Assessing Reliability in Faunal Analysis Using Blind Tests: A Case Study from Northwestern North America

RENO NIMS^a and VIRGINIA L. BUTLER^a

ABSTRACT

Zooarchaeologists have long recognized that assigning taxonomic identifications to animal remains is a subjective process, and recent studies have highlighted the need for data quality assurance standards in archaeofaunal research. Our study contributes to this growing interest in quality assurance by presenting simple quantitative methods for assessing reliability in analytic results through blind reanalysis of animal remains that we developed during analysis of fishbone from Tse-whit-zen, a large Native American village on the coast of Washington State, U.S.A. Given the large scale of the Tse-whit-zen project – over 112,000 fish remains were documented by five different analysts over three years – there was a real possibility that inconsistencies in laboratory practices affected analytic results (e.g., number of identified specimens, taxa present and relative abundance, elements identified, burning frequencies). To evaluate the overall reliability of the Tse-whit-zen fishbone data, and the possibility of “protocol drift” – changes in how specimens and bone modifications were identified over the course of analysis – we reanalyzed samples of fish remains that were previously documented during three discrete stages (beginning, middle, and end) of the Tse-whit-zen project. The original data and the reanalysis results show close agreement in each stage, with only minor differences in the numbers of recorded specimens, taxonomic representation at family- and finer taxonomic levels, and body part representation assessed for a single taxonomic order. Identifying burning on bone was not very reliable. Reliability studies such as this are useful for highlighting ambiguous identification criteria (e.g., in taxonomic assignment, bone modification), and could stimulate dialog among researchers about ways to address such issues in future studies. Increasing the implementation of these, and other, widely applicable quality control methods should improve zooarchaeological data quality and stimulate further research on quality assurance in archaeology overall.

Keywords: BLIND REANALYSIS; DATA QUALITY CONTROL; RELIABILITY; ZOOARCHAEOLOGY; AGREEMENT INDEX; PROTOCOL DRIFT

1. INTRODUCTION

Zooarchaeologists have long recognized that assigning taxonomic identifications to animal remains is a subjective process (Driver, 1992 [2011]; Gobalet, 2001; Lawrence, 1973; Wolverton, 2013). In one particularly striking demonstration of this fact, Gobalet (2001) and three other ichthyoarchaeologists separately analyzed the same, coastal California fishbone assemblage and found they had major disagreements regarding the number of species, number of identifiable specimens (NISP), and taxonomic identity of specific specimens represented in the collection. For example, one analyst identified 18 species in the assemblage, two researchers identified half that number, and one researcher only identified four species (Gobalet, 2001, p. 378). Such disagreements might be related to differences in the educational background and experience of individual analysts, analysts' familiarity with the biological community linked to the site's context (Gobalet, 2001, p. 384), or yet other factors. Overall the study highlights the pressing need for quality assurance practices in zooarchaeology.

Building on examples from environmental chemistry, Wolverton (2013) has provided a framework for *quality assurance* in zooarchaeology, which includes "activities intended to ensure that the analytical information produced meets the quality requisites...[of the field] in terms of accuracy and representativeness" (Perez-Bendito and Rubio 1999:39, *in* Wolverton 2013, p. 383-384). Wolverton breaks quality assurance into two components: *quality control* (QC) and *quality assessment* (QA). QC is an explicit set of steps that ensure laboratory practices are followed consistently, increasing replicability of results. QA includes practices that verify whether the results of analyses are valid. Wolverton (2013) provides numerous examples of methods that zooarchaeologists can use to attend to both QC and QA, which are necessary for supporting the long-term value of our data in archaeology and other research domains. As we seek to apply zooarchaeological research to conservation biology and policy debates, the need for rigorous analytic methods and documentation aligned with quality assurance increases in kind.

Our study contributes to the growing interest in quality assurance in zooarchaeology (e.g. Fisher, 2015; Morin et al., *in press*; Twiss et al., 2016) by introducing simple quantitative methods for assessing reliability in faunal analysis through reanalysis that we developed during study of fish remains from Tse-whit-zen, a large Native American village from coastal Washington State, U.S.A. (Figure 1) that was excavated extensively in 2004. The fishbone study is part of a larger analysis of mammal, bird, fish and shellfish remains that aims to understand how animal resources and in turn people were affected by great earthquakes and other environmental forces, such as climate change and local bay development, over the past ~2,200 years. Given the large scale of the Tse-whit-zen project – over 112,000 fish remains

were documented by five different analysts over three years – possible inconsistencies in laboratory practices may have affected analytic results (e.g., number of identified specimens, taxa present and relative abundance, elements identified, burning frequencies). We were concerned that temporal and spatial patterns documented in the fishbone data could reflect analytic error, rather than past human activity.

Figure 1 – Location of Tse-whit-zen in northwest Washington State, USA (drawn by Kristina Dick and Reno Nims).

Here we describe several simple quantitative methods for assessing reliability in faunal analysis that we applied in a ‘blind’ reanalysis of fishbone samples which were previously documented during three discrete stages of the Tse-whit-zen project. Though we focus on fish remains in particular and zooarchaeology generally, the methods we outline could also be applied to archaeological analysis broadly. In addition to assessing the overall reliability of our data, we evaluate whether increasing knowledge of, and experience with, the ichthyofaunal materials created ‘protocol drift’ – changes in

how specimens and human modifications to bone were identified over the course of analysis. Reliability studies such as ours also highlight ambiguous criteria (e.g., in taxonomic assignment, bone modification), and help stimulate dialog among researchers about ways to address such issues in future studies. Finally, it is our hope that increased focus on widely applicable QC methods such as these will help improve zooarchaeological data quality and stimulate further research on quality assurance in archaeology overall (e.g., Banning et al., 2016; Beck and Jones, 1989; Evans et al., 2014; Fish, 1978).

2. MATERIALS AND METHODS

The Tse-whit-zen fishbone assemblage comes from a Lower Elwha Klallam Tribe (LEKT) village in Port Angeles, WA at the base of Ediz Hook on the south shore of the Strait of Juan de Fuca (Figure 1) that was occupied for the past 2,800 years (Larson, 2006). In 2004, Larson Anthropological Archaeological Services (LAAS) and LEKT members excavated 261.4 m³ of the site's volume (estimated at 6,900 m³) in large open blocks of 1 x 1 m units by natural stratigraphic layers that were subdivided into 10 cm arbitrary levels (Reetz et al., 2006). All excavated matrix was collected in 10 L buckets and wet-screened with nested 1" (25.6 mm), 1/2" (12.8 mm), and 1/4" (6.4 mm) mesh. Every twentieth bucket from a given stratum was also screened to 1/8" (3.2 mm) mesh and designated a "complete" ("C") bucket (Reetz et al., 2006). All faunal remains from "C" buckets were sorted into four main animal groups (fish, bird, mammal, shellfish), but for various reasons they were not studied immediately. In 2010, Kristine Bovy (University of Rhode Island), Virginia Butler (Portland State University [PSU]), Sarah Campbell (Western Washington University), Michael Etnier (Western Washington University), and Sarah Sterling (PSU) initiated a large-scale zooarchaeological and geoarchaeological analysis of three structures at Tse-whit-zen.

2.1 Original Analysis Methods

Butler directed analysis of the fishbone assemblage at PSU between 2012 and 2015 following QC protocols that were consistent with Driver's (2011) recommendations (see also Wolverton, 2013): the universe of possible fish taxa was established at the beginning of the project using Strait of Juan de Fuca fisheries survey records (Miller et al. [1980] and north Pacific field guides [Hart, 1973]); identification criteria were specified and referred to over the course of the project; difficult to distinguish taxa and elements were specified; and a descriptive summary, a written document outlining the rules, protocols, and criteria used for assigning skeletal elements to taxon, is being prepared (Butler, *in prep.*). PSU Master's students Kathryn Mohlenhoff, Anthony Hofkamp, Shoshana Rosenberg, and Nims identified

fish remains using the PSU Anthropology Department comparative collection, which was supplemented with taxa loaned from the personal collections of Bob Kopperl (Willamette Cultural Resources Associates) and Ross Smith (University of Oregon) (Supplementary Table 1). Collections at the University of Victoria, Department of Anthropology, were also consulted with help from Rebecca Wigen (Supplementary Table 1). Butler verified, and often adjusted, all initial identifications under magnification (2x-40x power) with a binocular loupe or microscope.

