Effects of Picture References on Reproducibility of Entheseal Change Recordation

Dustin J. Lloyd
Illinois State University, USA

Abstract: Reproducibility and observer agreement are critical to the production of standardized and comparable data sets. The reproducibility of recording methodology is a current issue in entheseal change studies. The effects of a picture reference guide on current entheseal change methodology are explored in the hope of increasing observer agreement and thus overall applicability of the picture reference guide method in a variety of archaeological contexts. The picture guide seems marginally effective; however, it requires a few modifications and subsequent testing before full-scale implementation.

Keywords: Entheseal change, Schroeder Mounds

Introduction

Entheseal changes (EC), formally musculoskeletal stress markers and renamed at the Coimbra Conference of 2009 (Henderson et al. 2010, 2012, 2015; Villotte 2013), are the recordation of osteophytic change at an enthesis. An enthesis is any muscular origin or insertion on the bone. The results and conclusions drawn from EC have been criticized in the past concerning limitations in methodology and broader applicability of their findings and interpretation (Jurmain et al. 2011; Schlecht 2012); the criticism is not unfounded. Various methods of scoring coexist (Hawkey and Merbs 1995; Henderson et al. 2010, 2012, 2015; Mariotti et al. 2007), which certainly lowers comparability between studies. Many individuals pursue questions of EC because they are valuable in decoding past life activities, life courses, social dynamics, and health, through the study of the physical manifestations of activity and health changes at the sites of EC. The various methods concentrate on different, largely functional, aspects of EC (e.g., biomechanical stress and quantifying types of osseous reactions) and do not always provide a holistic picture of other causative or confounding factors (e.g., trauma, chronic disease [e.g., diffuse idiopathic skeletal hypertrophy or DISH]) affecting entheseal change (Beyeler et al. 1990; Utsinger et al. 1976). Others do not account for the complex osteobiological etiology of entheseal development (e.g., age, sex, injury, occupation) and insist that entheseal changes are activity driven.

Currently, paleopathological researchers use three methods to assess entheseal change: Hawkey and Merbs (1995), Mariotti et al. (2007), and Henderson et al. (2010, 2012, 2015). One aspect of comprehensive recording is the categorical quantification of progressive reactive change. Most categorical quantification strategies are attribute lists with examples of the most extensive reactive changes; the recorder, depending on their personal experience, can variably interpret these attributes. The most recent method (Henderson et al. 2012, 2015) uses images. Observer error can confound and obscure the cause (s) and mitigating factors of entheseal change (Jurmain et al. 2011; Schlecht

2012). Thus, the production of standardized and comparable data is an ongoing and crucial debate within the EC community. This paper strives to critically examine methods established by Henderson et al. (2012) by testing the usage and efficacy of a picture reference compendium on an archaeologically recovered skeletal sample using five independent observers. Results are further compared against previously published results.

Background

Hawkey and Merbs

Hawkey and Merbs (1995) is the oldest of the three protocols. Their protocol was described in the authors' publication of a study of joint reactive changes in the upper extremities (clavicle, scapula, humerus, radius, and ulna) of an Inuit population. The scoring method has a proven past and alleviates reliance on an observer's experience through use of pictures to describe scoring degrees. Inter- and intraobserver error differences are statistically insignificant ($P < 0.5$) in a number of studies (Hawkey 1988; Hawkey and Street 1992; Nagy and Hawkey 1993; Peterson 1994), demonstrating a high reproducibility. Hawkey and Merbs (1995) described three main categories: robusticity (osseous reaction to biomechanical stress), stress legions (pitting on the cortical surface of the bone), and ossification (exostosis). Each category contains four scoring grades: 0–4 with zero being absence of the trait.

