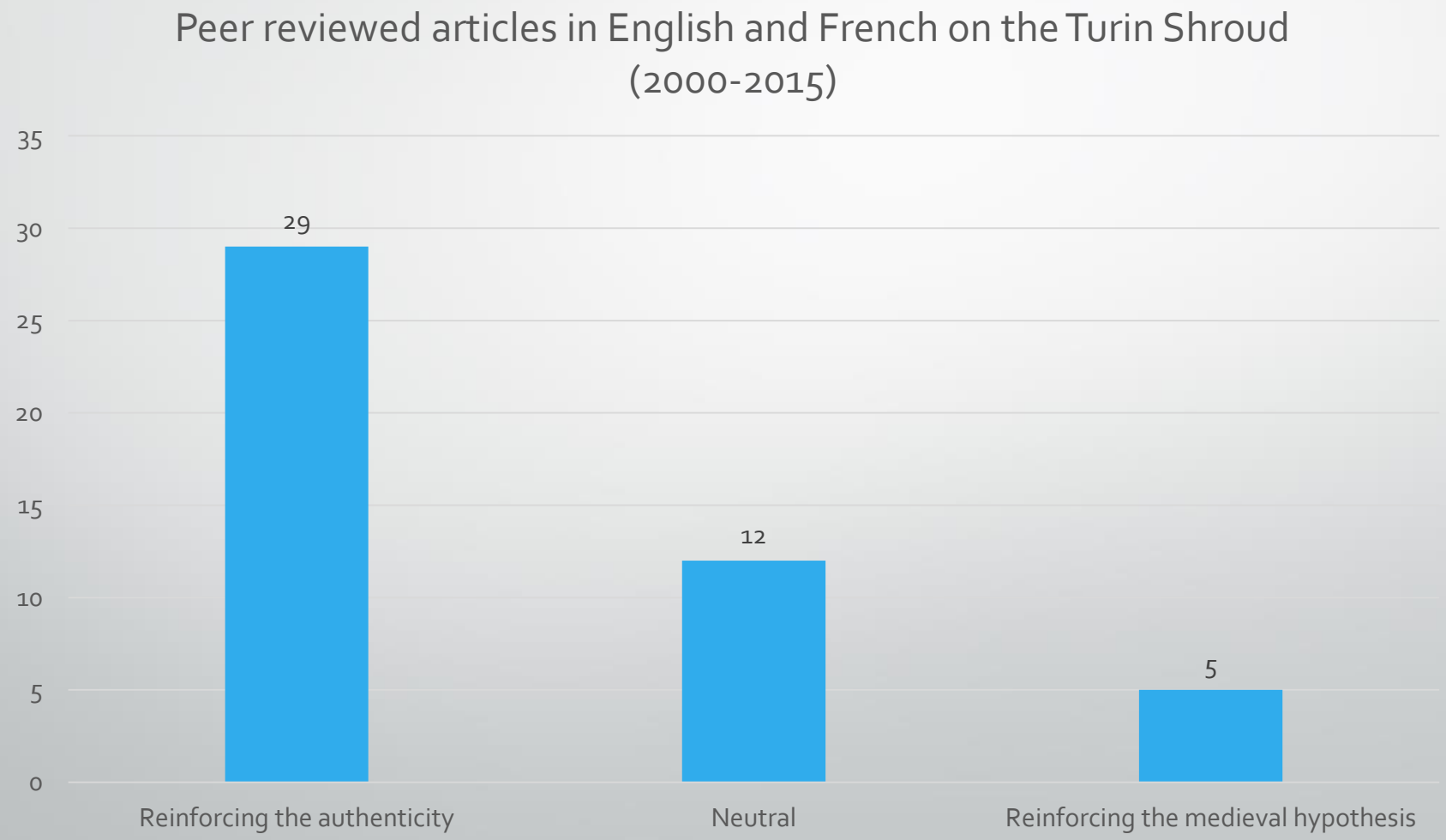
From 1260-1390! to 1260-1390?