For each specimen, we documented the finest possible taxon and skeletal element, whether the specimen included a morphological landmark, and whether the specimen was burned. For many specimens, it was difficult to determine whether a dark color reflected exposure to heat, flecks of charcoal attached to the surface of a specimen, or staining from the absorption of minerals found in the surrounding matrix. The decision to call a specimen burned was based on conservative criteria: only those specimens which were at least partially black (charred) or white-to-bluish (calcined) in color were called burned.

Given the large sample size and the small size of many specimens, it was impractical for each specimen to receive a unique catalog number or separate bag. Instead, we tallied and recorded the number of specimens from a given provenience and mesh size that shared the same attributes (e.g., taxon, skeletal element, burned) together. All fish remains from a given context with similar attributes were bagged together, with acid-free labels summarizing analytic information. Specimens which could not be identified to any taxon other than fish were recorded as unidentified, with one exception. The category 'non-salmonid' was used to tally unidentifiable vertebra fragments that were definitively not from salmonids in an effort to counter the potential over-documentation of salmon specimens, which can be easily identified to genus from very small vertebra fragments (e.g. Casteel, 1976; Ewonus, 2011; Grier and Lukowski, 2012). The primary counting units used for fish faunal analysis were NISP and number of specimens (NSP), which includes all remains assigned to fish, including unidentified fish and non-salmonid. In addition, we counted the minimum number of elements (MNE) using the number of specimens that bear morphological landmarks. Specimens identified as mammal, bird, or invertebrate were sent to the appropriate co-principle investigator of the Tse-whit-zen project for analysis and curation and not recorded during fishbone analysis. Specimens that we could not assign to a specific animal group, or that we determined were not animal at all (e.g. wood, rock), were not recorded, but were curated with the fish remains.

2.2 Reanalysis Methods

The goals of the reanalysis project were to evaluate overall reliability and whether reliability increased over time, which would reflect growing knowledge and/or more consistent application of identification protocols. Thus we selected samples for reanalysis from three ‘analytic stages’ of the project: the first three months (Stage 1 – Fall 2012), the middle (Stage 2 – Fall 2013), and the end (Stage 3 – Fall 2014) of analysis. To define our reanalysis populations for each analytic stage, we arbitrarily selected 4, 1m² units, representing between 200 and 380 L of site matrix each (Table 1). Based on NISP estimates from our original analysis, we reasoned that we needed to reanalyze remains from four or five 10 L, ‘C’ buckets from each stage to provide sufficient sample sizes (in terms of NISP and volume) for statistical comparisons. The sample of buckets for reanalysis (Table 1) was chosen at random from all buckets for each analytic stage, except that we excluded all buckets that produced less than 30 NISP in the original analysis from selection to further ensure we would reanalyze sufficiently large samples. This resulted in reanalysis of fish remains from 14, 10 L buckets (Table 1). Our choice to focus on ‘C’ buckets was intended to ensure the relative frequencies of small fish remains in the reanalysis sample are representative of the original analysis. For example, fish remains in the complete analysis of each analytic stage were overwhelmingly recovered in fine mesh screens, with over 19,000 NSP (94.8% NSP) recovered in 1/8” screens, the finest mesh size used during excavation, alone (Table 2).

Once the buckets were selected for reanalysis, Nims emptied fish remains from each context onto a tray and carried out analysis using the same protocols as the original study, noting taxon, element, presence of burning, etc. To ensure the original identifications would not bias results, and effectively create a blind test, Nims replaced the identification labels in the original sub-bags with unique letter labels and then waited at least three days before reanalyzing them. Following earlier protocols, Butler evaluated and adjusted all identifications. As specimens were recorded, Nims compared the original identification labels to the new identifications and recorded the discrepancies. In sum, Nims spent 96 hours carrying out the reanalysis, and Butler spent approximately eight hours verifying identifications.

All reanalyzed specimens were curated with their original identifications, despite any such discrepancies, since our goal was not to fix specific mistakes, but to evaluate the reliability of our analysis. Because we did not catalog each specimen individually we cannot compare the identifications that were attributed to any particular specimen. However, we did closely compare analytic results for a given bucket and specify the likely source of the discrepancy, using bag labels that record the original attributions.

Table 1 – Summary information of Tse-whit-zen buckets and fish remains selected for reanalysis by analytic stage. Fish remains were recovered from nested screens to minimum size of 1/8" mesh.

Complete Analysis					Reanalysis Sample			
	NISP	NSP	Vol. (L)		NISP	NSP	Vol. (L)	Bucket Catalog Numbers
Stage 1 – Area A4 (Fall 2012)								
Unit	17	811	2,018	80	106	253	10	WS-10039
	18	1,112	3,143	80	167	477	10	WS-10994
	19	481	1,413	50	56	277	10	WS-13621
	20	483	1,081	60	201	398	10	WS-18082
Subtotal	2,887	7,655	270		530	1,405	40	
Stage 2 - Area A4 (Fall 2013)								
Unit	14	419	754	50	372	616	10	WS-8034
	15	1,359	2,385	90	1,332	1,939	20	WS-9146; WS-9475
	16	3,707	5,564	110	0	0	0	(no bucket selected)
	21	1,164	1,980	130	142	298	20	WS-8097; WS-10468
Subtotal	6,649	10,683	380		1,846	2,853	50	
Stage 3 - Area A1 (Fall 2014)								
Unit	4	649	1,093	40	510	851	20	WS-691; WS-6903
	5	184	356	60	112	210	20	WS-549; WS-7897
	7	182	399	50	0	0	0	(no bucket selected)
	8	173	357	50	146	302	10	WS-189
Subtotal	1,188	2,205	200		768	1,363	50	
Grand Total	10,724	20,543	850		3,144	5,621	140	

Table 2 – Summary frequencies of fish specimens (NSP) recovered in each screen size fraction of "C" buckets from Tse-whit-zen in the original analysis of each analytic stage and reanalysis sample.

	Complete Analysis				Reanalysis Sample			
	1/2"	1/4"	1/8"	Total	1/2"	1/4"	1/8"	Total
Stage 1	3	514	7,138	7,655	1	108	1,296	1,405
Stage 2	7	397	10,279	10,683	6	103	2,744	2,853
Stage 3	0	156	2,049	2,205	0	52	1,311	1,363
Grand Total	10	1,067	19,466	20,543	7	263	5,351	5,621
% of Total	<0.1	5.2	94.8		0.1	4.7	95.2	

2.3 Comparing Results from Original Analysis and Reanalysis

We assessed reliability in overall fishbone counts, fragmentation measures, taxonomic identifications for family and finer taxa, in element identification for one taxon (flatfishes [Pleuronectiformes]), and in burning frequencies, using ordinal and interval scale approaches. We used total NSP to compare fishbone counts in the reanalyzed sample, and NISP to compare the sample size of

each analytic stage in the original analysis and the reanalysis. To compare measures of post-depositional destruction, we calculated a commonly used index of fragmentation intensity (NSP:NISP) (e.g., Grayson, 1991; Nagaoka, 2005) for each analytic stage.

To explore the reliability of identifications for individual specimens, we calculated the proportion of overall agreement, a simple agreement index that has been widely used in reliability studies of psychological diagnoses. First, we compared the taxonomic attributions made for specimens in each bag lot during the original analysis and reanalysis, and then summarized the results of this comparison as a $C \times C$ matrix (Uebersax, 2014), where C is equal to the total number of taxonomic groups we included in our taxonomic universe, including non-salmonid and unidentified fish specimens (Table 3). Here, n_{ij} represents the NSP assigned to group i in the reanalysis, and to group j in the original analysis, with i and $j = 1, 2, \dots, C$; and N is equal to the total NSP reanalyzed. Wherever a “.” appears in the subscript, it represents the sum of all specimens assigned to the corresponding taxonomic group i or j (e.g., $n_{1.}$ is the total NSP assigned to taxon 1 in the reanalysis). All values along the main diagonal (shaded grey) represent specimens that were assigned to the same taxonomic group in the original analysis and the reanalysis, while cases above and below the diagonal indicate disagreement.