The Hawkey-Merbs scoring protocol is implicit, even offering measurement requirements and restrictions. This mitigates experience-based errors. Furthermore, the method quantifies three types of reactive entheseal change: robusticity, stress lesions, and ossification. Robusticity is further scored for both periosteal and myoskeletal attachment, stress lesions, and ossification exostosis. Unfortunately, the authors assume that entheseal pathological change and reactive development always have a biomechanical etiology. They offer no other explanations for entheseal change. Their assumptions are a product of the then accepted paradigm. Current research and studies show a more complex etiology for EC (Cardoso and Henderson 2010; Henderson and Cardoso 2013; Henderson et al. 2012; Milella 2012; Niinimaki 2012; Schlecht 2012), shedding light on possible limitations and how certain aspects of EC were scored. Correspondingly, this method requires a precise situation regulated by three rules: relatively narrow time frame, cultural and genetic isolation, and a small number of known and specialized activities. While not as glaringly obvious as some weaknesses, the authors wanted to associate certain enthesopathies and entheseal developments with particular activities; thus, they needed to restrict their studies to well documented, small, and specialized activities. However, their goals and interpretive framework notwithstanding, the resulting protocol underscored a larger problem: lack of standardized scoring protocols. Without standardization, it is difficult to compare results from different studies. Further problematic issues include no controls for age at death and no clinically or archaeologically supported occupation corroboration.

Mariotti et al.

Mariotti et al. (2007) proposed a detailed standardized scoring method for 23 postcranial entheses. The scoring protocol was similar to Hawkey-Merbs; however, there are a few key differences. Observers score three aspects of the enthesis: robusticity (vs. Hawkey and Merbs' definition), osteophytic enthesopathies (OF: prevalence of osteophytic activity), and erosive osteolytic enthesopathies (OL: vs. Hawkey and Merbs' definition of stress lesions). The authors provide many descriptive pictures for all 23 entheses to aid in scoring by comparison. Their method has an intra- and interobserver error of 28% when applying their five degree method. This reduces to 20% when using three out of the five scoring degrees. The authors recommend only using the three degree method unless the sample is large. Since the scoring scale is flexibly tailored to both the observers' expertise and the sample size, an experienced observer is able to gather more nuanced data from an archaeological context (Mariotti 2001; Mariotti and Belcastro 2011). Their method attempts to control for age through the use of robusticity. According to research, entheseal reaction positively correlates with increased age. Robusticity acknowledges the correlation of EC with age and provides a good comparative model for younger versus older individuals.

The method also suffers from some weaknesses. Despite the pictures and descriptions, the method is not user friendly. The protocol is demanding of the observer; therefore, the data collected are only as good as the observer. A good, well trained, and practiced observer would excel with this method; however, a less well trained one would find difficulty in scoring. Moreover, robusticity scores all have pictures for comparison. OF and OL are only addressed through scant written description. OF and OL are important entheseal changes and require a more descriptive scoring methodology than the generic protocol, which the authors provide. Generic descriptions of the degrees of OF and OL further problematizes the issue of experienced vs. amateur observer. A final shortcoming is the choice of parameters. It has been argued that the scoring methodology was articulated without reference to medical literature concerning entheseal etiology, namely the lack of separation between fibrocartilaginous and fibrous entheses (Jurmain et al. 2011; Vilotte 2009).

Henderson et al.

The Henderson et al. (2010, 2012, 2015) method evolved from a workshop at a conference in Coimbra, Portugal. The conference focused on musculoskeletal stress markers and their uses in reconstructing past activity as well as on reassessing terminology, recording methods, and possible correlates of repetitive (possibly occupation-related) joint movement (Henderson 2012). Henderson and colleagues attempted to construct a definitive standardized data collection method for EC (Henderson et al. 2010). As one of the newest attempts at quantifying EC data, the Henderson et al. method splits the enthesis into two zones: Zone 1 = the margin opposite the acute angle of muscle insertion; Zone 2 = remaining margin and surface of insertion. Zone 1 is scored for bone formation (BF Z1) and erosion (ER Z1). Zone 2 is scored for bone formation (BF), erosion (ER), fine porosity (FPO), macroporosity

(MPO), and cavitation (CA). A sampling of scoring degrees for the biceps brachii is shown below (Figure 1).