- "at least 95% confidence for the linen of the Shroud of Turin of AD 1260–1390 [...] These results therefore provide conclusive evidence that the linen of the Shroud of Turin is mediaeval." (Damon et al., *Nature*, 1989, 614)
- From 2005, a growing contestation:
  - Is the sample representative of the whole cloth? (Rogers, *Thermochimica Acta*, 2005)
  - Is the dating statistically valid? (Riani et al., *Statistics and Computing*, 2013)
    - Regression analysis shows a significant statistical trend





A new trend


Peer reviewed articles in English and French on the Turin Shroud (2000-2015)
35
30
25
20
15
10
5
0
29
12
5
Reinforcing the authenticity
Neutral
Reinforcing the medieval hypothesis

More and more contested

- Contestation of Damon et al. acted even in *Nature* journals
  - "Results of radiocarbon measurements from distinct and independent laboratories yielded a calendar age range of 1260–1390 AD, with 95% confidence, thus providing robust evidence for a Medieval recent origin of TS. However, **two papers** [Rogers and Riani] **have highlighted some concerns about this determination and a Medieval age does not appear to be compatible with the production technology of the linen nor with the chemistry of fibers**…" (Barcaccia et al., "Uncovering the sources of DNA found on the Turin Shroud", *Scientific Reports*, 2015)

nature > scientific reports

SCIENTIFIC REPORTS

Raw data

- The statistic analysis has frequently been put into question.
  - As soon as 1988, multiple scholars repeatedly asked for the raw data.
  - However the three laboratories and the centralising institution, the British Museum, never answered favorably
- In 2017, in a Freedom of Information Act request, the British Museum released its documentation, more than 700 pages not ordered or classified
  - Among them, the data sent by the laboratories to the British Museum for its statistical analysis
- Raw data are critical in the understanding of this radiocarbon dating process