Table 3 – Generalized model for comparison of identifications made during the original analysis and the reanalysis of any number of taxonomic groups (adapted from Uebersax, 2014).

		Original Analysis					Total
		1	2	3	...	C	
Reanalysis	1	n_{11}	n_{12}	n_{13}	...	n_{1C}	$n_{1.}$
	2	n_{21}	n_{22}	n_{23}	...	n_{2C}	$n_{2.}$
	3	n_{31}	n_{32}	n_{33}	...	n_{3C}	$n_{3.}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	C	n_{C1}	n_{C2}	n_{C3}	...	n_{CC}	$n_{C.}$
Total		$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.C}$	N

The proportion of overall agreement (p_o) quantifies the amount of agreement that we observed between the two studies, and is equal to the NSP in the main diagonal divided by the total NSP, or

$$p_o = \frac{\sum n_{ii}}{N}$$

where n_{ii} is equal to the NSP attributed to a given group by both the original analysis and the reanalysis (Uebersax, 2014). This ratio represents the proportion of specimens that were recorded as the same taxon in both analyses, with values close to 1 indicating close agreement, and values close to 0 indicating minimal instances of agreement.

We used Spearman's rank order correlation coefficient (r_s) to test whether the rank order taxonomic abundances of each taxon are statistically correlated at the family level, and at the finest identified levels, in the original analysis and reanalysis. Because it is possible to observe nonrandom correlations that are extremely weak, particularly when a large number of cases are compared (Wolverton et al., 2016), we also calculated the *effect size* for Spearman's correlation test. While Spearman's correlation test determines whether there is a statistically significant correlation between variables that is not the product of chance, measures of effect size help determine whether such relationships are *practically* significant. For Spearman's correlation test, effect size is measured with the coefficient of determination (r_s^2), which signifies what amount of the variation in one variable can be explained by the other. A coefficient equal to or less than 10% indicates a trivial relationship exists between the variables, 25% is moderate, and coefficients of 50% or greater demonstrate a strong relationship (Wolverton et al., 2016).

To examine how relative abundance changed for each identified taxon between the two analyses, we calculated % NISP for each taxon in the original analysis and the reanalysis, and the absolute difference in % NISP between the two taxa. The results estimate the uncertainty of our identifications for each taxon, and can be used in future intra-site analyses as a minimum threshold of difference that must be exceeded to rule out the possibility of analytic error.

Skeletal element identifications are an important source of information about body part representation, which zooarchaeologists use to explore butchery and transport decisions, and to evaluate the intensity of post-depositional destruction (for fishes, see Butler and Chatters, 1994; Çakırlar et al., 2014; Partlow and Kopperl, 2011; Smith et al., 2011). We assessed reliability in skeletal element identifications by focusing on flatfishes (Pleuronectiformes), which had a sufficient sample size and range of element types represented in the original analysis for study. In particular, we compared element representation in the original and reanalysis studies for each analytic stage, and use the results to suggest sources of error.

Zooarchaeologists commonly document thermally altered bone using visual criteria – especially color – on the bone surface to study food preparation, human disposal practices, or overall taphonomic history, and burning was recorded for Tse-whit-zen fish remains. To assess whether criteria for identifying thermal alterations were applied consistently, we compared the proportions of burned specimens to the total NSP of fish remains in the original analysis and the reanalysis of each analytic stage. We also used the z-test as if each analysis was an independent sample to determine if proportions of burned specimens in the reanalysis results are statistically different from the original results, and

measured effect size as the absolute difference between the proportions.

3. RESULTS

The original and reanalysis results show close agreement overall, with some minor differences in the numbers of recorded specimens. In the original analysis, a total of 5,621 NSP was recorded from the 14, 10-L buckets selected for reanalysis (Table 1, 4). During reanalysis, we recorded 5,515 NSP, with a net decrease of 97 fish specimens in Stage 1, a net decrease of 10 NSP in Stage 2, and a net increase of 1 NSP in Stage 3 (Table 4). Besides possible counting and recording errors, the specimens that appear to be missing from Stages 1 and 2 in the reanalysis were likely assigned to “indeterminate animal type” and not recorded, per our protocol. These 107 unrecorded specimens in the reanalysis were almost certainly recorded as unidentified fish in the original analysis, as the count of unidentified fish specimens decreased in these two analytic stages, while the numbers of non-salmonid and identified fish specimens increased. The one specimen in Stage 3 that was added in the reanalysis could reflect a simple counting error, or it may indicate that a specimen originally assigned to “indeterminate animal” was subsequently recorded as unidentified fish. Remarkably, the vast majority of disagreements in fishbone frequencies for each analytic stage occurred in the materials that are between 1/4” and 1/8” in size (Table 5). There are only four specimens from Analytic Stage 1 unaccounted for in the 1/4” sample, while all other discrepancies occurred in the 1/8” sample. These differences in NISP and total NSP are notable, but they have minimal effects on our measures of the sample size and fragmentation among fish remains. Close agreement between the total numbers of *identified* specimens recorded originally and in the reanalysis suggests there was little change in our protocols that would affect sample size as measured by total NISP. The increase of 18 NISP in the reanalysis of Stage 1 represents an increase of 3.3% NISP, and the increase of 70 NISP for Stage 2 represents a similarly small increase of 3.7% NISP (Table 4). In the reanalysis of Stage 3, NISP decreased by only five specimens, or 0.7% NISP. Furthermore, fragmentation intensity (NSP:NISP) shows little change as a result of differences in the number of specimens for each analytic stage. Stage 1 exhibits the greatest amount of fishbone destruction in both the original analysis (2.7 NSP:NISP) and reanalysis (2.4 NSP:NISP), with very little change between the two analyses. Meanwhile, Stages 2 and 3 show only minor difference in fragmentation intensity at all, with ratios of 1.5 NSP:NISP and 1.7 NSP:NISP respectively in each analysis. Therefore, any differences in fishbone abundance or fragmentation intensity that might be observed in comparisons within the Tse-whit-zen fishbone assemblage cannot be explained by inconsistencies in our laboratory practices.

Table 4 – Summary frequencies of fish remains recorded in the original analysis and reanalysis of the Tse-whit-zen fishbone assemblage.

Analytic Stage	Group	Original Count	%NSP	Reanalysis Count	%NSP	Difference NSP %NSP*	
Stage 1	Total NISP	530	37.7	548	41.9	+18	3.3
	Non-salmonid	96	6.8	121	9.3	+25	23.0
	Unidentified	779	55.4	639	48.9	-140	19.7
	Total	1,405	100.0	1,308	100.0	-97	7.2
Stage 2	Total NISP	1,846	64.7	1,916	67.4	+70	3.7
	Non-salmonid	165	5.8	176	6.2	+11	6.5
	Unidentified	842	29.5	751	26.4	-91	11.4
	Total	2,853	100.0	2,843	100.0	-10	0.4
Stage 3	Total NISP	768	56.3	763	55.9	-5	0.7
	Non-salmonid	122	9.0	122	8.9	0	0.0
	Unidentified	473	34.7	479	35.1	+6	1.3
	Total	1,363	100.0	1,364	100.0	+1	0.1
Overall	Total NISP	3,144	55.9	3,227	58.5	+83	2.6
	Non-salmonid	383	6.8	419	7.6	+36	9.0
	Unidentified	2,094	37.3	1,869	33.9	-225	11.4
	Grand Total	5,621	100.0	5,515	100.0	-106	1.9

*Calculated as %Difference = $\frac{|NSP_2 - NSP_1|}{\overline{NSP}} \times 100\%$, where NSP_1 is the Original Count; NSP_2 is the Reanalysis Count for a given analytic stage, and $\overline{NSP}$ is the average of NSP_1 and NSP_2 .

Table 5 – Summary frequencies of fish specimens (NSP) recorded in the original analysis and reanalysis of the Tse-whit-zen fishbone assemblage by mesh size.

