METAL 2010

International Conference On Metal Conservation

Interim Meeting Of The International Council Of Museums Committee For Conservation Metal Working Group

October 11-15, 2010
Charleston, South Carolina, USA

EDITED BY:
PAUL MARDIKIAN
CLAUDIA CHEMELLO
CHRISTOPHER WATTERS
PETER HULL

CLEMSON CONSERVATION CENTER
Where Materials Science and Engineering Meet Archaeology and Conservation

CLEMSON UNIVERSITY
RESTORATION INSTITUTE

ICOM
COMMITTEE FOR CONSERVATION
COMITÉ POUR LA CONSERVATION
COMITÉ PARA LA CONSERVACIÓN

ISBN 978-0-9830399-0-7

Photograph: Detail from ‘Bridge No. 2’ from the series *Rust Never Sleeps*, John Moore, 1996

AN EVALUATION OF INTER-LABORATORY REPRODUCIBILITY FOR QUANTITATIVE XRF OF HISTORIC COPPER ALLOYS

Arlen Heginbotham[1*], Aniko Bezur[2], Michel Bouchard[3], Jeffrey M. Davis[4], Katherine Eremin[5], James H. Frantz[6], Lisha Glinsman[7], Lee-Ann Hayek[8], Duncan Hook[9], Vicky Kantarelou[10], Andreas Germanos Karydas[10, 11], Lynn Lee[5], Jennifer Mass[12], Catherine Matsen[12], Blythe McCarthy[13], Molly McGath[13], Aaron Shugar[14], Jane Sirois[15], Dylan Smith[7], Robert J. Speakman[16]

[1] Decorative Arts and Sculpture Conservation Department
J. Paul Getty Museum
1200 Getty Center Drive, Suite 1000
Los Angeles, CA 90049-1687
USA

[2] The Museum of Fine Arts, Houston

[3] The Getty Conservation Institute

[4] The National Institute of Standards and Technology

[5] Harvard Art Museum

[6] The Metropolitan Museum of Art

[7] National Gallery of Art, Washington, D.C.

[8] The Smithsonian Institution

[9] The British Museum

[10] NCSR "Demokritos" – Institute of Nuclear Physics – Athens, Greece

[11] Nuclear Spectrometry and Applications Laboratory, International Atomic Energy Agency (IAEA) – Vienna, Austria

[12] The Winterthur Museum/ University of Delaware

[13] The Freer Gallery of Art and Arthur M. Sackler Gallery

[14] Buffalo State College

[15] The Canadian Conservation Institute

[16] Museum Conservation Institute, Smithsonian Institution

* Corresponding author: aheginbotham@getty.edu

Abstract

This paper reports the results of a study conducted to evaluate the current state of inter-laboratory reproducibility when conducting quantitative XRF analysis of historic copper alloys. Fourteen institutions, primarily from the museum community, participated in the study, using a total of 19 X-ray fluorescence instruments. The design of the study was based largely on ASTM standard E1601, Standard Practice for Conducting an Interlaboratory Study to Evaluate the Performance of an Analytical Method. *In addition to addressing overall inter-laboratory reproducibility, we also attempt to evaluate the accuracy of individual laboratories. By determining correlations between accurate results and experimental methods and procedures, we are able to propose recommendations regarding best practice and ways in which reproducibility might be improved.*

Keywords: inter-laboratory reproducibility, X-ray fluorescence, copper alloys, fundamental parameters, ASTM E1601

Introduction

Since at least the late 1950s, a number of papers have been published that report quantitative analyses of historic copper alloys based on X-ray Fluorescence Spectroscopy (XRF). With the recent widespread introduction and adoption of relatively low-cost, portable XRF spectrometers, the pace of publication of such data is increasing and is likely to accelerate further. Although we welcome these advances, the rapid proliferation and publication of XRF data raises a host of important questions concerning the accuracy and inter-laboratory comparability and reproducibility of published data. While within laboratory conclusions based on quantitative XRF analysis may be interesting and instructive, comparing data between laboratories, or even between different instruments within a laboratory, can be problematic. Traditionally, such issues have been dealt with via formal or informal inter-laboratory analyses in which common reference materials are measured (e.g., (Glascock 1999, Hein 2002)). Since W. T. Chase's signal 1974 paper, 'Comparative analysis of archaeological bronzes' (Chase 1974), we know of only one other published study that has attempted to evaluate the inter-laboratory reproducibility of quantitative XRF on historic copper alloys (Northover and Rychner 1998). Neither of these publications focused primarily on XRF, but rather on reproducibility *between* techniques. Both publications also focused on copper alloys where the primary alloying metals were tin and lead.

Building on an earlier workshop and XRF round-robin organized by the Getty Conservation Institute, the National Gallery of Art in Washington hosted a seminal meeting in 2007 of representatives from seven museums to address issues surrounding the sharing and comparability of quantitative XRF data between institutions. That meeting, sponsored by Robert H. Smith and the Center for Advanced Study in the Visual Arts, focused on these issues particularly as they relate to the analysis of Renaissance bronze sculpture. Moderated by then Senior Curator of Sculpture, Nicholas Penny and Head of the Object Conservation Department, Shelley Sturman, the participants agreed that the ability to compare data would be valuable, but enumerated a host of problems and obstacles to be overcome before meaningful inter-laboratory comparisons could be made. This study is a direct product of that encounter.

The program described here is an attempt to evaluate the current state of inter-laboratory reproducibility of quantitative XRF analysis of copper alloys. We conducted, interpreted and summarized data generated from a carefully designed study informed by ASTM standard methodology (ASTM 2006, ASTM 2003). By quantifying the extent of reproducibility, we hope to provide valuable quantitative guidelines for practitioners who might wish to compare their own quantitative data with that generated by other laboratories, or who might wish to pursue meta-studies based on the work of many laboratories.

Our study sought participation primarily from laboratories in the museum community whose interests include a focus on historic copper alloys. In addition, we sought to include a variety of instrument types, supported by a variety of quantification procedures and software.

In addition to addressing overall inter-laboratory reproducibility, we also attempt to evaluate the accuracy of individual laboratories. By determining correlations between accurate results and experimental methods and procedures, we are able to propose some recommendations regarding best practice.

Methods

Research Design

Seventeen institutions agreed to participate in the study. Of these, many hoped to produce multiple data sets by using more than one instrument or by processing data from one instrument using multiple methods. Therefore, the maximum number of data sets anticipated was 30. In order to maintain anonymity throughout the study, each institution was assigned a laboratory number for each anticipated data set. This number was known only to the members of that institution and to the program coordinator (Heginbotham). Fourteen, or 82%, of the institutions turned in complete results and the total number of data sets included in the study is 19. In one case, the same instrument was used to produce three data sets by processing the same raw spectra using three different methods[1].