archaeometry

Archaeometry **, ** (2019) **–**

doi: 10.1111/arcm.12467

RADIOCARBON DATING OF THE TURIN SHROUD: NEW EVIDENCE FROM RAW DATA*

T. CASABIANCA†

Ajaccio 20000, France

E. MARINELLI

Collegamento pro Sindone, Rome, Italy

G. PERNAGALLO

Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania CT, Italy

and B. TORRISI

Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania CT, Italy

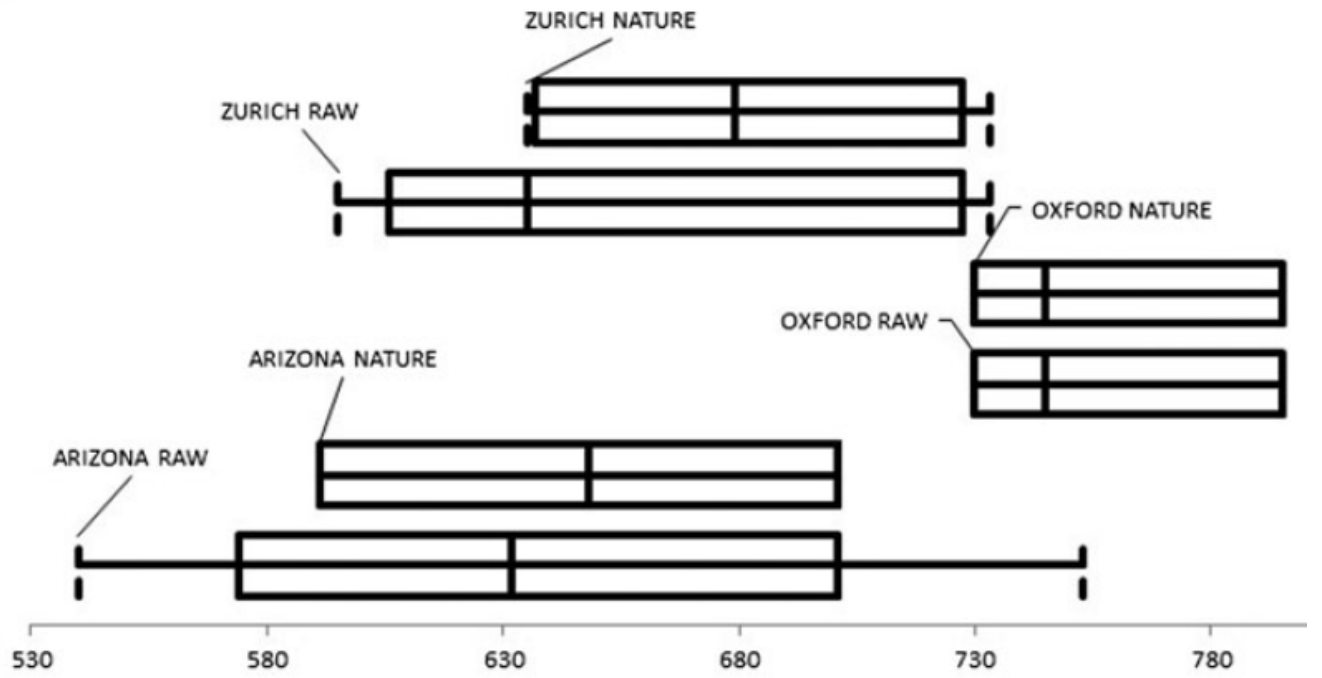
Turin Shroud Raw Radiocarbon Dates

- **Arizona** made 40 measurements (5 x 8) for **8 radiocarbon dates** (vs. 4 in Nature)
  - Two uncertainties corrected (difference between Raw 1 and Raw 2)
- **Oxford** performed 5 measurements, a mean was given for 2 radiocarbon dates, the errors are corrected (counting error contribution)
- **Zürich** performed 4 x 10 measurements for each sub sample, sent a revised report due to a current dependent effect for its two last dates

Arizona Raw 1	Arizona Raw 2	Arizona Nature	Oxford Raw	Oxford Nature	Zürich Raw	Zürich Nature
606 ± 41	606 ± 41	591 ± 30	795 ± 53	795 ± 65	733 ± 61	733 ± 61
574 ± 45	574 ± 45		730 ± 30	730 ± 45	722 ± 56	722 ± 56
753 ± 51	753 ± 51	690 ± 35	745 ± 46	745 ± 55	635 ± 57	635 ± 57
632± 49	632± 49				617 ± 47	639 ± 45
676 ± 40	676 ± 59	606 ± 41			595 ± 46	679 ± 51
540 ± 37	540 ± 57					
701 ± 47	701 ± 47	701 ± 33				
701 ± 47	701± 47					

The *Nature* radiocarbon dates

- Classical tests are concordant and show heterogeneity
  - ANOVA. P-Value = 4,0% (significant for Arizona-Oxford, but also for Oxford-Zürich)
  - Ward and Wilson test (significant: 8,60 > 5,99)
  - OxCal: Poor overall agreement (significant: 41,8% < 60%)
- Trend in the results
  - The closer to to center of the cloth, the younger are the radiocarbon dates.
- Lack of validity of the results



Raw data: focus on Arizona

- 2 series of 8 measurements (Arizona Raw 1/ Raw 2)
  - Unusual change in two errors between 'Arizona Raw 1' and 'Arizona Raw 2'
  - Two radiocarbon dates of 'Arizona Raw 1' should not have been merged.
- Justification: made the same day with same standards. But the counts of detected C14 atoms for the 4 groups also show strong heterogeneity (p-value < 0,0001).