Analytic Stage	Screen Size	Original Count	Reanalysis Count	Diff.
Stage 1	1/2"	1	1	0
	1/4"	108	104	-4
	1/8"	1296	1203	-93
	Total	1405	1308	-97
Stage 2	1/2"	6	6	0
	1/4"	103	103	0
	1/8"	2744	2734	-10
	Total	2853	2843	-10
Stage 3	1/2"	0	0	0
	1/4"	52	52	0
	1/8"	1311	1312	+1
	Total	1363	1364	+1
Grand Total		5621	5515	-106

Agreement indices for each analytic stage demonstrate that the vast majority of reanalyzed specimens were attributed to the same taxon as in the original analysis, which we illustrate with reanalysis results from Stage 2. The main diagonal of Table 6 – highlighted in grey – shows the number of specimens that we attributed to the same group in the original analysis and the reanalysis. The number of specimens on the diagonal far out-weigh the number of specimens for which our taxonomic attributions were not in agreement (above and below the diagonal, Table 6), as demonstrated by the proportion of overall agreement: $p_o = 0.95$. Similar tables for Stage 1 (Supplementary Table 2) and Stage 3 (Supplementary Table 3) materials exhibit similarly high levels of agreement, with p_o 's of 0.91 and 0.98, respectively. In all three analytic stages, most disagreements between identifications made in the original analysis and the reanalysis are related to whether particular fish specimens are counted as an identifiable specimen, or a non-salmonid or unidentified fish specimen. In the example of Stage 2, a total of 86 specimens that were originally attributed to non-salmonid or unidentified fish were later identified at the family level or finer during reanalysis, while 11 specimens originally considered identifiable were assigned to non-salmonid or unidentified fish during reanalysis (Table 6). However, the low number of these specimens, relative to the 2,853 NSP reanalyzed from Stage 2, observed alongside high overall agreement in each analytic stage, suggests we applied our laboratory protocols consistently throughout the original analysis.

Comparisons of family-level taxonomic representation show that minor differences in the results of the original analysis and the reanalysis do not affect rank order abundance in any of the three analytic stages (Table 7). In Stage 1, the first five most abundant of the 11 represented taxa show identical rank order abundance in each analysis, and the differences in the remaining six taxa are minor (Table 7). The correlation between the original and reanalysis results is statistically significant ($r_s = 0.980$; $n = 11$; $p < 0.001$) and extremely high ($r_s^2 = 96\%$). In Stage 2, there is slightly less agreement between the original and reanalysis results, with minor differences in rank order abundance apparent in seven out of the 13 represented taxa (Table 7). However, there is still a statistically significant ($r_s = 0.973$; $n = 13$; $p < 0.001$), and extremely high ($r_s^2 = 95\%$), correlation between taxonomic representation in each analysis. Finally, Stage 3 shows only two taxa have minor differences in rank order abundance between the original analysis and the reanalysis (Table 7), and the results have a statistically significant ($r_s = 0.994$; $n = 13$; $p < 0.001$), near-perfect ($r_s^2 = 99\%$) correlation. Overall, these patterns suggest laboratory practices had minimal effect on family-level taxonomic representation.

Table 6 – Agreement matrix showing pair-wise family level identifications made for individual specimens in the original analysis (columns) and reanalysis (rows) of Stage 2 materials. Highlighted cells represent specimens identified to the same family in both analyses.

		Original Taxonomic Family Attribution																
		Squ.	Raj.	Ela.*	Chi.	Clu.**	Sal.	Gad.	Emb.	Sco.	Ano.	Hex.	Cot.	Plu.†	N.S.	Unid	Total	
Reanalysis Taxonomic Family Attribution	Squ.	30		1													31	
	Raj.		7														7	
	Ela.	4	1	3											1	1	10	
	Chi.				4												4	
	Clu.					1,535									1	45	1,581	
	Sal.						18								1	2	21	
	Gad.						1	44							7	1	53	
	Emb.								7						1		8	
	Sco.									1						2	3	
	Ano.										56				1		57	
	Hex.											2		1	2	1	6	
	Cot.									1			49		10	4	64	
	Plu.						2				1				62	3	3	71
	N.S.								1					2	3	136	32	174
	Unid				1		1		3							3	732	740
Total		34	8	5	4	1,539	18	48	7	1	58	2	51	66	166	823	2,830	
Reanalysis specimens originally identified as non-fish											13	Total in Reanalysis				2,843		
Specimens from original analysis missing from reanalysis											10	Total in Original Analysis				2,853		

Squ. = Squalidae; Raj. = Rajidae; Ela. = Elasmobranchii; Chi. = Chimaeridae; Clp. = Clupeiformes; Sal. = Salmonidae; Gad. = Gadidae; Emb. = Embiotocidae; Sco. = Scorpaenidae; Ano. = Anoplopomatidae; Hex. = Hexagrammidae; Cot. = Cottidae; Plu. = Pleuronectiformes; N.S. = Non-salmonid; Unid = Unidentified Fish.

* Sub-class Elasmobranchii is included because some fragmentary dogfish shark (family Squalidae) and ray (family Rajidae) vertebrae cannot be definitively identified to either family.

**Herrings, anchovies, and sardines are reported at the order level (Clupeiformes) because the elements of the otic series could not be identified to the family level with confidence. Given their numerical dominance, Clupeiformes specimens likely derive from the Pacific herring (*Clupea pallasii*).

† Flatfishes are reported at the order level (Pleuronectiformes) because skeletal elements from member families Paralichthyidae and Pleuronectidae cannot always be distinguished with confidence.

Table 7 – Rank order comparison of family taxonomic abundance (NISP) between the original analysis and reanalysis of a sample of the Tse-whit-zen fishbone assemblage by analytic stage.

Family	Stage 1				Stage 2				Stage 3			
	Original NISP (%)	Rank	Reanalysis NISP (%)	Rank	Original NISP (%)	Rank	Reanalysis NISP (%)	Rank	Original NISP (%)	Rank	Reanalysis NISP (%)	Rank
Squalidae	16 (3.0)	6	20 (3.6)	7	34 (1.8)	6	31 (1.6)	6	9 (1.1)	7	9 (1.1)	7
Rajidae	2 (0.3)	10	2 (0.3)	11	8 (0.4)	9	7 (0.3)	10	4 (0.5)	9	4 (0.5)	9
Elasmobranchii*	6 (1.1)	8	6 (1.0)	8.5	5 (0.2)	10	10 (0.5)	8	1 (0.1)	13	1 (0.1)	13
Chimaeridae	-		-		4 (0.2)	11	4 (0.2)	12	6 (0.7)	8	6 (0.7)	8
Clupeiformes**	282 (53.2)	1	282 (51.4)	1	1,540 (83.4)	1	1,581 (82.5)	1	437 (56.9)	1	433 (56.7)	1
Salmonidae	33 (6.2)	4	42 (7.6)	4	19 (1.0)	7	21 (1.0)	7	27 (3.5)	5	22 (2.8)	6
Gadidae	32 (6.0)	5	41 (7.4)	5	48 (2.6)	5	53 (2.7)	5	82 (10.6)	3	80 (10.4)	3
Embiotocidae	4 (0.7)	9	6 (1.0)	8.5	9 (0.4)	8	8 (0.4)	9	3 (0.3)	10.5	3 (0.3)	10.5
Scorpaenidae	-		-		1 (<0.1)	13	3 (0.1)	13	2 (0.2)	12	2 (0.2)	12
Anoplopomatidae	15 (2.8)	7	23 (4.1)	6	58 (3.1)	3	57 (2.9)	4	21 (2.7)	6	23 (3.0)	5
Hexagrammidae	1 (0.1)	11	3 (0.5)	10	2 (0.1)	12	6 (0.3)	11	3 (0.3)	10.5	3 (0.3)	10.5
Cottidae	76 (14.3)	2	65 (11.8)	2	51 (2.7)	4	64 (3.3)	3	74 (9.6)	4	72 (9.4)	4
Pleuronectiformes†	63 (11.8)	3	58 (10.5)	3	67 (3.6)	2	71 (3.7)	2	99 (12.8)	2	105 (13.7)	2
Total NISP	530 (100.0)		548 (100.0)		1,846 (100.0)		1,916 (100.0)		768 (100.0)		763 (100.0)	

* Sub-class Elasmobranchii is included because some fragmentary dogfish shark (family Squalidae) and ray (family Rajidae) vertebrae cannot be definitively identified to either family.