Eight instruments were used in the study. These include Bruker/Keymaster Tracer, Bruker/Roentec Artax, EDAX Eagle 3, Elva-X light, Innov-X XT-260, Niton Gold, Spectrace Omega 5, and laboratory-built models. While many laboratories chose to use the manufacturer's proprietary software for quantification, many others used or created customized solutions, ranging from spreadsheet-based analysis, to complete programs written in-house, to the use of X-ray analysis software available on the Internet.

The design of the study was based largely on ASTM standard E1601, *Standard Practice for Conducting an Interlaboratory Study to Evaluate the Performance of an Analytical Method*, following Test Plan A. Each participating laboratory was asked to analyze a set of 12 samples of metal (designated A-L). The same sample set was circulated to each participant via a traceable shipper over the course of eight months. The test samples consisted of three types: 1- cuttings obtained from reference materials[2] (RMs) n=4; 2- pieces of historic metal, n=6; 3- small ingots prepared by the lead author, n=2. The range of elemental compositions included in these samples was tailored to imitate the broad range can be found in historic copper alloy artifacts from the Bronze Age through the 19th century. A table presented in the *Results* section below provides brief descriptions of the 12 samples, their approximate compositions, and the range of concentrations determined for each element.

On each sample, a circular site was selected for analysis with a diameter of approximately 9 mm. These sites were first flattened with 220-grit silicon carbide abrasive paper. They were then polished with successively finer grades of Micro-mesh™ abrasive cloths, finishing with 4,000-grit. All polishing was done wet in ethanol, and fresh abrasive was used for each sample. The sites designated for analysis were clearly circumscribed on each sample with a stylus, ensuring that the material analyzed would be the same across laboratories. The samples were individually bagged and placed in a padded case for transport. Participants were asked not to touch or otherwise disturb the sample sites.

Per ASTM standard E1601, each laboratory was asked to conduct triplicate analyses of each area on each sample according to their standard in-house procedures. Participants were asked to conduct analyses that would yield a result representative of the entire area. In addition, the three measurements were to be acquired in immediate succession with as little variation in procedure as possible.

Data Recording and Accumulation for Analysis

Each participating laboratory completed a standardized reporting form in spreadsheet format for each instrument used. If the same instrument was used in conjunction with more than one quantification method, a separate form was completed for each method employed. For every analysis, participants were asked to report on a minimum of 12 elements. These elements were Mn, Fe, Ni, Cu, Zn, As, Ag, Cd, Sn, Sb, Pb and Bi. Space was provided to report on additional elements if they were detected. If a quantitative result for a requested element could not be obtained, analysts were asked to choose from the following responses: BDL – below detection limit/not detected; Trace – element present in a small amount but not quantifiable; Present – element present in a significant amount but not quantifiable; N/A – element not analyzed for/not detectable by this instrument. Data for each sample from all laboratories were compiled in a master database for evaluation.

The reporting form provided to the participants also requested extensive detail about the instrument, software, and procedures utilized in each laboratory. Participants were asked to provide information about their instrument manufacturer, model, anode material, and detector type. Participants also reported on operating parameters, including voltage (kV), current (mA), measurement time, spot size, filters, typical number of live (valid) counts collected by the detector per second, and average dead time. Participants also reported the software and methodology used for quantification. This included the full name and version of software, the type of method used, the number of standards used, and the frequency of calibration checks and recalibration. In addition, participants were asked to report their errors and detection limits for each of the 12 elements listed above, and to specify how these values were determined.

Assessment Methods

Reproducibility Statistics

In general, our evaluation followed the guidelines presented in the ASTM E1601 (test plan A). For each set of triplicate results for a particular element in an individual sample, the mean result ($\bar{x}$) was calculated. The overall group mean ($\bar{X}$) was then calculated as

$$\bar{X} = (\Sigma \ \bar{x}) \ / \ p$$

where p = the number of laboratories reporting a quantitative result for that element. For each $\bar{x}$, the laboratory difference (d) was calculated as

$$d = \ \bar{x} - \ \bar{X}$$

The standard deviation of all laboratory differences ($s_{\bar{x}}$) was then calculated for each element in each sample as

$$S_{\bar{x}} = [\Sigma(d^2) \ / \ (p\text{-}1)]^{1/2}$$

These preliminary calculations allowed the calculation of a between-laboratory consistency statistic, designated as h, that provides a normalized measure of the difference between the reported result and the overall mean value of all laboratories' results for the same element and standard:

$$h = d \ / \ s_{\bar{x}}$$

Comparison of the h statistics to a table of critical values allowed outlying results, that is, results that deviated significantly from the overall group mean, to be identified and flagged for follow-up. Laboratories with flagged results were contacted and asked to check their records to see if any errors in procedure, analysis, or transcription of results could be identified. If any such errors were identified, the data were corrected, but if no errors were found, the data were retained as originally reported. Of 1,718 h statistics that were calculated, 48 (2.8%) were flagged as identifying outliers and 20 corrections were made by four laboratories.

Once errors were corrected, reproducibility statistics were calculated for each of the 12 requested elements in each sample. The reproducibility index (R) is a measure of precision and represents the expected variability of results when a method is used in different laboratories. Specifically:

> *Use R to predict how well your results should agree with those from another laboratory: First, obtain a result…, then add R to, and subtract R from, this result to form a concentration confidence interval. Such an interval has a 95% probability of including a result obtainable by the method should another laboratory analyze the same sample. For example, a result of 46.57% was obtained. If R for the method at about 45% is 0.543, the 95% confidence interval for the result (that is, one expected to include the result obtained in another laboratory 19 times out of 20) extends from 46.03 to 47.11%* (ASTM 2003).

The reproducibility index was calculated as:

$$R = 2.8\{(s_{\bar{x}})^2 + [(\Sigma(s^2) \ / \ p) \ (n\text{-}1) \ / \ n]\}^{1/2}$$

where s = the standard deviation of each laboratory's replicate measurements and n = the number of replicates (in this case, three[3]).

Finally, the percent relative reproducibility index (R_{rel}%), which represents R as a percentage of the overall mean, was calculated according to the formula:

$$R_{\mathrm{rel\%}} = 100R \ / \ \bar{X}$$

Lower Limits

A lower limit (L) was calculated for each element (with the exception of copper) below which the method is not considered reliable. This calculation was made according to the formula

$$L = 100R \ / \ e_{\max}$$

where R = element reproducibility index determined for the sample with the lowest concentration of the specific element, and $e_{\max}$ = maximum acceptable percent relative error. In this case, $e_{\max}$ was set to 50% based on ASTM guidelines.