Arizona Raw 1	*Arizona Raw 2*	*Arizona Nature*
606 ± 41	606 ± 41	
574 ± 45	574 ± 45	591 ± 30
753 ± 51	753 ± 51	
632 ± 49	632 ± 49	690 ± 35
676 ± 40	676 ± 59	
540 ± 37	540 ± 57	606 ± 41
701 ± 47	701 ± 47	
701 ± 47	701 ± 47	701 ± 33

	Ward and Wilson test (critical value in brackets)	*OxCal 4.3 overall agreement index (number of individual dates below 60% in brackets)*
Arizona *Nature* vs. Oxford	8.60	41.8%
Nature vs. Zürich *Nature*	(5.99 for 3–1 df)	(3/12)
Arizona Raw 1 vs. Oxford	10.75	18.1%
Nature vs. Zürich *Nature*	(5.99 for 3–1 df)	(6/16)
Arizona Raw 2 vs. Oxford	8.55	28.4%
Nature vs. Zürich *Nature*	(5.99 for 3–1 df)	(5/16)
Arizona Raw 1	19.24	21.4%
	(14.07 for 8–1 df)	(2/8)
Arizona Raw 2	14.45	34.6%
	(14.07 for 8–1 df)	(2/8)

An Unreliable Radiocarbon Dating

- Statistical results are supported by the amount of foreign material.
- No ‘conclusive evidence’ that the 1260-1390 AD interval is reliable or representative of the whole cloth.
- How was this failure possible?

The radiocarbon dating of the Turin Shroud in the reproducibility crisis

- An explanation of the 'Carbon-dating fiasco' would be a wonderful topic for historians of science and sociologists (Thomas de Wesselow, art historian)
- Radiocarbon dating of the Turin Shroud offers a practical example of the ongoing reproducibility crisis in science
  - "That, in the light of concerns about the 'reproducibility crisis' (the difficulty of replicating a large number of published claims), is no surprise – but it's troubling nonetheless." (Philip Ball, 2019)

Some facets of the reproducibility crisis

- Reluctance to release the data:
  - "the only case I know of authors of an article refusing to provide data that would allow other scientists to repeat the calculation and verify whether it was done correctly" (Paolo di Lazzaro, 2018)
- Confirmation bias
- Pressure to publish
- Data dredging:
  - Different methods between control samples and TS sample (p-hacking?)
  - 5% significance for TS should have been a red flag
- Peer review failure
  - Not specific to the Turin Shroud.
    - A medical journal put 8 errors in an article. Received 221 reviews.
      - Median number of errors detected 2.
      - 16% of the reviewers failed to detect any (Godlee et al., JAMA, 1998)
    - "journal editors should not assume that their reviewers will detect most major flaws in manuscripts [...] improvements after training were minor despite using the types of papers easiest to review for errors" (Schroter et al., JRSM, 2008)

Table 3 Comparison of Interlaboratory Scatter with Quoted Errors

	Sample 1	2	3	4
Observed standard error on weighted mean (scatter)	29	4	17	22
Combined quoted error on mean	16	16	20	20
Chi-squared value (2 d.f.)	**6.4**	**<1**	**1.3**	**2.4**
Significance[1]	**<5%**	**>25%**	**>25%**	**>25%**