**Herrings, anchovies, and sardines are reported at the order level (Clupeiformes) because the elements of the otic series could not be identified to the family level with confidence. Given their numerical dominance, Clupeiformes specimens likely derive from the Pacific herring (*Clupea pallasii*).

† Flatfishes are reported at the order level (Pleuronectiformes) because skeletal elements from member families Paralichthyidae and Pleuronectidae cannot always be distinguished with confidence.

Turning to taxonomic representation at the finest identified levels, comparisons of the original and reanalysis results again show striking similarities (Table 8). Most taxa in Stage 1 show very similar rank order abundances in the results of each analysis (Table 8). Only the taxon Cottidae shows a substantial deviation between the two analyses, with a change from 20 NISP (7th most abundant taxon) to 1 NISP (tied for 21st most abundant taxon). This decrease reflects an increase in our confidence to identify sculpin vertebrae to finer taxonomic categories between the beginning and end of the Tse-whit-zen project. Other notable differences in rank order for great sculpin (*Myoxocephalus polyacanthocephalus*), which were previously recorded at the family level, and Pacific cod (*Gadus macrocephalus*) result from the addition of only 6-7 NISP. The overall appearance of similarity is borne out statistically, as the original and reanalysis results are significantly ($r_s = 0.834$; $n = 21$; $p < 0.001$) and highly ($r_s^2 = 70\%$) correlated. The results for Stage 2 are more highly correlated, with 11 out of the 22 taxa represented in both the original analysis and the reanalysis showing identical rank order abundance (Table 8). This correlation is statistically ($r_s = 0.869$; $n = 22$; $p < 0.001$) and practically ($r_s^2 = 76\%$) significant. Notable shifts did occur for some sculpin taxa, and for Elasmobranchii, but they result from only minor changes in NISP that are in turn related to reassignments within these taxonomic groups. Finally, in Stage 3, only the Irish lords (*Hemilepidotus* sp.) show a substantial change in rank order abundance, which results from a difference of 2 NISP (Table 8). Fewer taxa show identical abundances than in Stage 2, but the amount of difference between ranks is much smaller overall, resulting in a significant ($r_s = 0.972$; $n = 23$; $p < 0.001$) and extremely strong ($r_s^2 = 94\%$) correlation.

Regarding proportional representation, differences in taxonomic abundance are minor overall (Figure 2, Table 8). In Stage 1, there is a change of 3.6% NISP for Cottidae, while *Clupea pallasii*, Pleuronectiformes, *Leptocottus armatus*, *Oncorhynchus* sp., *Anoplopoma fimbria*, *Myoxocephalus polyacanthocephalus*, and *Gadus macrocephalus* all show a change in % NISP between 2% and 1% (Figure 2). All remaining taxa in Stage 1, and all taxa in the other two analytic stages, have an absolute difference in % NISP less than 1%. In sum, we observed some inconsistencies in identification protocols for several sculpin taxa, but the rank order correlations in each analytic stage are strong enough, and the differences in proportional representation are small enough, to rule out the possibility that protocol drift explains patterns of taxonomic representation at the finest identified levels.

400 **Table 8** - Comparison of rank order abundance (NISP) for finest taxonomic level between the original analysis and reanalysis of a sample of the
 401 Tse-whit-zen fishbone assemblage. Cottidae and finer taxa of Cottidae are highlighted and discussed in text.
 402

Finest Taxon	Stage 1						Stage 2						Stage 3					
	Original NISP (%)		Rank	Reanalysis NISP (%)		Rank	Original NISP (%)		Rank	Reanalysis NISP (%)		Rank	Original NISP (%)		Rank	Reanalysis NISP (%)		Rank
<i>Squalus acanthias</i>	16	(3.0)	8	20	(3.6)	8	34	(1.8)	5	31	(1.6)	5	9	(1.1)	10	9	(1.1)	10.5
<i>Raja</i> sp.	2	(0.3)	16.5	2	(0.3)	18	8	(0.4)	12.5	7	(0.3)	15	4	(0.5)	13	4	(0.5)	13.5
<i>Elasmobranchii</i>	6	(1.1)	12	6	(1.0)	14.5	5	(0.2)	15.5	10	(0.5)	11.5	1	(0.1)	21.5	1	(0.1)	22.5
<i>Hydrolagus colliei</i>	-			-			4	(0.2)	18	4	(0.2)	18	6	(0.7)	12	6	(0.7)	12
<i>Clupea pallasii</i>	196	(36.9)	1	197	(35.9)	1	1,416	(76.7)	1	1,461	(76.2)	1	262	(34.1)	1	258	(33.8)	1
<i>Clupeiformes</i>	86	(16.2)	2	85	(15.5)	2	124	(6.7)	2	120	(6.2)	2	175	(22.7)	2	175	(22.9)	2
<i>Oncorhynchus</i> sp.	33	(6.2)	5	42	(7.6)	4	19	(1.0)	8	21	(1.0)	8	27	(3.5)	6	22	(2.8)	7
<i>Microgadus proximus</i>	22	(4.1)	6	22	(4.0)	7	29	(1.5)	6	29	(1.5)	6	59	(7.6)	4	57	(7.4)	4
<i>Gadus macrocephalus</i>	2	(0.3)	16.5	8	(1.4)	12	5	(0.2)	15.5	4	(0.2)	18	7	(0.9)	11	9	(1.1)	10.5
<i>Gadidae</i>	8	(1.5)	11	11	(2.0)	10	14	(0.7)	9	20	(1.0)	9	16	(2.0)	9	14	(1.8)	9
<i>Embiotocidae</i>	4	(0.7)	14	6	(1.0)	14.5	9	(0.4)	11	8	(0.4)	14	3	(0.3)	15.5	3	(0.3)	15.5
<i>Sebastes</i> sp.	-			-			1	(<0.1)	22.5	3	(0.1)	20.5	2	(0.2)	18.5	2	(0.2)	18.5
<i>Anoplopoma fimbria</i>	15	(2.8)	9	23	(4.1)	6	58	(3.1)	4	57	(2.9)	4	21	(2.7)	7	23	(3.0)	6
<i>Ophiodon elongatus</i>	-			1	(0.1)	21	1	(<0.1)	22.5	1	(<0.1)	22.5	1	(0.1)	21.5	1	(0.1)	22.5
<i>Hexagrammos</i> sp.	-			1	(0.1)	21	-			5	(0.2)	16	2	(0.2)	18.5	2	(0.2)	18.5
<i>Hexagrammidae</i>	1	(0.1)	20	1	(0.1)	21	1	(<0.1)	22.5	-			-			-		
<i>Leptocottus armatus</i>	36	(6.7)	4	32	(5.8)	5	21	(1.1)	7	26	(1.3)	7	44	(5.7)	5	48	(6.2)	5
<i>Enophris bison</i>	12	(2.2)	10	15	(2.7)	9	7	(0.3)	14	16	(0.8)	10	20	(2.6)	8	15	(1.9)	8
<i>Myoxocephalus poly.</i>	2	(0.3)	16.5	9	(1.6)	11	4	(0.2)	18	9	(0.4)	13	3	(0.3)	15.5	4	(0.5)	13.5
<i>Hemilepidotus</i> sp.	1	(0.1)	20	1	(0.1)	21	8	(0.4)	12.5	3	(0.1)	20.5	3	(0.3)	15.5	1	(0.1)	22.5
<i>Hemi./Scorp.*</i>	5	(0.9)	13	7	(1.2)	13	1	(<0.1)	22.5	10	(0.5)	11.5	3	(0.3)	15.5	3	(0.3)	15.5
<i>Cottidae</i>	20	(3.7)	7	1	(0.1)	21	10	(0.5)	10	-			1	(0.1)	21.5	1	(0.1)	22.5
<i>Pleuronectidae</i>	1	(0.1)	20	3	(0.5)	16.5	2	(0.1)	20	1	(<0.1)	22.5	1	(0.1)	21.5	2	(0.2)	18.5
<i>Citharichthys</i> sp.	2	(0.3)	16.5	3	(0.5)	16.5	4	(0.2)	18	4	(0.2)	18	-			2	(0.2)	18.5
<i>Pleuronectiformes</i>	60	(11.3)	3	52	(9.4)	3	61	(3.3)	3	66	(3.4)	3	98	(12.7)	3	101	(13.2)	3
Total	530	(100.0)		548	(100.0)		1,846	(100.0)		1,916	(100)		768	(100)		763	(100)	

403 *Recorded as a joint category given the close similarity of *Hemilepidotus* sp. and *Scorpaenichthys marmoratus* vertebrae.