Accuracy of Overall Median

It was hypothesized that the overall group median $\bar{\chi}$ would likely be a good approximation of the true concentration of each element in a sample. If true, then $\bar{\chi}$ values could be used to gauge the accuracy of individual laboratories for samples A-H. In order to verify this hypothesis, the accuracy of $\bar{\chi}$ values was evaluated for the four RMs (samples I-L). For each certified value (X) in the RMs, the percent error of the median was calculated:

$$\% \text{ error} = 100(\ \bar{\chi} - \mathrm{X}) \ / \ \mathrm{X}$$

Certified values that fell below the method's calculated lower limit (L) for that specific element were not considered in evaluating accuracy. The mean percentage error for all elements in the RMs was calculated using the absolute values of all percentage errors where X > L.

Ranking of Laboratories

The accuracy of each laboratory/instrument combination was evaluated on an element-by-element basis. For each quantitative result from a given laboratory, the laboratory difference from the assumed 'true' value (d_t) was calculated. For the four RMs (samples I-L), this was calculated as

$$d_t = \bar{x} - X$$

(recall that $\bar{x}$ = the laboratory's mean result and X = the certified value). For the non-reference samples (A-H) d_t was calculated as

$$d_t = \bar{x} - X_m$$

where X_m = the median value of all laboratory results.

If $X_m < L$ (the method's lower limit as defined above), then X_m was considered to be unreliable as a measure of the true value; therefore no *d* values were calculated and the element was not used for ranking purposes. As an added precaution, if fewer than 10 laboratories reported data for an element in a given standard, no *d* values were calculated and the element was not used for ranking purposes.

A normalized accuracy statistic (h_a) was then calculated by dividing the laboratory difference by the standard deviation of laboratory differences.

$$h_a = d_t /(\Sigma(d_t^2) / (p\text{-}1))^{1/2}$$

where *p* = the number of laboratories reporting a quantitative results for the element in the given sample.

For each laboratory, all h_a values for a given element were combined to generate a mean accuracy score ($S_{element}$) for that element according to the formula

$$S_{element} = \Sigma(h_a^2) / n$$

where *n* = the number of quantitative results reported for the given element for all 12 samples. Scores close to zero reflect results that are consistently close to the assumed true value[4].

All 19 laboratories reported quantitative results for Cu, Zn, Sn and Pb (hereafter referred to as the 'major elements'). An aggregate score for major elements (S_{major}) was calculated:

$$S_{major} = S_{Cu} + S_{Zn} + S_{Sn} + S_{Pb}$$

Only 15 laboratories reported quantitative results for all four of the elements Fe, Ni, As and Sb (hereafter referred to as the 'minor elements'). An aggregate score for minor elements (S_{minor}) was calculated for these laboratories:

$$S_{minor} = S_{Fe} + S_{Ni} + S_{As} + S_{Sb}$$

Only eight laboratories reported quantitative results for Bi, so S_{Bi} was not included in the calculation of S_{minor}. S_{Ag} also was rejected for inclusion in S_{minor} because the reproducibility of results for Ag was so poor that the median results (X_m) were not considered to be valid indicators of the true value.

Mn and Cd were sporadically reported by only a few laboratories, making any meaningful comparisons or calculation impossible. Consequently, discussion of these elements is omitted.

Correlations Between Accuracy Scores and Methods

Accuracy scores were compared with the descriptions of instrument specifications, operating parameters and methodology provided in the laboratories' reporting forms. In an attempt to identify 'best practices', we sought to identify characteristics that were common to the most accurate laboratories. No attempts were made to be quantitative in this assessment. Rather, general correlations were identified by simple graphical plotting of the data.

Results

Table 1 provides a summary of laboratory data collected in the reporting forms. Table 2 gives brief descriptions of the 12 samples along with their approximate compositions, and the range of concentrations covered by the set as a whole. For samples A-H, the values are based on the overall group median; for samples I-J, values are as listed by the manufacturer of the RM. Lower limits for samples A-H were defined as described in METHODS. The complete quantitative data reported by all laboratories is available at the following address:

http://www.getty.edu/museum/conservation/papers.html

Reproducibility Statistics and Lower Limits

Summary statistics as per ASTM for the eight most commonly identified elements are presented as a group in Table 3. For each element, the samples, or test materials, are sorted by overall mean weight percent. The method's lower limit (*L*) for each element is shown on the right side of the relevant sub-table except in the case of copper, for which no lower limit was calculated. A dashed line through the center of each sub-table separates materials whose overall mean concentration falls below *L* (above the line) from those where the mean is greater than *L* (below the line). The latter group constitutes the samples for which the method is considered valid. The mean value of the $R_{rel\%}$ statistics for these samples is shown at the bottom right of each sub-table. This statistic provides the most succinct summary, for each element, of the analytical reproducibility that may be currently anticipated within this group of laboratories, based on a 95% confidence interval.

Evaluation of Accuracy

Data for the four RMs are presented in Table 4. This table shows the group's overall median ($\bar{\chi}$) for all elements where reference or certified values are given. Percent errors are shown for elements where $\bar{\chi} > L$. The results show that, on average, $\bar{\chi}$ falls within 5% of the certified value in cases where $\bar{\chi}$ lies in the range of validity for the method. It was determined that $\bar{\chi}$, if greater than *L*, could be used as a reasonable approximation of the true value for the purposes of evaluating the accuracy of individual laboratories.