Notes

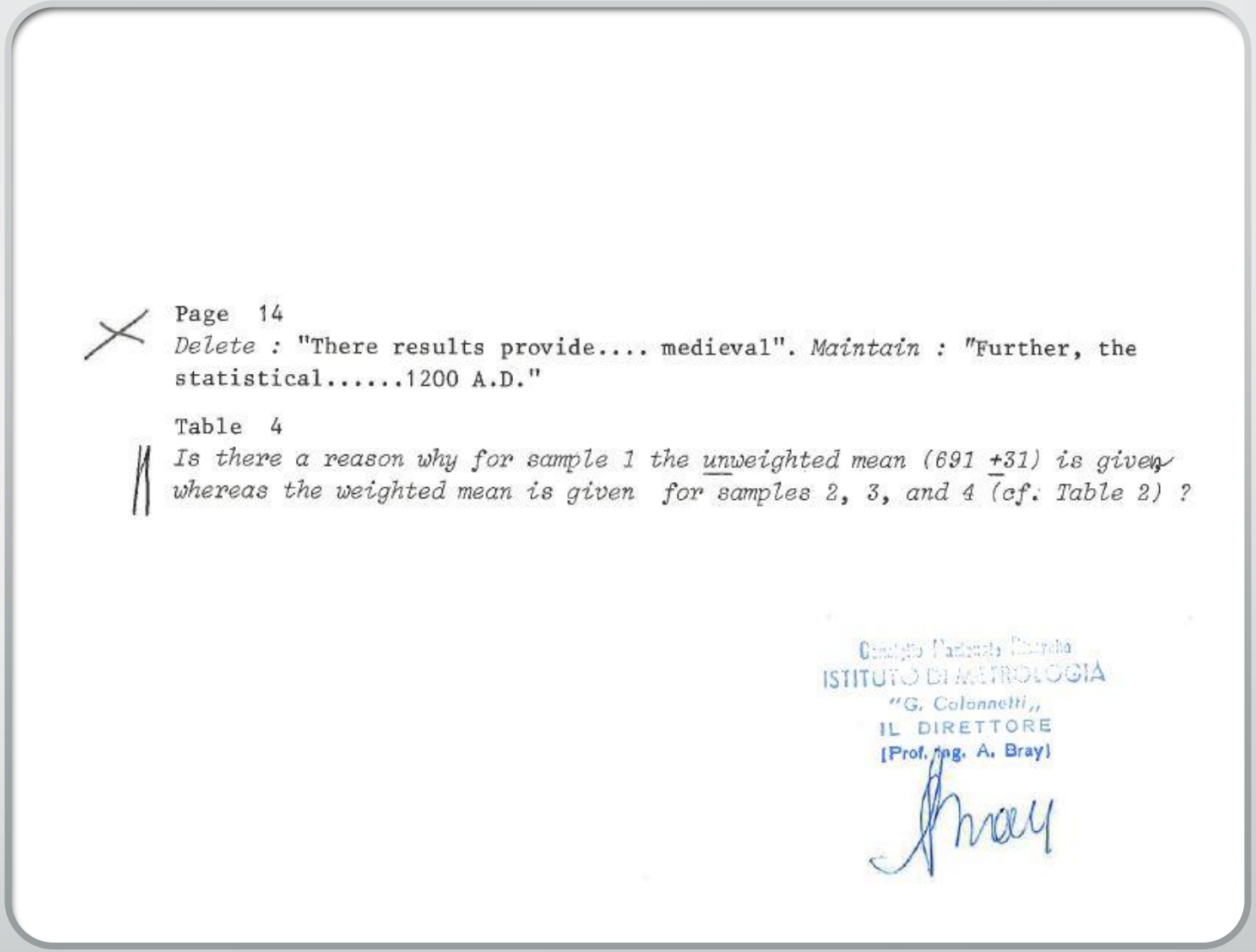
(1) Assuming that the quoted errors are a true reflection of all sources of random variation, the significance level is the probability of obtaining, by chance, a scatter among the three dates as high as that observed.

Table 2 Summary of mean radiocarbon dates and assessment of interlaboratory scatter

Sample	1	2	3	4
Arizona	646 ± 31	927 ± 32	1,995 ± 46	722 ± 43
Oxford	750 ± 30	940 ± 30	1,980 ± 35	755 ± 30
Zurich	676 ± 24	941 ± 23	1,940 ± 30	685 ± 34
Unweighted mean*	691 ± 31	936 ± 5	1,972 ± 16	721 ± 20
Weighted mean†	689 ± 16	937 ± 16	1,964 ± 20	724 ± 20
χ^2 value (2 d.f.)	6.4	0.1	1.3	2.4
Significance‡ level (%)	5	90	50	30

Internal peer review

- Luigi Gonella
  - Gonella to Tite: provide some informations requested by Tite (14 November 1988)
- Anthos Bray
  - Focus on the final results
  - Relies on the results provided
  - Wanted to delete the sentence: « These results provide conclusive evidence that the linen of the Turin Shroud is mediaeval »
  - Insists on the 68% probability that the shroud lies within the range 1270-1290 AD