Figure 2 – Comparison of abundance (% NISP) for finest taxon assignments, in original and reanalysis of Tse-whit-zen fish remains by analytic stage. Taxa listed in order of abundance in the original analysis.

Note: *Clupea pallasii* and Clupeiformes are excluded from this figure as their high abundance (together representing between 53% and 83% NISP of the original sample in each analytic stage), overwhelm the patterns for less abundant taxa. For stage 1, 2, and 3, respectively, disagreement was 0.6%, 0.2%, and 0.0% for *Clupea pallasii*; and 0.0%, 0.1%, and 0.0%, for Clupeiformes.

Squ. = *Squalus acanthias*; Raj. = *Raja* sp.; ELA = Elasmobranchii; Hyd. = *Hydrolagus coliei*; Clu. = *Clupea pallasii*; CLU = Clupeiformes; Onc. = *Oncorhynchus* sp.; Gad. = *Gadus macrocephalus*; Mic. = *Microgadus proximus*; GAD = Gadidae; EMB = Embiotocidae; Seb. = *Sebastes* sp.; Ano. = *Anoplopoma fimbria*; Oph. = *Ophiodon elongatus*; Hex. = *Hexagrammos* sp.; HEX = Hexagrammidae; Lep. = *Leptocottus armatus*; Eno. = *Enophrys bison*; Myo. = *Myoxocephalus polyacanthocephalus*; Hem. = *Hemilepidotus* sp.; H./S. = *Hemilepidotus/Scorpaenichthys*; COT = Cottidae; PLE = Pleuronectidae; Cit. = *Citharichthys* sp.; PLU = Pleuronectiformes

Element frequencies for Pleuronectiformes in the reanalysis match the original frequencies almost perfectly in each analytic stage. In the reanalysis of Stage 1 materials, there was a net decrease of five specimens from the original 63 NISP as some elements originally considered identifiable (1 posttemporal, 6 vertebrae) were recorded as unidentified in the reanalysis (Table 9). In addition, one vertebra originally recorded as flatfish was attributed to tom cod (*Microgadus proximus*) during reanalysis, and several other elements (1 hyomandibula, 1 1st vertebra, and 1 scale) were identified as flatfish from specimens previously recorded as unidentified. In Stage 2, there was a net increase of four specimens from the 64 NISP originally recorded (Table 9). The additional premaxilla, angular/articular, and cleithrum in the reanalysis were originally recorded as unidentified fish. There was also a net increase of one indeterminate vertebra after four flatfish vertebrae were re-recorded as non-salmonid, one flatfish vertebra was re-recorded as greenling (*Hexagrammos* sp.), and one sablefish (*Anoplopoma fimbria*) and two Pacific herring (*Clupea pallasii*) vertebrae were re-recorded as Pleuronectiformes. In Stage 3, the reanalysis sample increased by six specimens from the original 99 NISP after one posttemporal was added from unidentified fish, one vertebral fragment and three indeterminate vertebrae were added from non-salmonid, and one indeterminate vertebra originally recorded as tom cod was re-recorded as flatfish. Discrepancies in the frequencies for vomer and basioccipital in Stage 3 are also striking, where three specimens originally attributed to vomer were identified as basioccipital in the reanalysis (Table 9). Given that these elements show substantial morphological differences, we believe these discrepancies reflect recording errors, which are certainly a source of error in any analysis.

Overall, these minor differences in element frequencies have little effect on body part representation. When element frequencies are totaled for the cranium, the paired fins, and the vertebral column, we find no differences in ordinal scale representation of each segment, and very little change in proportional representation (Table 9, Figure 3). In both the original and reanalysis results, Pleuronectiformes is represented overwhelmingly by vertebrae and a small number of cranial elements in all analytic stages, while elements of the paired fins are rarely present. Therefore, any differences in Pleuronectiformes body part representation that might be observed in intra-site comparisons with the Tse-whit-zen fishbone assemblage cannot be explained by inconsistent laboratory protocols.

Table 9 – Body-part and element representation (NISP and MNE) for Pleuronectiformes in the original analysis and reanalysis of Tse-whit-zen fish remains by analytic stage.

Body Part	Element	Stage 1				Stage 2				Stage 3			
		Original		Reanalysis		Original		Reanalysis		Original		Reanalysis	
		NISP	MNE	NISP	MNE	NISP	MNE	NISP	MNE	NISP	MNE	NISP	MNE
Cranial	Premaxilla							1	1	2	2	2	2
	Maxilla					1	1	1	1				
	Dentary									1	1	1	1
	Angular/Articular							1	1	1	1	1	1
	Quadrate	1	1	1	1					2	2	2	2
	Hyomandibula	2	1	3	3					3	3	3	3
	Vomer	1	1	1	1					4	4	1	1
	Basioccipital	2	2	2	2	1	1	1	1			3	3
	Urohyal					1	1	1	1	1	1	1	1
	Interhaemal												
	Spine	1		1	1								
	Subtotal	7	5	8	8	3	3	5	5	14	14	14	14
Paired Fins	Posttemporal	1	1							3	3	4	4
	Cleithrum							1					
	Subtotal	1	1	0	0	0	0	1	0	3	3	4	4
Vertebrae	1st Vertebra	2	2	3	3	2	2	2	2	4	4	4	4
	Vertebra Fragment									1		2	
	Indet. Vertebra	53	53	46	46	62	62	63	63	77	77	81	81
	Subtotal	55	55	49	49	64	64	65	65	82	81	87	85
Other	Scale*			1	1								
	Grand Total	63	61	58	58	67	67	71	70	99	98	105	103

*Recorded only for starry flounder (*Platichthys stellatus*), which has distinctive, spiny scales.

Comparing frequencies of burned specimens shows that significantly more burned specimens were identified during reanalysis than in the original analysis for Stage 1, while discrepancies for Stage 2 and 3 were minimal (Table 10), suggesting that our criteria for identifying burning shifted over time. In particular, 8.7% more specimens were identified as burned in Stage 1 ($p < 0.001$), while the differences of less than 1% in Stages 2 ($p = 0.407$) and 3 ($p = 0.881$) were not significant. Considering that we did not apply consistent burning criteria throughout the project, future comparisons of burning within the Tse-whit-zen fishbone assemblage will have to account for the possibility that differences in thermal alteration are related to unreliability in the data, and not human behavior or site formation patterns. We suggest that the ~9% difference in burning rates for Stage 1 specimens should be used as a crude margin of error, and that any observed differences in burning will have to exceed this value to reject the hypothesis that the differences are the result of analytic error.

Figure 3 – Proportional body-part representation (% MNE) for all elements recorded in the flatfish skeleton (with MNE values), and all Pleuronectiformes specimens in the original analysis and reanalysis of Tse-whit-zen fish remains by analytic stage (see Table 9 for all MNE counts).

Table 10 – Number and relative abundance of burned specimens in the original analysis and reanalysis of Tse-whit-zen fish remains by analytic stage.

		Stage 1	Stage 2	Stage 3
Original	Burned NSP	43	132	108
	Total NSP	1,405	2,853	1,363
	% Burned	3.1	4.6	7.9
Reanalysis	Burned NSP	164	145	106
	Total NSP	1,308	2,843	1,364
	% Burned	12.5	5.1	7.8
Difference %		9.5	0.5	-0.2
Z-score		9.21	0.83	-0.15
p value		<0.001*	0.407	0.881

*Difference is significant at the 0.05 level

4. DISCUSSION AND CONCLUSION

As demonstrated in previous studies, the *a priori* assumption that any two zooarchaeologists will produce comparable results is not defensible (Fisher, 2015; Gobalet, 2001), even when they follow

the same detailed identification protocols (Twiss et al., 2016). Similarly, we argue the methods and protocols used by an individual or research team to generate data over the course of a multi-year project have the potential to ‘drift,’ producing results from different stages of the project that are not comparable. In light of these observations, and the subjective nature of faunal identification, zooarchaeologists should assess the reliability of their data before presuming that observed trends are related to past human activity, environmental change, or post-depositional processes.