Laboratory Number	Tube target	Detector Type	kV	mA	Acqusition Time (s)	Spot size (mm)	Filters (element)	Counting rate (cps)	Quantification method	Number of Standards
1	Rh	PIN	40	2.5	90	6	Al Ti Cu	3500	Empirical	27
2	W	PIN	45	7.5	100	8	Ni Al	4800	FP	0
3	Re	PIN	40	1	400	6	Al Ti	5000	FP w/stds	29
6	Rh	SDD	50	0.6	600	0.05	Ti Co Pd	8000	FP w/stds	19
7	Re	PIN	40	1	400	6	Al Ti	6000	FP	0
8	Mo	SDD	50	0.8	300	0.9	None	30000	FP w/stds	4
9	Rh	Si-Li	40	0.1-0.3	300	0.054	None	10000	FP	0
10	Rh	Si-Li	45	1	100	8.5	Rh	6800	FP w/stds	8
12	Mo	SDD	50	0.6	150	0.07	none	4800	Empirical	12
13	Au	SDD	40	40	400	8	Ag	95000	FP w/stds	8
14	Rh	PIN	40	1.4	120	6	Al Ti	6000	Empirical	73
15	Rh	PIN	40	1.8	60	6	Al Ti	6300	Empirical	45
18	Rh	PIN	40	1.8	180	6	Al Ti	7300	Empirical	46
19	Rh	PIN	40	0.1	600	2.6	Ni V	700	FP w/stds	19
22	Re	PIN	40	1.5	400	6	Al	6500	Empirical	36
23	W	SDD	50	0.2	200	1.5	Ni	16000	Empirical	5
24	Ag	PIN	35	6	60	10	Al	5000	FP w/stds	5
27	Rh	PIN	40	0.003	300	5	Al Ti Cu	6250	Empirical	125
28	Rh	SDD	50	0.35	200	1.5	None	60000	Empirical	15

Table 1. Summary of laboratory data.

Sample:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**		
Description	Chinese Coin (unknown date)	Italian Upholstery Tack (17th Century?)	British Door Knob (18th Century?)	American Screwdriver Ferrule (19th Century)	British Auger Cover Plate (19th Century)	Laboratory-cast Ingot	Laboratory-cast Ingot	Dutch East India Company Coin (1754)	Brammer C934 (RM)	CITF B32 (CRM)	MBH 31X B27 A (CRM)	BNF C71.34-3 (CRM)	Minimum	Maximum
Fe	0.55	0.22	<0.17	<0.17	0.41	0.82	0.41	<0.17	0.01	0.1	0.31	0.29	0.01	0.82
Ni	<0.35	0.35	<0.35	<0.35	<0.35	0.96	1.1	<0.35	0.49	1.49	0.042	-	0.04	1.5
Cu	71	82	75	85	70	53	72	98	82.64	74.85	78.2	87.23	53	98
Zn	<0.79	9.3	22	3.6	28	34	3.0	<0.79	0.17	1.15	19.9	1.55	0.17	34
As	0.47	<0.25	<0.25	<0.25	0.29	2.52	0.93	0.25	-	0.0056	0.03	0.18	0.01	2.5
Ag	<0.15	<0.15	<0.15	<0.15	<0.15	<0.15	0.17	<0.15	-	-	-	0.025	0.03	0.17
Sn	4.1	4.6	<0.27	8.5	0.53	2.8	16	<0.27	8.1	5.9	0.92	8.2	0.53	16
Sb	0.22	0.12	<0.12	0.13	<0.12	3.0	1.9	0.87	0.14	0.13	0.04	0.071	0.04	3.0
Pb	24	3.3	1.9	2.3	<1.22	1.4	3.9	<1.22	8.45	16.1	0.24	2.47	0.24	24
Bi	<0.12	<0.12	<0.12	<0.12	<0.12	0.32	0.18	<0.12	-	-	0.055	0.029	0.03	0.32

Table 2. Compositions and descriptions of the 12 samples (A-L) used in the study. For samples A-H, values are based on the overall group median; for samples I-J, values are as certified by the manufacturer. Lower limits for samples A-H were defined as described in METHODS.

Iron - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
I	7	0.023	0.083	353	Calculated Lower Limit (L)
H	11	0.029	0.062	216	
J	17	0.126	0.192	153	**0.165**
C	18	0.135	0.156	115	
D	19	0.151	0.213	141	
B	19	0.236	0.303	128	
L	19	0.283	0.356	126	
K	19	0.363	0.412	113	
E	19	0.420	0.459	109	Mean $R_{rel\%}$ for $\bar{X}>L$
G	18	0.427	0.569	133	
A	19	0.592	0.696	118	**121%**
F	19	0.902	1.101	122	

Arsenic - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
K	10	0.041	0.074	179	Calculated Lower Limit (L)
I	5	0.054	0.142	262	
J	4	0.067	0.280	417	**0.246**
D	11	0.144	0.573	399	
C	14	0.146	0.231	159	
B	7	0.176	0.764	435	
L	13	0.213	0.746	350	
H	16	0.247	0.249	101	
E	16	0.291	0.337	116	Mean $R_{rel\%}$ for $\bar{X}>L$
A	10	0.444	0.489	110	
G	15	0.908	0.873	96	**110%**
F	16	2.558	3.261	127	

Nickel - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
K	12	0.074	0.177	238	Calculated Lower Limit (L)
C	10	0.082	0.240	292	
E	15	0.092	0.174	189	**0.354**
D	7	0.100	0.281	281	
A	12	0.145	0.211	146	
L	4	0.164	0.584	356	
H	14	0.197	0.267	136	
B	18	0.378	0.273	72	
I	17	0.462	0.242	52	Mean $R_{rel\%}$ for $\bar{X}>L$
G	17	1.040	0.582	56	
F	18	1.066	0.732	69	**61%**
J	18	1.475	0.820	56	

Tin - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
H	5	0.053	0.136	255	Calculated Lower Limit (L)
C	13	0.112	0.132	118	
E	19	0.529	0.235	44	**0.271**
K	19	0.866	0.315	36	
F	18	3.092	2.176	70	
A	18	4.320	1.660	38	
B	19	4.687	1.295	28	
J	19	5.951	2.330	39	
D	19	8.543	1.804	21	Mean $R_{rel\%}$ for $\bar{X}>L$
I	19	8.554	2.225	26	
L	19	8.608	3.953	46	**40%**
G	18	17.166	9.082	53	

Copper - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
F	19	53.249	10.289	19	Calculated Lower Limit (L)
E	19	69.855	3.031	4	
A	19	70.508	13.163	19	**n/a** **(see note a)**
G	18	71.481	8.980	13	
J	19	73.935	7.083	10	
C	19	75.236	3.123	4	
K	19	78.234	2.590	3	
I	19	81.758	2.701	3	
B	19	81.795	3.930	5	Mean $R_{rel\%}$ for $\bar{X}>L$
D	19	85.266	2.493	3	
L	19	86.209	6.725	8	**8%**
H	19	98.130	2.694	3	

Antimony - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
E	5	0.026	0.064	247	Calculated Lower Limit (L)
C	6	0.029	0.064	220	
K	8	0.030	0.060	199	**0.120**
I	13	0.126	0.252	200	
B	13	0.156	0.400	257	
D	12	0.157	0.259	165	
J	13	0.171	0.418	244	
L	11	0.177	0.932	528	
A	14	0.211	0.277	131	Mean $R_{rel\%}$ for $\bar{X}>L$
H	16	0.882	0.473	54	
G	16	1.857	0.738	40	**185%**
F	17	3.020	1.450	48	