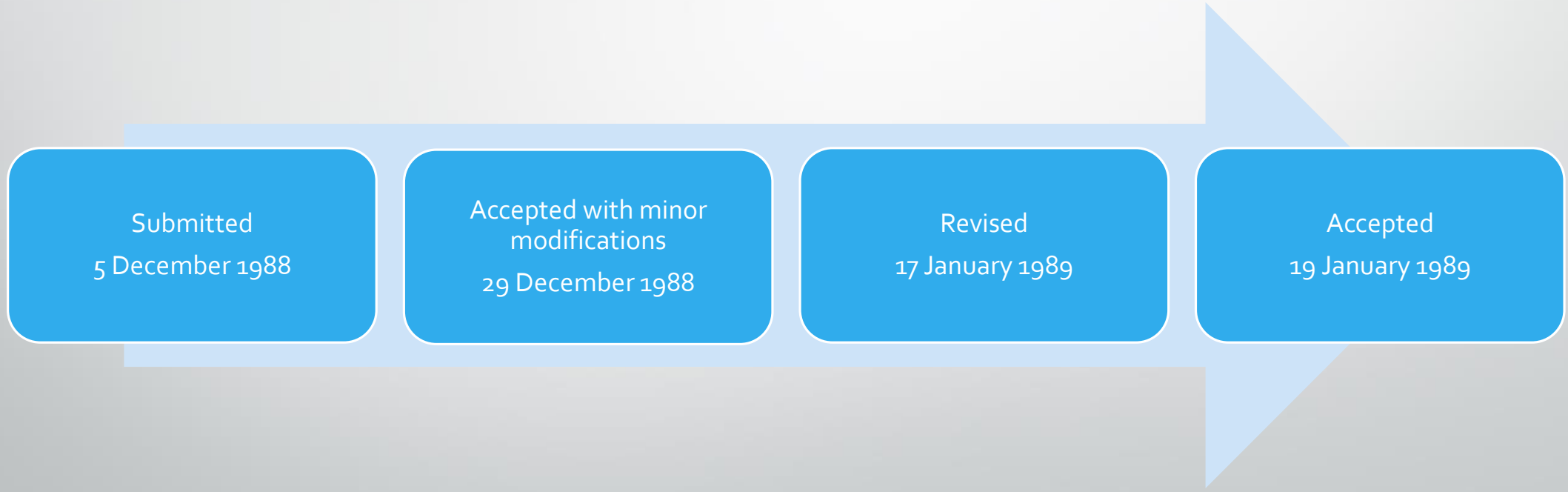


Page 14
Delete : "There results provide.... medieval". *Maintain :* "Further, the statistical.....1200 A.D."

Table 4
Is there a reason why for sample 1 the unweighted mean (691 ±31) is given whereas the weighted mean is given for samples 2, 3, and 4 (cf. Table 2) ?