In this study, we outline several simple QC methods for assessing the reliability of Tse-whit-zen fish faunal data, which involved blind reanalysis of samples that were originally documented over three stages of the project. This research design allowed us to study overall reliability in the Tse-whit-zen fishbone data, and to assess whether any evidence of ‘protocol drift,’ (or, more simply, ‘learning curve’) exists. We introduced a simple agreement index that is widely applied in psychology, and used it to quantify agreement in faunal identifications, which proved a valuable tool for isolating particular sources of disagreement. This method showed that some discrepancies in NSP counts are related to disagreements about whether fish remains belong to unidentified fish or indeterminate vertebrate. A highly disproportionate number of these disagreements were found in the 1/8” mesh fishbone samples, which is unsurprising given that relatively small remains recovered in fine mesh have smaller surface area, which is key for distinguishing vertebrate type by texture and morphology. However, because there are few such discrepancies relative to the sample sizes in each analytic stage, these disagreements did not affect the utility of NSP-based fragmentation indices or statistical comparisons of NSP counts.

On the other hand, our research highlights two areas of inconsistency that will need to be considered in future studies of this assemblage. While our identifications within the family Cottidae were reliable in terms of proportional representation, rank-order abundance was unreliable for at least one sculpin taxon in every analytic stage. This source of uncertainty has the potential to affect future research of fishing practices and local habitat change over time depending on which quantitative approaches are used to explore these patterns, highlighting the subjective nature of error and reliability. Determining reliability is contingent on the particular research questions that are being asked, and on the details of the analytic methods that are employed. To determine if the minor inconsistencies in Tse-whit-zen sculpin identifications will affect the conclusions of a given inquiry, it will be necessary to repeat any quantitative tests using the original and reanalysis results of this study to evaluate the reliability of the fishbone data in that context.

More importantly, classifying burned bone was not very reliable, especially in Stage 1, where there is a difference of over 9% NSP in the proportion of burned specimens for each analysis.

Archaeologists have long used color to identify thermal alterations to bone (Bennett, 1999; Clark and Ligouis, 2010; Lyman, 1994; Nicholson, 1993; Shipman et al., 1984; Stiner et al., 1995), though the continua of color change and surface modifications are affected by many variables (e.g., temperature, duration of exposure, animal type, proximity to fire, burial medium). Given the complexity of the relationship between burning and color, and the enormous size of the Tse-whit-zen fishbone assemblage, we tried to streamline our analysis of burning by only recording specimens with black (charred) and white (calcined) color changes, which are linked to the extreme end of the burning continuum. However, our reanalysis of Tse-whit-zen fishbone shows that even these conservative criteria were inconsistently applied. Because of our reanalysis, we are able to control for this error and study patterns of fishbone burning within the site, so long as we take the large margin of uncertainty ($\pm 9.5\%$) into account. Our results suggest that future zooarchaeology projects that use burned bone to track cultural processing and site formation should conduct quality control studies, such as blind reanalysis, to increase confidence in the analytic results.

Having reanalyzed samples of the Tse-whit-zen assemblage, we can now argue with confidence that our data are reliable, overall. The high correlations between ordinal scale taxonomic abundance at each stage of analysis show high agreement. Similarly, strong correlations between original and reanalysis results for sample size, fragmentation rates, and flatfish element representation suggest these data are adequately reliable for intra-site comparisons.

The overall reliability of our data stands in sharp contrast to the results of Gobalet's (2001) blind tests, which revealed great discrepancies between the work of four analysts who studied the same coastal California archaeological assemblage independently of one another. To account for the discrepancies in his study of inter-analyst comparability, Gobalet suggests these disagreements reflect differences in training or experience among analysts; lack of familiarity with the fish taxa local to the archaeological site; the speed with which the analysts worked; and differences in the comparative collections that analysts used. Though we acknowledge the importance of these factors, we suggest another factor that Gobalet did not mention likely contributed to the disagreements: researchers did not use the same analytic protocol, a set of rules and conventions that guides study. In zooarchaeology, such protocols establish which skeletal elements should be identified to a given taxonomic level, which morphological criteria are used to identify an element to taxon, the conventions for recording fragmentary specimens, and so forth. In our study, a single person, Butler, checked all attributions made by lab personnel over the course of the project. Thus our reliability study is different from Gobalet's project, which evaluated whether the results of several analysts working on the same assemblage "as

they normally would” (Gobalet, 2001, p. 378) are comparable. However, we assert that outlining analytic protocols at the start of the Tse-whit-zen project was a key factor contributing to the demonstrated, overall consistency in identifications for the ~5,500 specimens in our study. Building on Driver’s (2011) and Wolverton’s (2013) recommendations, we argue that increasing the transparency of analytic methods used in zooarchaeology would produce much more robust records of archaeological animal remains.

The high level of disagreement revealed in Gobalet’s study raises obvious questions about the comparability of results generated by different researchers and different laboratories that may be employing different analytic protocols. For ichthyoarchaeology in the Pacific Northwest of North America, research protocols are rarely published, making it difficult to assess the role analytic decisions play in accounting for faunal data (Nims 2016). While we are not promoting standardized laboratory practices per se, since methods have to be adapted for particular research objectives and practical considerations, research would be greatly improved by the adoption of minimum data quality standards, and – even more importantly – publication of analytic protocols, so that differences in analytic methods can be considered in inter-site comparisons of faunal records (Atici et al., 2013; Faniel et al., 2013). Furthermore, individual researchers should always test the reliability of their results to demonstrate that protocols have been applied consistently.

In sum, we strongly recommend that QA/QC should form an integral part of every zooarchaeological project research design. Our blind reanalysis of Tse-whit-zen samples shows that it is possible to address QC goals with only modest investments of time relative to overall project size. Rather than wait until a project was completed to study reliability, we recommend that blind reanalysis be incorporated into the early stages of a project, which would highlight current identification protocols that are ambiguous or insufficiently specific while there is still an opportunity to modify protocols. Blind reanalysis could also be performed by external researchers who have been provided with the project’s protocols as an additional QC check on reliability. Future studies could use quantitative methods for comparing ratings made by of any number of researchers that were developed in the statistics of diagnostic agreement (Uebersax, 2014) to compare taxonomic identifications made by multiple researchers and evaluate the reliability of zooarchaeological methods as they are practiced in a given region. Routine use of such QC practices in zooarchaeological project design and analysis, and further QA research, would make our records more robust overall and thereby contribute to their long term value in archaeology, paleobiology, and conservation biology.

ACKNOWLEDGMENTS

We thank the Lower Elwha Klallam Tribe for their on-going support. Laura Phillips (Burke Museum, University of Washington) facilitated the loan of faunal remains. Washington State Department of Transportation helped subsidize loan costs. Dennis Lewarch and Lynn Larson (formerly of LAAS, Ltd.) directed site excavations. Current Tse-whit-zen project co-PI's Kris Bovy (University of Rhode Island), Sarah Campbell (Western Washington University [WWU]), Mike Etnier (WWU), and Sarah Sterling (Portland State University [PSU]), provided leadership on the faunal-geoarchaeological study. Kristina Dick and Laura Syvertson (both PSU) provided assistance with database management and map production. Ross Smith (University of Oregon), Bob Kopperl (Willamette Cultural Resources Associates, Ltd.), and Iain McKechnie and Rebecca Wigen (both University of Victoria) loaned comparative fish skeletons and/or assisted with identifications. Former and current PSU students Tony Hofkamp, Kathryn Mohlenhoff, Emily Rocha, Shoshanna Rosenberg, and Laura Syvertson assisted in the lab. Martyn Tranter (Department of Geography, University of Bristol, UK) provided work space for Butler during manuscript preparation. Work was funded by the National Science Foundation [grant numbers 1219468 and 1353610].