Zinc - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
A	6	0.209	0.710	340	Calculated Lower Limit (L)
H	9	0.240	0.396	165	
I	12	0.315	0.401	127	**0.792**
J	19	1.132	1.016	90	
L	19	1.653	0.927	56	
G	18	3.005	1.127	38	
D	19	3.669	1.121	31	
B	19	9.376	1.799	19	
K	19	19.873	1.762	9	Mean $R_{rel\%}$ for $\bar{X}>L$
C	19	22.312	2.003	9	
E	19	27.733	1.968	7	**31%**
F	19	33.600	6.026	18	

Lead - Statistical Summary

Test Material	Number of laboratories (n)	Overall Mean ($\bar{X}$)	Reproducibility Index (R)	Percent Relative Reproducibility Index ($R_{rel\%}$)	
H	17	0.178	0.609	343	Calculated Lower Limit (L)
K	19	0.269	0.397	148	
E	19	1.033	0.693	67	**1.217**
C	19	1.939	0.670	35	
F	19	2.234	6.701	300	
D	19	2.283	0.659	29	
L	19	2.860	1.909	67	
B	19	3.247	1.210	37	
G	18	3.816	3.084	81	Mean $R_{rel\%}$ for $\bar{X}>L$
I	19	8.650	2.642	31	
J	19	17.346	10.256	59	**77%**
A	19	24.628	13.570	55	

a. Lower limits are only calculated where the element is to be analyzed near the lower end of its effective concentration range

Table 3. Statistical summaries; for each element, the samples are sorted by overall mean weight percent.

Ranking of Laboratories

Laboratories S_{major} and S_{minor} scores are shown in Table 5, ranked in order of highest to lowest accuracy.

Correlations with Performance

Quantification Method

Clearly, the strongest correlation between laboratory characteristics and accuracy was based on the type of method employed to convert raw elemental intensities into a quantitative result (see Figures 1a, 1b and Table 5).

Three major categories of method were reported by the participating laboratories: standardless fundamental parameter (FP); fundamental parameter calibrated with standards (FP w/standards); and algorithms using empirical coefficients (empirical).

FP methods are based on mathematical models that predict the intensity of fluorescent radiation from a sample of known composition. The models incorporate knowledge of many instrument parameters, such as incidence and take-off angles (for both anode and sample), anode material, detector area and thickness, voltage, attenuators (such as windows, filters and air path), etc. FP models generally account for matrix effects, such as absorption and secondary fluorescence (in which some portion of the characteristic photons

Sample I (C934)	# of Labs (*p*)	Reference Value	Overall Median ($\tilde{x}$)	% error
Fe	7	0.01	0.01	*
Ni	17	0.49	0.46	-7%
Cu	19	82.64	81.91	-1%
Zn	12	0.17	0.29	*
Sn	19	8.07	8.46	5%
Sb	13	0.14	0.10	-27%
Pb	19	8.45	8.82	4%

Sample J (CITF B32)	# of Labs (*p*)	Certified Value	Overall Median ($\tilde{x}$)	% error
Fe	17	0.10	0.11	*
Ni	18	1.49	1.50	1%
Cu	19	74.85	74.53	-0.4%
Zn	19	1.15	1.10	-4%
As	4	0.0056	0.0328	*
Sn	19	5.92	5.78	-2%
Sb	13	0.13	0.12	-5%
Pb	19	16.10	16.76	4%

Sample K (MBH 31X B27 A)	# of Labs (*p*)	Certified Value	Overall Median ($\tilde{x}$)	% error
Mn	11	0.045	0.046	2%
Fe	19	0.31	0.33	7%
Ni	12	0.042	0.054	*
Cu	19	78.2	78.4	0.3%
Zn	19	19.9	19.78	-1%
As	10	0.03	0.04	*
Sn	19	0.92	0.84	-9%
Sb	8	0.04	0.03	*
Pb	19	0.24	0.24	*
Bi	8	0.055	0.046	*

Sample L (BNF C71.34-3)	# of Labs (*p*)	Certified Value	Overall Median ($\tilde{x}$)	% error
Mn	12	0.05	0.05	0.0%
Fe	19	0.29	0.25	-13%
Cu	19	87.230	86.592	-1%
Zn	19	1.55	1.62	5%
As	13	0.18	0.17	-6%
Ag	10	0.025	0.038	*
Sn	19	8.20	8.43	3%
Sb	11	0.071	0.119	*
Pb	19	2.47	2.76	12%
Bi	4	0.029	0.025	*
			Mean error (median > *L*)	5%

*** certified value below *L***

Table 4. Comparison of certified values to group medians.

Ranking (major elements)	SCORE (S_{major})	Lab #	Quant Method
1	0.2	13	FP w/stds
2	0.6	24	FP w/stds
3	0.7	6	FP w/stds
4	0.9	19	FP w/stds
5	0.9	3	FP w/stds
6	1.2	8	FP w/stds
7	1.5	2	FP
8	1.8	23	Empirical
9	2.2	15	Empirical
10	3.2	14	Empirical
11	3.3	10	FP w/stds
12	3.7	28	Empirical
13	4.7	27	Empirical
14	5.7	1	Empirical
15	5.7	18	Empirical
16	9.1	12	Empirical
17	10.1	7	FP
18	11.1	22	Empirical
19	14.6	9	FP

Ranking (minor elements)	SCORE (S_{minor})	Lab #	Quant Method
1	0.3	19	FP w/stds
2	0.7	3	FP w/stds
3	0.8	13	FP w/stds
4	0.8	6	FP w/stds
5	1.9	8	FP w/stds
6	2.7	15	Empirical
7	2.8	1	Empirical
8	2.9	28	Empirical
9	3.3	23	Empirical
10	3.9	14	Empirical
11	3.9	2	FP
12	4.5	18	Empirical
13	5.2	22	Empirical
14	8.4	9	FP
15	14.1	7	FP

Table 5. Laboratories' S_{major} and S_{minor} scores, ranked in order of highest to lowest accuracy.

excited by incident x-rays cause enhanced fluorescence in the sample) through theoretically-derived mathematical equations. The complex calculations involved in this method rely on knowledge of many physical constants, such as mass-attenuation coefficients, fluorescence yields, absorption jump ratios, relative line intensities, absorption edges, etc. (de Vries and Vrebos 2002). Some FP applications allow for the use of pure element standards to help model the spectral distribution of the tube output (de Viguerie et al. 2009) or to model transmission efficiency by polycapillary lenses (Karydas et al. 2008). The use of pure element standards in this manner is still considered 'standardless' FP for the purposes of this study, as the standards are not used directly to generate scaling coefficients for analytes.