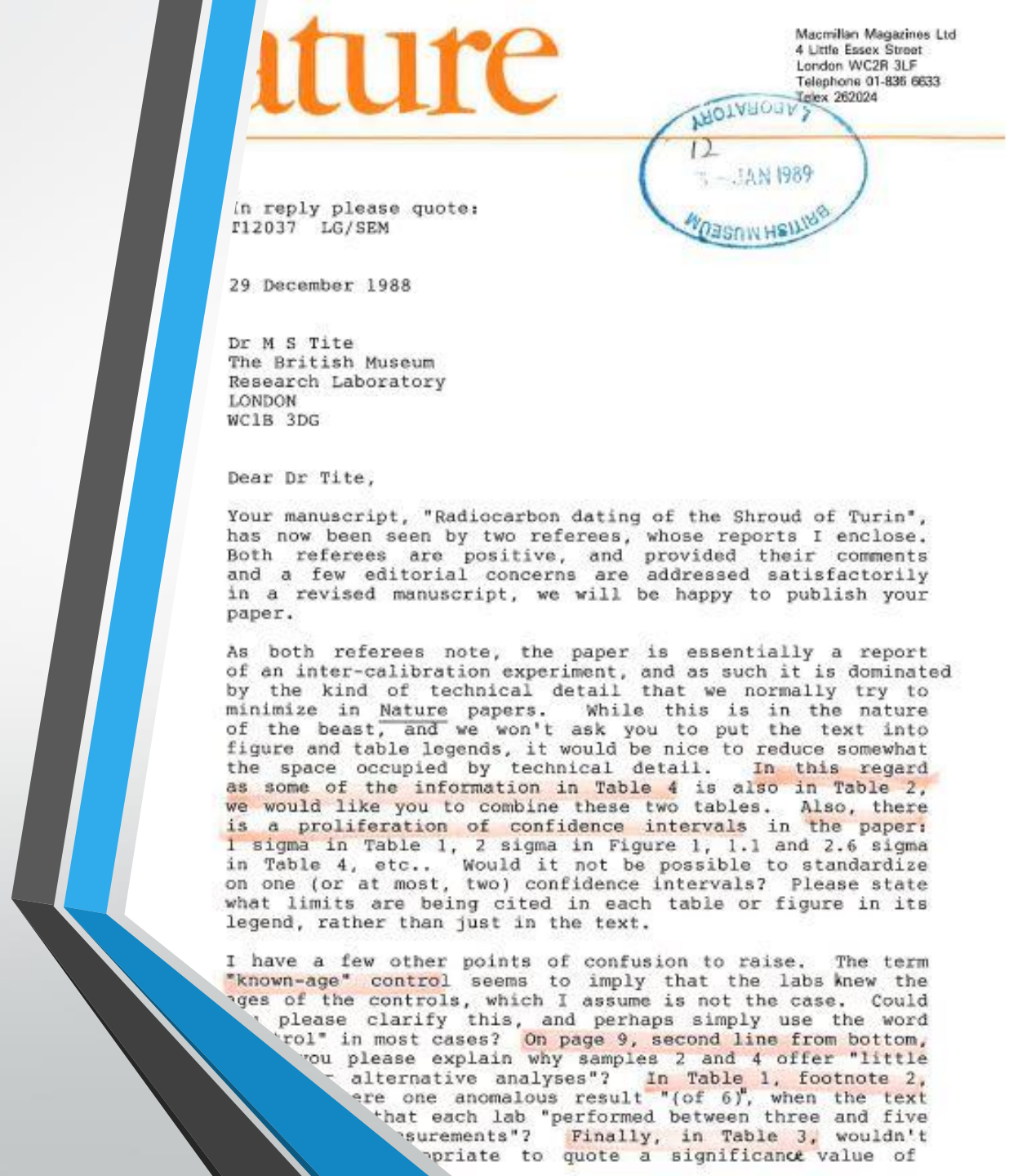
Consiglio Nazionale delle Ricerche
ISTITUTO DI METROLOGIA
"G. Colonnetti"
IL DIRETTORE
(Prof. Ing. A. Bray)

Peer review process of Damon et al. to *Nature*



Nature Review

- 2 referees and one editor (2 pages letter)
- Positive but not detailed reviews from the referees (one half page for each)
  - From one reviewer: « I feel that in general the data treatment has been appropriately carried out. I would suspect that a statistician could raise some technical questions but that is not the point of the paper »
- Revised version:
  - Paper shortened (abstract)
  - Two tables combined
  - Blind test procedures are not « relaxed » anymore but « abandoned »
  - The different sigma levels not modified
    - « We have tried to clarify the different sigma levels used. However, in order to determine the required 68% and 95% confidence limits, it is necessary to use 1 and 2 sigma for [control samples] but 1,1 and 2,6 sigma for [the Turin Shroud] » (Tite to Nature editor, 17 January 1989)



ature

Macmillan Magazines Ltd
4 Little Essex Street
London WC2R 3LF
Telephone 01-836 6633
Telex 262024