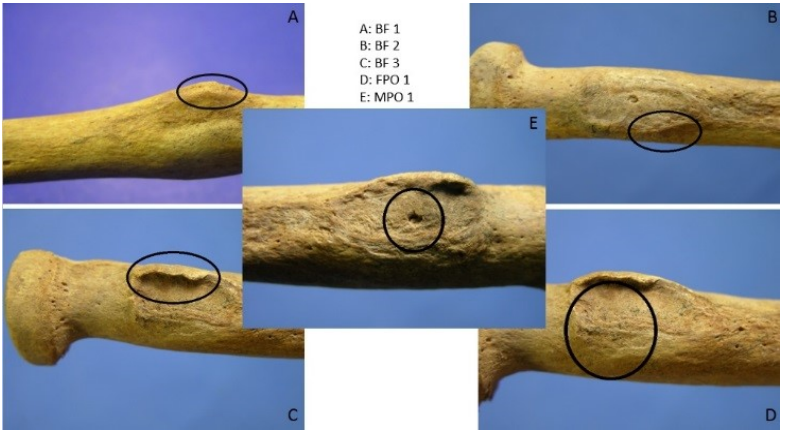

Figure 1: Sampling of Entheseal Variation from picture reference guide of proximal radial attachment of biceps brachii. A) Bone Formation Degree 1: Notice slight osteophytic nodule in circled area. B) Bone Formation Degree 2: Notice very slight raised ridge on under 50% of Zone 1 in circled area. C) Bone Formation Degree 3: Notice very well developed ridge on more than 50% of Zone 1 in circled area. D) Fine Porosity Degree 1: Notice fine pin prick–like depressions within the circled area. E) Macroporosity Degree 1: Notice enlarged porosity (> 1 mm) in circled area.

Their method excels at quantifying the types of reactive changes on an enthesis. Henderson et al. (2012) suggest that a closer examination of the suite of changes may provide more information on the age correlation that the other methods note but do not quantify (Hawkey and Merbs 1995; Mariotti 2007). Their method allows for the collection of data, which can differentiate the types of changes in different ages groups (e.g., whether certain changes take place more often or in greater severity in old or young individuals). According to the authors, bone formation seemed to be closely correlated with age, which makes sense since an older individual would have more opportunity to incur microtrauma resulting in increased bone formation; however, the study sample was small so this may not hold true for a larger sample. A wide array of features recorded helps to diversify the data so that more complex questions can be asked since activity patterns are easier to recreate. Overall, the error rate was 20% (Henderson et al. 2015), though error on certain entheses ranged from 30–40%.

A major problem of the Henderson et al. (2012) method is the issue of reproducibility; there was a systematic disagreement between observers one and four in their study (Henderson et al. 2012). Observer one consistently scored higher than four. The authors attribute this to each observer's previous creation of their own scoring protocol. However, such disagreements underscore the larger issue of lack of standardized scoring methods. That withstanding, a systematic disagreement between the same observers may also point to a programmatic flaw in their method.

Some scoring points (i.e., MPO) have rather banal scoring protocol,

which is easy to understand; others (i.e., BF and ER) have more vague and overlapping scoring criteria, which rely on the observer's judgment to segregate. The three-degree scoring levels encompass the extremes and middle option. To reduce error, Henderson et al. (2015) recommended decreasing the scoring degrees for BF and ER from three to two. Decreasing options will certainly reduce the error. However, combining the middle and upper extreme or middle and lower extreme would lead to data loss and possibly more interpretive confusion. For example, a skeletal sample that was characterized as robust but normal on the old scoring system may now present scores in a higher range and support a false assumption of advanced entheseal development.

The authors suggested an accompanying picture compendium, which outlines what each scoring point looks like. The addition of pictures strengthened Mariotti et al.'s (2007) method and would do the same for the Henderson et al. method. Visual description should theoretically decrease inter- and intraobserver error, which would be a great step toward the attainment of a standardized scoring method.

Methods and Materials

The collection utilized for this test is a Late Woodland (AD 900–1150) skeletal sample from the Schroeder Mounds site (11He177), which is currently housed at Illinois State University. The skeletal elements scored are non-pathological and include 41 elements: 15 humeri, 18 radii, and 8 scapulae. Five observers (identified by numbers one through five) with osteological experience assessed the 41 elements at four entheses: biceps brachii, triceps brachii, infraspinatus, and supraspinatus. Biceps brachii was scored at the long head attachment on the radial tuberosity. Triceps brachii was scored at the medial head attachment on the distal posterior aspect of the humerus slightly superior to and surrounding the olecranon fossa. Infraspinatus and supraspinatus were scored at their scapular attachment sites of infraspinous fossa and supraspinous fossa respectively.