WORKS CITED

- Atici, L., Kansa, S.W., Lev-Tov, J., Kansa, E.C., 2013. Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. *J. Archaeol. Method Theory* 20, 663–681. doi:10.1007/s10816-012-9132-9
- Banning, E.B., Hawkins, A.L., Stewart, S.T., Hitchings, P., Edwards, S., 2016. Quality Assurance in Archaeological Survey. *J. Archaeol. Method Theory*. doi:10.1007/s10816-016-9274-2
- Beck, C., Jones, G.T., 1989. Bias and Archaeological Classification. *Am. Antiq.* 54, 244–262.
- Bennett, J.L., 1999. Thermal Alteration of Buried Bone. *J. Archaeol. Sci.* 26, 1–8. doi:10.1006/jasc.1998.0283
- Butler, V.L., Chatters, J.C., 1994. The Role of Bone Density in Structuring Prehistoric Salmon Bone Assemblages. *J. Archaeol. Sci.* 21, 413–424.
- Çakırlar, C., Ikram, S., Gates, M.-H., 2014. New Evidence for Fish Processing in the Ancient Eastern Mediterranean: Formalised *Epinephelus* Butchery in Fifth Century bc Kinet Höyük, Turkey. *Int. J. Osteoarchaeol.* 16, 3–16. doi:10.1002/oa.2388
- Casteel, R.W., 1976. *Fish Remains in Archaeology and Paleo-Environmental Studies*. Academic Press, New York.
- Clark, J.L., Ligouis, B., 2010. Burned Bone in the Howieson's Poort and post-Howieson's Poort Middle Stone Age Deposits at Sibudu (South Africa): Behavioral and Taphonomic Implications. *J. Archaeol. Sci.* 37, 2650–2661. doi:10.1016/j.jas.2010.06.001
- Driver, J.C., 2011. Identification, Classification and Zooarchaeology. *Ethnobiol. Lett.* 2, 19–39.
- Evans, A.A., Lerner, H., Macdonald, D.A., Stemp, W.J., Anderson, P.C., 2014. Standardization, Calibration, and Innovation: A Special Issue on Lithic Microwear Method. *J. Archaeol. Sci.* 48, 1–4.

- doi:10.1016/j.jas.2014.03.002
- Ewonus, P.A., 2011. Social Zooarchaeology of a Northwest Coast House. *J. Isl. Coast. Archaeol.* 6, 72–97. doi:10.1080/15564894.2010.504806
- Faniel, I., Kansa, E., Kansa, S.W., Barrera-Gomez, J., Yakel, E., 2013. The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. *JCDL 2013 Proc. 13th ACM/IEEE-CS Jt. Conf. Digit. Libr.* 295–304.
- Fish, P.R., 1978. Consistency in Archaeological Measurement and Classification: A Pilot Study. *Am. Antiq.* 43, 86–89.
- Fisher, J.L., 2015. Faunal Quantification and the Ascendancy of Hunting Debate: Reevaluation of the Data from Southeastern California. *Am. Antiq.* 80, 767–775. doi:10.7183/0002-7316.80.4.767
- Gobalet, K.W., 2001. A Critique of Faunal Analysis: Inconsistency Among Experts in Blind Tests. *J. Archaeol. Sci.* 28, 377–386. doi:10.1006/jasc.2000.0564
- Grayson, D.K., 1991. Alpine Faunas from the White Mountains, California: Adaptive Change in the Late Prehistoric Great Basin? *J. Archaeol. Sci.* 18, 483–506. doi:10.1016/0305-4403(91)90039-R
- Grier, C., Lukowski, S., 2012. On Villages, Quantification, and Appropriate Context: A Comment on “Social Zooarchaeology of a Northwest Coast House” by Paul A. Ewonus. *J. Isl. Coast. Archaeol.* 7, 430. doi:10.1080/15564894.2012.683513
- Hart, J.L., 1973. *Pacific Fishes of Canada*. Fisheries Research Board of Canada, Volume 180, Ottawa, Ontario.
- Larson, L.L. (editor), 2006. Data Recovery Excavation and Archaeological Monitoring at the Tse-whit-zen Site (45CA523), Clallam County, Washington. Larson Anthropological/Archaeological Services Limited, Gig Harbor, WA. Submitted to Washington State Department of Transportation, Olympic Region, Tumwater, WA.
- Lawrence, B., 1973. Problems in the Inter-site Comparison of Faunal Remains, in: Matolcsi, J. (Ed.), *Domestikationsforschung Und Geschichte Der Haustiere*. Akademiai Kiado, Budapest, pp. 397–402.
- Lyman, R.L., 1994. *Vertebrate Taphonomy*. Cambridge University Press, Cambridge.
- Miller, B.S., Simenstad, C.A., Cross, J.N., Fresh, K.L., Steinfert, S.N., 1980. Nearshore Fish and Macroinvertebrate Assemblages Along the Strait of Juan de Fuca Including Food Habits of the Common Nearshore Fish. *Marine Ecosystems Analysis, Puget Sound Region, Seattle*. Submitted to Environmental Protection Agency and National Oceanic and Atmospheric Administration.
- Morin, E., Ready, E., Boileau, A., Beauval, C., Coumont, M.-P., *in press*. Problems of Identification and Quantification in Archaeozoological Analysis, Part I: Insights from a Blind Test. *J. Archaeol. Method Theory*.
- Nagaoka, L., 2005. Declining Foraging Efficiency and Moa Carcass Exploitation in Southern New Zealand. *J. Archaeol. Sci.* 32, 1328–1338. doi:10.1016/j.jas.2005.04.004
- Nicholson, R.A., 1993. A Morphological Investigation of Burnt Animal Bone and an Evaluation of its Utility in Archaeology. *J. Archaeol. Sci.* 20, 411–428.
- Partlow, M.A., Kopperl, R.E., 2011. Processing the Patterns: Elusive Archaeofaunal Signatures of Cod Storage on the North Pacific Coast, in: Moss, M.L., Cannon, A. (Eds.), *The Archaeology of North Pacific Fisheries*. University of Alaska Press, Fairbanks, pp. 195–218.
- Reetz, E.C., Lewarch, D.E., Trudel, S.E., Gillis, N., Kanipe, H.E., Sterling, S.L., Tatum, D.E., Kekkonen, S., 2006. Chapter 4: Field Techniques, in: Larson, L.L. (Ed.), *Data Recovery Excavation and Archaeological Monitoring at the Tse-Whit-Zen Site (45CA523), Clallam County, Washington*. Report submitted to Washington State Department of Transportation, Olympic Region. Larson Anthropological Archaeological Services Limited., Gig Harbor, Washington.
- Shipman, P., Foster, G., Schoeninger, M., 1984. Burnt Bones and Teeth: An Experimental Study of Color, Morphology, Crystal Structure and Shrinkage. *J. Archaeol. Sci.* 11, 307–325.
- Smith, R.E., Butler, V.L., Orwoll, S., Wilson-Skogen, C., 2011. Pacific Cod and Salmon Structural Bone

- Density, in: Moss, M.L., Cannon, A. (Eds.), *The Archaeology of North Pacific Fisheries*. University of Alaska Press, Fairbanks, pp. 45–56.
- Stiner, M.C., Kuhn, S.L., Weiner, S., Bar-Yosef, O., 1995. Differential Burning, Recrystallization, and Fragmentation of Archaeological Bone. *J. Archaeol. Sci.* 22, 223–237. doi:10.1006/jasc.1995.0024
- Twiss, K.C., Wolfhagen, J., Madgwick, R., Foster, H., Demirergi, G.A., Russell, N., Everhart, J.L., Pearson, J., Mulville, J., 2016. Horses, hemiones, hydruntines? Assessing the reliability of dental criteria for assigning species to southwest Asian equid remains. *Int. J. Osteoarchaeol.* doi:10.1002/oa.2524
- Uebersax, J., 2014. Raw Agreement Indices [WWW Document]. *Stat. Methods Diagnostic Agreem.* URL <http://www.john-uebersax.com/stat/raw.htm#introd> (accessed 6.17.16).
- Wolverton, S., 2013. Data Quality in Zooarchaeological Faunal Identification. *J. Archaeol. Method Theory* 20, 381–396. doi:10.1007/s10816-012-9161-4
- Wolverton, S., Dombrosky, J., Lyman, R.L., 2016. Practical Significance: Ordinal Scale Data and Effect Size in Zooarchaeology. *Int. J. Osteoarchaeol.* 26, 255–265. doi:10.1002/oa.2416