In reply please quote:
T12037 LG/SEM

29 December 1988

Dr M S Tite
The British Museum
Research Laboratory
LONDON
WC1B 3DG

Dear Dr Tite,

Your manuscript, "Radiocarbon dating of the Shroud of Turin", has now been seen by two referees, whose reports I enclose. Both referees are positive, and provided their comments and a few editorial concerns are addressed satisfactorily in a revised manuscript, we will be happy to publish your paper.

As both referees note, the paper is essentially a report of an inter-calibration experiment, and as such it is dominated by the kind of technical detail that we normally try to minimize in Nature papers. While this is in the nature of the beast, and we won't ask you to put the text into figure and table legends, it would be nice to reduce somewhat the space occupied by technical detail. In this regard as some of the information in Table 4 is also in Table 2, we would like you to combine these two tables. Also, there is a proliferation of confidence intervals in the paper: 1 sigma in Table 1, 2 sigma in Figure 1, 1.1 and 2.6 sigma in Table 4, etc.. Would it not be possible to standardize on one (or at most, two) confidence intervals? Please state what limits are being cited in each table or figure in its legend, rather than just in the text.

I have a few other points of confusion to raise. The term "known-age" control seems to imply that the labs knew the ages of the controls, which I assume is not the case. Could please clarify this, and perhaps simply use the word rol" in most cases? On page 9, second line from bottom, you please explain why samples 2 and 4 offer "little alternative analyses"? In Table 1, footnote 2, ere one anomalous result "(of 6)", when the text that each lab "performed between three and five surements"? Finally, in Table 3, wouldn't opriate to quote a significance value of

Comments by the two referees

A.

Comments on "Radiocarbon Dating of Shroud of Turin"

The medieval age of the Shroud of Turin based on AMS ^{14}C analysis has already been "published" in the popular media. The scientific value of this paper is not, in my view, the age of a piece of medieval textile but the inter-/intra-laboratory calibration experiment for three AMS laboratories which has been performed. I have gone over this data in detail and have several minor questions, but I feel that in general the data treatment has been appropriately carried out. I would suspect that a statistician could raise some technical questions but that is not the point of the paper.

I have only one major comment. It is extremely unfortunate that the original blind test protocols were not followed. Some more detailed explanation needs to state exactly as to why they were not carried out. The statement on page 4 that pretreatment cleaning with unravelled or shredded samples would have been more difficult and wasteful of sample is not a compelling in light of the original test protocols that took this into account. The experiments which were carried out were in no sense blind tests. The blind test procedures were not relaxed (page 4: ". . . it was decided to relax blind test procedures. . ."). In fact, they were abandoned. I would think this wording needs to be modified.

The data standing behind the public announcement needs to be published as soon as possible since it is of interest to a wide spectrum of disciplines.

B

The sampling strategy, the technical aspects of the measuring process, the statistical interpretation and the scientific analysis all are in good shape.

The report would simply be an inter-calibration project were it not for the religious aspects of the shroud's history. Given these aspects, the report should be of interest to a wide readership.

It would be useful to devote a paragraph or two to the difference (if any) expected in age calibration when materials formed in a single year (presumably the shroud fibers) are calibrated against a bi-decadal calibration curve.

The summary should be drastically shortened. Less so the remaining text, but there is some verbosity that needs an editorial touch.

Lessons From Failure

- Our paper
  - supports the growing contestation about the reliability of the Turin Shroud radiocarbon dating
  - No ‘conclusive evidence’
  - supports the hypothesis of a reproducibility crisis, even in physical sciences
    - This weakness partly explains the ongoing controversies about the Turin Shroud
    - Shows also the potential fragility of our knowledge of the cloth
    - This replication crisis should be taken into account in the development of robust protocols

Thank you for your attention!