FP w/standards methods can take several forms. As the name implies, they are based on mathematical predictions of fluorescent intensity and provide a theoretical accounting for matrix effects as discussed above. However, these methods also perform corrections to the model, using spectra generated by the instrument in question, from reference standards of composition similar to that of the analyte. The corrections can be performed in a variety of ways (discussion of which is beyond the scope of this paper), but by and large they attempt to account for instrument-related factors (de Vries and Vrebos 2002).

Empirical calibrations are derived through the measurement of standards that are similar to the unknown. Ideally, the compositional standards have the same elements as the unknown, although the composition may be different. Comparing each elemental fluorescence intensity in the standard to the corresponding composition and fitting a regression between known points, analysts can interpolate between known values. The fluorescence intensity of the unknown is then compared to the calibrated regression, and the composition is derived. Empirical models typically account for absorption, secondary fluorescence and other matrix effects using empirically derived correction coefficients based on regression analysis (de Vries and Vrebos 2002).

For both major and minor elements, it is very clear that laboratories using fundamental parameters software calibrated with standards (labs #3, 6, 8, 10, 13, 19, 24), performed consistently more accurately than laboratories using other methods. Remarkably, of these seven laboratories, no two used the same type of instrument or the same brand of software.

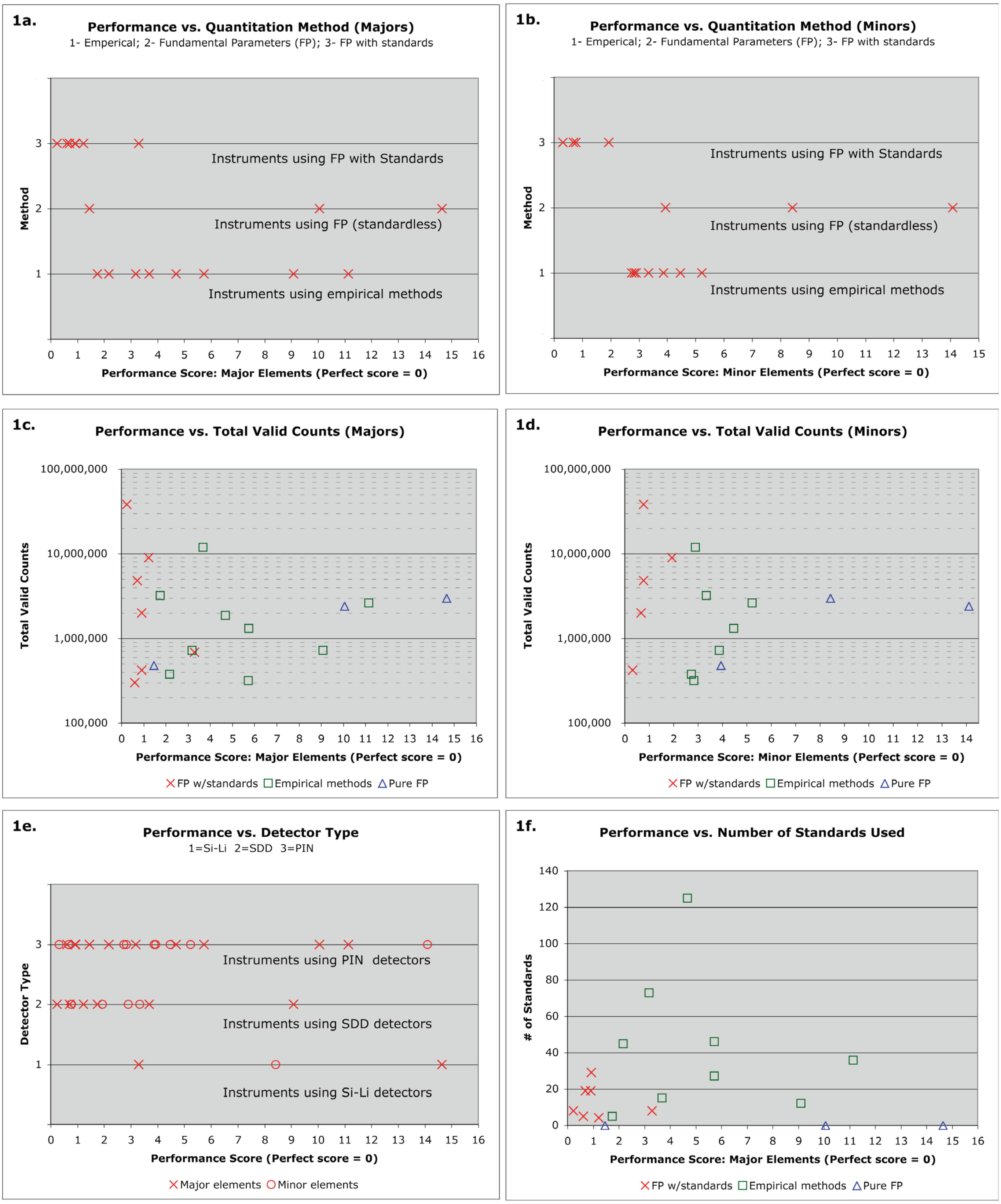

Figure 1. Performance scores (S_{major} and S_{minor}) are plotted against selected laboratory characteristics. High performance is reflected by a score close to zero (i.e. on the left side of the charts).

Among the laboratories using FP with standards, the majority had S_{major} and S_{minor} scores that were tightly clustered near the perfect score of zero. One of these laboratories (#10), however, appears to have performed noticeably less well than the others in the group (though their scores were still better than almost all other laboratories using empirical or standardless FP methods). In their data reporting form, the analysts for this instrument noted that 'While we did the analysis on this instrument, the stability of the instrument is doubtful and we would be cautious about reporting numbers from this instrument at the present time'. They also reported that the last full calibration of the instrument had been performed on the instrument more than four years ago. These observations may explain the difference in performance.

Only seven of the 19 data sets in the study (2, 3, 6, 7, 8, 13, 19) were able to consistently report quantitative results for the four major elements, plus six minor elements (Fe, Ni, As, Ag, Sb and Bi). Of these seven, all used FP or FP w/standards methodology.