A scoring pamphlet (Figure 1) was given to all observers. It contained a list of definitions for each enthesis, picture references for each scoring feature and degree, and two examples of a scored enthesis for each scoring feature and degree (Henderson et al. 2010). Observers were given preliminary instructions on scoring, which included an overview of the definitions of bone formation (BF), erosion (ER), macroporosity (MPO), and fine porosit (FPO). Charts and tables use the previous abbreviations for scoring features: Z1 and Z2 are used for Zone 1 and Zone 2 respectively. Observers scored each enthesis without input from the author or other observers. Observers scored twice, a week apart, to provide data for both inter- and intraobserver agreement.

The scoring definitions mirrored the work of Henderson et al. (2015), and the scoring degrees came from their 2012 publication (Table 1). Furthermore, the definition of cavitation makes macroscopic identification and photography difficult since the opening must be smaller than the subcortical cavity. Cavitation was also not scored since no good examples existed in the Schroeder Mound sample, which made it difficult to test the usefulness of the picture reference guide. Textural change was not recorded since it is

scored on a presence or absence basis, which is not a robust enough scoring protocol. As such, cavitation and textural change were excluded from this research.

Table 1: Scoring definitions and degrees of expression (Henderson et al. 2010, 2015).

	Scoring Feature	Degree of Expression
Zone 1: Margin opposite acute angle of fiber attachment	Bone Formation (BF Z1): See degrees of expression. Normal morphological smooth-rounded or mound-like (check by touching) margins, even if the margin is elevated, should be scored as 0.	1 = small, nodular or slightly raised margin < 1 mm 2 = distinctive, sharp crests or other enthesophytes ≥ 1 mm but < 50% of margin 3 = distinctive, sharp crests or other enthesophytes ≥ 1 mm but ≥ 50% of margin
	Erosion (ER Z1): Depression or excavations of any shape and involving discontinuity of the lesion greater in width and depth with irregular margins. Only erosions >1mm where the floor can be clearly seen were recorded. This does not include pores (rounded margins). Score erosions if they occur on bone formation.	1 = < 25% margin 2 = 25 to 50% margin 3 = > 50% of margin

Table 1 cont.

Zone 2: Remaining margin and surface	Bone Formation (BF Z2): Any bone production from roughness of surface to true exostoses (e.g., distinct bone projections of any form, like bony spurs, bone nodules, and amorphous bone formation).	1 = roughness/rugosity change is diffuse not a distinct structure 2 = distinct structure measuring > 1 mm, affecting < 50% of surface 3 = distinct structure measuring > 1 mm, affecting $\geq$ 50% of surface
	Erosion (ER Z2): Depression or excavations of any shape (but not covered by the definition of macroporosity) and involving discontinuity of the floor of the lesion greater in width than depth with irregular margins. Only erosions > 2 mm were recorded. MPO and FPO occurring within an erosion should not be recorded separately. Bone formation is only scored if it exceeds the height of the depression (do not score woven bone). Score erosions if they occur on bone formation.	1 = < 25% of surface 2 = 25–50% of surface 3 = > 50% of surface

	Fine Porosity (FPO): Small, round to oval perforations with smooth, rounded margins < 1 mm. These should be visible to the naked eye and be in a localized area. Do not score if they are at the base of an erosion of if they occur as part of woven bone.	1 = < 50% of surface 2 = $\geq$ 50% of surface
	Macroporosity (MPO): Small, round to oval perforations with smooth, rounded margins about 1 mm or larger in size with the appearance of a channel, but the internal aspect is rarely visible. Do not score if they are at the base of an erosion.	1 = one or two pores 2 = > 2 pores

Interobserver agreement was tested by feature and enthesis. Each observer was compared to other observers for the two scoring sessions to obtain the rate of agreement between observers and between the two tests of the same observer. Agreements at each enthesis and scoring degree were compared to produce percent agreement scores (Figures 1&2). An exact agreement was counted as a match. These agreement percentages were then averaged for each observer pair across either entheses or scoring degrees into a composite agreement percentage (Table 2). A composite score represents the average agreement between observers at each enthesis or scoring feature. Inter- and intraobserver agreement was calculated for both scoring rounds. Intraobserver agreement was calculated with the same averaging methodology as interobserver agreement. Since the observers had no previous experience with the methods and since the project goal was to test the methods, the interobserver error and their potential statistical significance were calculated from the results of the second scoring session.