Detector Type
Another laboratory characteristic that was evaluated for correlation with performance was detector type. The vast majority of laboratories in this study (84%) used silicon drift detectors (SDD) or silicon-PiN diode detectors (PIN). These are clearly the two dominant types of detectors on the market today. Only three instruments in the study used lithium drifted silicon detectors (Si-Li). Figure 1e plots S_{major} and S_{minor} against the three detector types in the study. While more of the poorly performing laboratories seemed to use PIN or Si-Li detectors than SDDs, it is perhaps more significant to note that the six top ranked laboratories (#3, 6, 8, 13, 19, 24) were equally divided between PIN and SDD detectors. It would appear then that very strong performance can be achieved with either of these detector types. Si-Li detectors did not appear to perform as well as the other types, but with only two laboratories using these detectors, the results should not be given too much weight.

Valid Counts per Analysis
A surprising result of the study is that, within the range employed in this study, the total number of valid counts per analysis (vca) was not positively correlated with performance, either for major or minor elements. The vca is given here as the product of the typical valid count rate per second (as reported by each laboratory and accounting for detector deadtime) and the number of seconds that the analysis was allowed to run. For both major and minor elements (Figure 1c and 1d), there appears to be no correlation between vca and performance. It is interesting and perhaps instructive to note that if the laboratories are grouped by software/analytical method, many of the best results for each group were attained with relatively low total counts, on the order of 300,000.

Number of Standards
The data also suggest that increasing the number of standards used for quantification does not necessarily improve the accuracy of results (Figure 1f). In fact, the vast majority of the best performing laboratories used 20 standards or fewer, and most used fewer than 10. Even among the labs using empirical methods, increasing the number of standards was not a guarantee of improved results, and the best performing laboratory of this group (for major elements) used only five standards per analysis.

Conclusions
This study evaluates the current state of inter-laboratory reproducibility of quantitative XRF analysis of historic copper alloys based upon a representative group drawn from the art and archaeology community, primarily in the United States. Nine members of the working group met for two days of intensive meetings to evaluate the results of the inter-laboratory study. The conclusions and recommendations of this sub-group were reviewed and commented on by the wider group. What follows is a summary of the overall findings of the working group.

Reproducibility
The overall reproducibility of the group's results is relatively poor. Even if one considers only results where the median result is above the calculated lower limits (*L*), the average percent relative reproducibility ($R_{rel\%}$) is greater than 50% for all elements except Cu, Zn, and Sn. The ASTM standard practice stipulates that the working group may determine the degree of precision that may be considered acceptable for a given method, based on the context within which the results are to be used. However, an upper threshold for $R_{rel\%}$ is set at 50% above which the methods reproducibility must be considered unacceptable. While not bound by ASTM guidelines, it was the consensus of the working group that the current reproducibility of XRF analysis of historic copper alloys within the art and archaeology community is, in general, *not* sufficient for any but the most broad comparisons to be made between laboratories.

Two examples drawn directly from table 3 may help serve to illustrate the point. Assume that a laboratory arrives at a result of 8.6% tin in a bronze alloy. Based of the current state of affairs, there is a 95% chance that another laboratory, measuring the same bronze, would arrive at a result somewhere between 4.6% and 12.6%. Similarly, for a result of 33% zinc in brass, the 95% confidence interval ranges from 27% to 39%. Also, considering that tin and zinc are among the elements with the best reproducibility in the study, the group agreed that concerted efforts should be made to improve the situation.

The reproducibility results reported here should evoke a strong sense of caution in those who might wish to publish data, compare their own data with that generated by other laboratories, or pursue meta-studies based on the work of multiple laboratories.

The group also found that the lower limits determined by this study (below which reproducibility rapidly deteriorates) are considerably higher than could be wished for. It was agreed that analyte concentrations below the lower limits are frequently of interest and significance to scientists engaged in the study of historic copper alloys.

Quantification Method

Through this study, one common characteristic of higher-performing laboratories has become clear: the use of fundamental parameters software, calibrated with standards. In comparison, all other factors examined in this study appear to be relatively poorly correlated with laboratory accuracy. The consensus of the group is that options should be explored for ways in which existing instruments that currently use empirical or standardless FP methods could be upgraded to use FP with standards.

A sense of the magnitude of improvement that such a change, if widely adopted, might bring about can be gleaned from Table 6. This presents the method's lower limits (*L*) and the percent relative reproducibility ($R_{rel\%}$) for all participating laboratories compared to the same statistics calculated based only on six laboratories using FP with standards (laboratory 10 was excluded from the group based on their self-described instrumental irregularities). On average, the sub group using FP with standards[5] had a reduction in lower limits of 65% from the overall group limits, reflecting a substantial improvement in their ability to compare results when element concentrations are low. Similarly, the subgroup's $R_{rel\%}$ values were, on average, 55% less than those of the group as a whole. While these levels may still leave something to be desired, it seems apparent that as a first step, movement toward the wider adoption of quantification methods utilizing FP with standards offers the possibility of significant improvements in interlaboratory reproducibility.

		All Participants	Participants using FP with Standards
Iron	Calculated Lower Limit (*L*)	**0.165**	**0.063**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**121%**	**41%**
Nickel	Calculated Lower Limit (*L*)	**0.354**	**0.057**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**61%**	**47%**
Copper	Calculated Lower Limit (*L*)	**n/a**	**n/a**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**8%**	**2%**
Zinc	Calculated Lower Limit (*L*)	**0.792**	**0.147**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**31%**	**15%**
Arsenic	Calculated Lower Limit (*L*)	**0.246**	**0.193**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**110%**	**64%**
Tin	Calculated Lower Limit (*L*)	**0.271**	**0.121**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**40%**	**14%**
Antimony	Calculated Lower Limit (*L*)	**0.120**	**0.037**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**185%**	**83%**
Lead	Calculated Lower Limit (*L*)	**1.217**	**0.226**
	Mean $R_{rel\%}$ for $\bar{X}>L$	**77%**	**27%**

Table 6. 'Method's Lower Limits' and 'Percent Relative Reproducibility Indices' as calculated for all 19 data sets compared with the same statistics calculated for the six top performing data sets, all using FP with standards software.

It was suggested by some members of the working group that the use of a common, open source FP software package[6] used in conjunction with a common and readily available set of reference materials could further improve the reproducibility of results within the group.

Many participants expressed a desire to have a set of certified reference materials, replicated for the various institutions that wish to share data, which includes a range of major and trace elements appropriate for historic alloys. Although a selection of available standards might fill a portion of this range, such a set would certainly require some standards to be newly manufactured.