A Fisher's exact test was also done to determine the statistical significance of this paper's findings relative to the original tests (Henderson et al. 2012). Data were rounded up or down to the closest whole number to adhere to the parametric standards of the test. The data was also calculated using proportional fractions to ensure quality of the rounding method. The results of significance or nonsignificance were the same from both methods. Below results are from the rounded up or down figures.

The choice of entheses in this study varies from the entheses scored in the Henderson et al. (2012) scoring tests. Entheses utilized in this test mirror the entheses that will be used for a later master's thesis work. Data collected here serve to evaluate the effects of a picture reference guide on observer agreement and as a pilot study for methods in future data collection. Thus, it was more important to test the method on those entheses than to mirror the entheses of Henderson et al. (2012). The results are still comparable since both tests assessed percent agreement and not the agreement of a specific enthesis.

Results

Highest interobserver agreement for an enthesis is the supraspinatus at 87.8% (Figure 2); however, results may be slightly skewed by the relatively low number of scapulae in the sample (n = 8) relative to humeri (n = 15) and radii (n = 18). Discounting the scapulae score, the highest interobserver agreement by enthesis is the biceps brachii at 72.6% (Figure 2). Composite average interobserver agreement for all entheses is 76%.

Highest interobserver agreement by scoring feature was MPO at 95.2% (Figure 3). Composite average agreement by feature was 72.6%. Highest interobserver agreements for all scoring features were between observer one and observer three (89.5%) and between observer one and observer five (82.3%) (Figure 3). Composite average interobserver agreement by scoring feature is 72.9%. Overall interobserver agreement of combined feature and enthesis was 74.45% (Table 2).

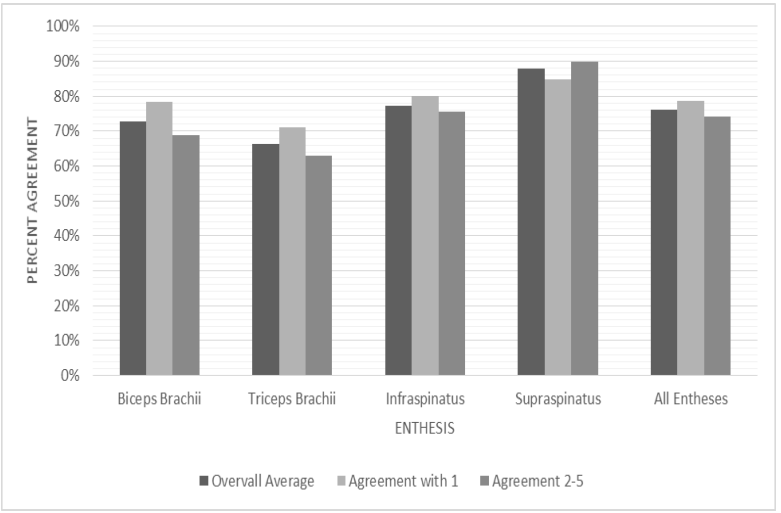

Figure 2: Interobserver agreement by enthesis.

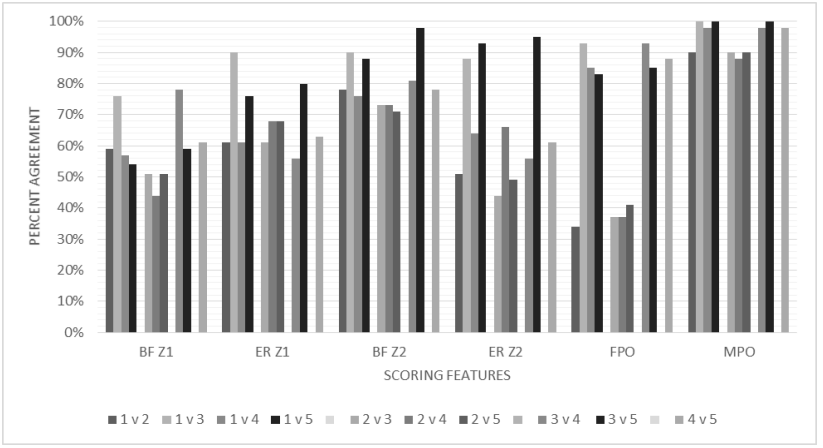

Figure 3: Interobserver agreement by scoring feature.

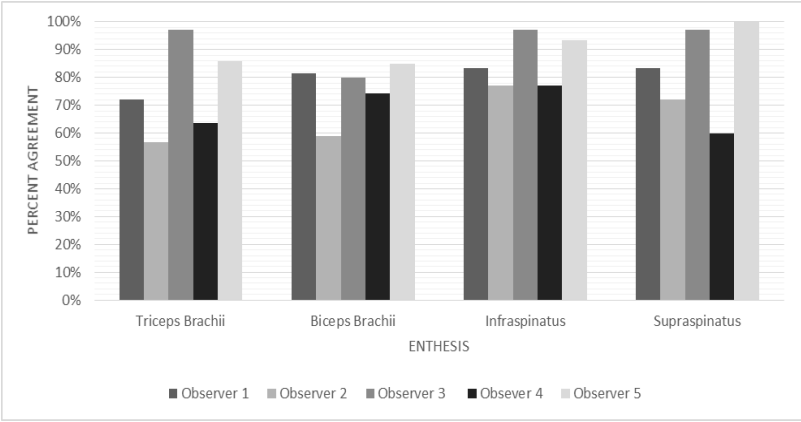

Figure 4: Intraobserver agreement by enthesis.

Table 2: Composite interobserver agreement scores by feature and enthesis: roman font (upper right) are the feature composite scores; italicized (lower left) are enthesis composite scores.

Observer	1	2	3	4	5
1		62.17%	89.5%	73.5%	82.33%
2	*67.12%*		59.33%	62.67%	61.67%
3	*87.29%*	*65.17%*		77.0%	86.67%
4	*74.67%*	*67.74%*	*81.96%*		74.83%
5	*85.33%*	*65.0%*	*89.29%*	*77%*	

The composite intraobserver agreement by scoring feature is 76.6% (Table 3). Composite intraobserver agreement by enthesis is 79.7% (Table 4). Overall intraobserver agreement by feature and enthesis is 78.15% (Tables 3&4). Fisher's exact test revealed significant statistical variation at ER Z1, BF Z2, and MPO. Table 5 provides the breakdown of the Fisher's exact test results of interobserver agreement by scoring feature and entheseal comparison.

Table 3: Intraobserver agreement by scoring

Scoring Zone	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5
BF Z1	53.7%	61%	78%	27%	61%
ER Z1	75.6%	76%	85%	66%	78%
BF Z2	82.9%	73%	83%	66%	97.5%
ER Z2	75.6%	59%	88%	73%	88%
FPO	80.4%	51%	97.5%	85.5%	88%
MPO	100%	54%	97.5%	100%	100%
Average	78.03%	62.33%	88.17%	69.5%	85.42%

Table 4: Intraobserver agreement by enthesis; parentheses indicate number of each enthesis scored.

Enthesis	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5
triceps brachii (18)	72.07%	56.72%	97%	63.5%	86%
biceps brachii (15)	81.5%	59%	80%	74.17%	85%
infraspinatus (5)	83.3%	77%	97%	77%	93.3%
supraspinatus (3)	83.3%	72%	97%	60%	100%
Average	80.04%	66.18%	92.75%	68.67%	91.08 %

Table 5: Fisher's exact test by scoring feature (interobserver) and Entheseal Comparison.