Error and Detection Limits

Reporting of error and detection limits was inconsistent among the participating laboratories. Several laboratories did not report errors or detection limits at all. Many laboratories reported errors calculated from their software based on counting statistics. While these values have meaning, they generally reflect the error associated with repeated analyses by the same instrument (or instrumental precision) rather than expected error with respect to the true value (instrumental accuracy).

The laboratories that produced the most meaningful and reliable error values relative to true values did so by analyzing multiple reference materials and conducting a regression analysis of certified vs. calculated values. These laboratories used the 'standard error' associated with the regression to define meaningful confidence intervals relative to the estimated true value. This strategy was employed both by laboratories using both FP w/standards and empirical methods.

Detection limits were, if anything, less consistently reported than errors. Some laboratories did not report detection limits while others estimated them for selected elements based on experience with standards. Several participants derived their detection limits based on analysis of multiple reference materials with certified values at or near zero. A regression analysis was performed and a value of two or three times the standard error was used to estimate the nominal detection limit. The consensus of the group was that this empirical approach provides useful results in a relatively straightforward manner though other means are possible (Ziebold 1967, Long and Winefordner 1983).

Other Suggestions

The working group suggests that, in instances where data are to be published or shared between laboratories, standard practice should include publication (perhaps separately) of a detailed and comprehensive reporting of the laboratory method along with the presentation of empirically derived error and detection limit values. In addition, it was suggested that publication of data include results for one or two control samples (e.g., reference materials analyzed during the analysis, but which are not part of the calibration).

In some areas, the raw data generated in this study has only been superficially evaluated and many more conclusions may be possible with further data analysis. A number of significant subjects possibly could be addressed using the data already collected. For instance, the relative advantages and disadvantages of different variants of the FP w/standards method; the factors affecting detection limits; factors affecting within-laboratory precision; the effects of filtration; and the importance of careful manual inspection of spectra.

Clearly, much work remains on the issue of inter-laboratory reproducibility of XRF data generated for historical metals. As we have shown above, results among laboratories vary widely, not only for minor elements, but also for major elements. These differences highlight the problems associated with trying to compare data from multiple laboratories and the need for common standards and quantification approaches. Future research in this area should focus on addressing these issues.

Endnotes

[1] The results from this instrument are designated with laboratory numbers 3, 7, and 22.

[2] Three of the reference materials were certified (so-called CRMs) based on analysis by multiple laboratories (samples J, K, and L); one, (sample I) has no certificate of analysis.

[3]A complete explanation of this calculation is given in the ASTM E1601 sections 10.4.5 to 10.4.8. The validity of the formula is contingent upon the result being larger than the method's minimum standard deviation, which was true in every instance in this study.

[4]Using the square of h_a has the dual advantages of making all values positive and of emphasizing the negative impact of occasional poor scores. It would be equally valid to rank based on the absolute value of h_a; in practice, the rank order changes very little.

[5] Six is the minimum number of participating laboratories required by ASTM E1601. The results calculated for this subgroup may therefore be considered as 'valid' based on the standard procedure.

[6] Several such software packages are available, such as PyMCA (European Synchrotron Radiation Facility) and AXIL-QXAS (International Atomic Energy Agency).

References

Hein, A., A. Tsolakidou, I. Iliopoulos, H. Mommsen, J. Buxeda i Garrigós, G. Montanac, and V. Kilikogloua. 'Standardisation of Elemental Analytical Techniques Applied to Provenance Studies of Archaeological Ceramics: An Inter Laboratory Calibration Study', Analyst 127 (2002) 524-53.

ASTM, 'Designation: E 1601 – 98 (Reapproved 2003) ^{E1}: Standard Practice for Conducting an Interlaboratory Study to Evaluate the Performance of an Analytical Method'. West Conshohocken: ASTM International (2003).

ASTM, 'Designation: E 1763 – 06: Standard Guide for Interpretation and Use of Results from Interlaboratory Testing of Chemical Analysis Methods'. West Conshohocken: ASTM International (2006).

Chase, W.T., 'Comparative Analysis of Archaeological Bronzes', In *Archaeological chemistry: a symposium sponsored by the Division of the History of Chemistry at the 165th meeting of the American Chemical Society, Dallas, Tex., April 9 - 10, 1973*, American Chemical Society, Washington, DC (1974) 148-85.

de Viguerie, L., A. Duran, A. Bouquillon, V. A. Solé, J. Castaing, and P. Walter. 'Quantitative X-Ray Fluorescence Analysis of an Egyptian Faience Pendant and Comparison with Pixe', *Anal Bioanal Chem* 395 ((2009) 2219-25.

de Vries, J. L., and B. A. R. Vrebos. 'Quantification of Infinitely Thick Specimens by Xrf Analysis', in *Handbook of X-Ray Spectrometry*, ed. R. van Grieken and Andrzej Markowicz, New York: Marcel Dekker, (2002) 341-406.

Glascock, M.D., 'An Inter-Laboratory Comparison of Element Compositions for Two Obsidian Sources', *IAOS Bulletin* 23 (1999) 13-25.

Karydas, A.G., D. Anglos, and M.A. Harith. 'Mobile Spectrometers for Diagnostic Micro-Analysis of Ancient Metal Objects', in *Metals and Museums in the Mediterranean: Protecting, Preserving and Interpreting*, ed. V. Argyropoulos, Athens: TEI of Athens, (2008) 141-77.

Long, G.L., and J.D. Winefordner. 'Limit of Detection - a Closer Look at the Iupac Definition', *Analytical Chemistry* 55 ((1983) 712A-24A.

Northover, P., and V. Rychner. 'Bronze Analysis: Experience of a Comparative Programme', In *Bronze '96: L'Atelier du bronzier en Europe du XXe au VIIIe siècle avant Notre Ère*, edited by C. Mordant, M. Pernot and V. Rychner, 19-40. Neuchâtel: Editions du CTHS, (1998) 19-40

Ziebold, T.O., 'Precision and Sensitivity in Electron Microprobe Analysis', *Analytical Chemistry* 39 (8), (1967) 858–61.

Author

Arlen Heginbotham received his B.A. in East Asian Studies from Stanford University and his M.A. in Art Conservation from Buffalo State College. He is currently Associate Conservator of Decorative Arts and Sculpture at the Getty Museum where he is currently writing technical essays for catalogs of the Museum's collections of French furniture. Arlen's research interests include the history and analysis of 17th century East Asian export lacquer, the history of metallurgy, the use of X-ray fluorescence spectroscopy as a tool for authenticating and interpreting gilded bronzes, microscopic and chemical wood identification, immunochemical analysis, and the history of wood dyes.