Scoring Zone	P Value	Entheseal Comparison	P Value
BF Z1	.7725	biceps brachii vs. biceps brachii	1.000
ER Z1	.0332*	triceps brachii vs. iliopsoas	1.000
BF Z2	.0001***	supraspinatus vs. common extensor	.0279*
ER Z2	.4406	infraspinatus vs. achilles	.7425
FPO	.4587	infraspinatus vs. iliopsoas	.02648 *
MPO	.0012**	supraspinatus vs. iliopsoas	.0008***
SD = .321606; 90% CI = .2161 - .6719; 95% CI = .2007 - .7887; 99% CI = .1757 – 1.1207.		SD = .46604; 90% CI = .31320 - .97368; 95% CI = .29090 – 1.14302; 99% CI = .25463 – 1.62405.	

* = significant ** = very significant *** = extremely significant.

Discussion

Decrease in Average Score

Outside of the scores for ER Z1/2 and FPO for observer two and BF Z1 for observer five, the observers tended to score lower in the second scoring session. Extraneous factors like differential lighting between testing or psychological effects like mental fatigue are unlikely to be responsible for the decreases in scoring in the second session because the sessions were conducted in the same classroom and the sessions were brief. The most logical interpretation is that the observers scored less the second time because of familiarity with the scoring protocol. Observers scored much faster the second attempt. In the first attempt, the observers were recorded as taking between 45–60 minutes to complete the task; however, the second attempt only took each about 20–30 minutes, supporting the idea that they had gained familiarity with the method.

Decrease in average scores between sessions mirrors the Henderson et al (2012) results. In summary, the more often an observer uses the method, the more they agree with each other and themselves. The decrease makes sense since the observers have now seen double the amount of bones and can assess the differences between each scoring degree. Experience will help to determine a score of one or three. If an observer is comfortable with the extremes, then they determine two by elimination.

Variance in Agreement at Entheses

Results indicate a stark difference between entheseal agreements. The difference between the triceps brachii and supraspinatus is statistically significant at $p = .0285$. The method may simply be better at describing one enthesis over another, or the scoring protocol may better describe entheseal change at the supraspinatus than at the triceps brachii. Different types of reactions may take place at each enthesis. Typical entheseal changes at the supraspinatus may be more easily recorded by this method than the changes at the triceps brachii.

Another possible explanation is that changes unrecordable with this method took place at the triceps brachii, which caused observers to find a category that best fit the observed reactive change. Observers may have scored a change in an inappropriate category. One observer noted reactive change on the scapular spine that was outside of the area of intended scoring. Perhaps future iterations of this method need to consider changes that happen to the surrounding cortical bone. A major function of an enthesis is to dissipate stress from the muscle body down into the enthesis and the surrounding cortical bone. The cortical bone may also have potentially scorable reactive changes.

The difference may also indicate an overall lack of experience with each enthesis. The highest entheseal agreement is the biceps brachii, which has one insertion at the radial tuberosity. Most osteologists have more extensive experience with the radial tuberosity through normal osteological data collection. The one scored attachment for the tricep brachii was on the distal posterior humerus, which is not a typical area for osteological data collection. All of the observers in this trial normally collect paleopathological data, and the distal posterior humerus rarely presents any pathological condition in isolation. Experience with this surface may be limited, meaning some could have scored

normal bone surface as entheseal reaction. Additional trials targeting entheses of lesser and greater knowledge would confirm or deny this hypothesis.

Another possible explanation is that the variance between individual entheses does not matter as much as looking at the overall composite muscle groups for arms, legs, torso, etc. Weiss (2003) suggests composite entheseal scoring by muscle groupings. Looking at entheses as collaborative groups may reveal new archaeological applications. An action is not performed with one muscle and therefore involves multiple entheses. Shoulder abduction, a mundane action, requires the supraspinatus, deltoid, trapezius, and serratus anterior. Consideration of multiple entheses and the creation of composite scores for muscle groups may assist in recreating the types of actions of past people within an archaeological context.

Variance in Interobserver Agreement

Wide variance in agreement speaks to the reproducibility issues found in the previous testing attempts (Henderson et al. 2012). Addition of pictures was intended to alleviate this issue and bring the agreement scores closer to the 80% score of the newest Henderson et al. method (2015). Pictures may still have a role; however, a reference book may not be the ideal place.

An odd trend emerged during the second scoring sessions. Observers extensively used the picture references during the first scoring session, which was the intended use for the guides; however, only one observer used the reference guide during the second session, and they only used the guide twice. Every observer produced higher agreement scores during the second test.

One possible explanation is that the books are better suited as a teaching tool than a reference guide. Observers received no extensive instruction using the guide prior to their first scoring. They used the book to reference each bone with pictures or attempted to match the reaction on the bone to one of the pictures. On the second scoring sessions, the observers seemed to have a mental image of what each scoring degree entailed and felt no need for the scoring pamphlets; this conjecture requires more experiments on new observers unfamiliar with the protocols to validate this conclusion.

Ways to Improve the Picture Guide

If the picture guide is to function as both a reference and teaching tool, then the pictures need to change in quantity and quality. The observers' guide contained examples of each scoring feature and degree; however, the guide did not contain pictures of every scoring degree for every enthesis. The book was heavily weighted toward the radial insertion of the biceps brachii. One observer remarked that pictures of feature degrees by enthesis would assist her. Inclusion of each enthesis may also decrease the chances that an observer will score normal bone as entheseal change. The largest incongruence is between scores of zero and a positive value. Some observers seem to have been scoring normal bone as entheseal change; this phenomenon may be related to inexperience in both the scoring protocol or with the enthesis. Observer three had the most osteological experience and also had the highest composite intraobserver agreement at 90.4%. It seems that unfamiliarity with

the method may play a bigger role than unfamiliarity with an ethesis. Inclusion of entheseal development at each enthesis may increase inter- and intraobserver agreement in future testing and decrease scoring normal osseous development as entheseal change.

Another possible addition to the reference guide is the inclusion of various pictures of each scoring feature degree. Entheseal change is not a discrete variable but rather a continuous one making scoring a combination of experience and standards. Various states of change could be scored the same. For example, FPO is scored on a percentage basis: 10%, 25% and 50% FPO are all scored as a one. Adding pictures that reflect the multiple forms that a score of one can take would increase observer agreement. Entheseal change is progressive and the scoring protocols need to reflect this aspect of entheseal change. Jacobi and Danforth (2002) suggest a similar idea for scoring porotic hyperostosis and cribra orbitalia.

Results of Fisher's Exact Tests Relative to Picture Guide

Results of the Fisher's exact tests are very interesting. They suggest that the pictures may be useful at scoring features ER Z1, BF Z2, and MPO. The aforementioned features all had results with statistically significant variation from the 2012 Henderson et al. scoring trials. Pictures may assist observers in more often agreeing on the types of changes happening in those zones; however, the 2012 sample was much larger than this sample. It is uncertain if these results are an artifact of smaller relative sample size or if they would hold true for a larger sample.

The results also indicate that the pictures may be more helpful at correctly identifying changes at an enthesis of relative unfamiliarity. Comparisons between common extensor and supraspinatus ($p = .0279$) and iliopsoas and supraspinatus ($p = .0008$) are both statistically significant, indicating that the pictures may assist observers, who are unfamiliar with a particular enthesis. The common extensor is a more commonly encountered entheseal surface than the supraspinatus. The iliopsoas and supraspinatus are both relatively unfamiliar. In both cases, the results of this study were positively statistically significant relative to the 2012 results. Therefore, pictures seem to assist in identifying entheseal change at an unfamiliar enthesis.

Conclusions

The goal of this paper was to test the effects of a picture reference on the existing Henderson et al. (2012) methodology and to assess inter- and intraobserver error and repeatability. The study suggests that the picture references may be more useful at certain entheses (biceps brachii and supraspinatus) and at certain scoring features (ER Z1, BF Z2, and MPO). Various additions and modifications to the picture reference book were also explored to increase its practicality. Results and observer experience advocate for an increase in quantity and quality of pictures. The picture reference should include pictures of all entheses at all scoring stages and multiple pictures of scoring degrees scored on a percent present basis.

The picture compendium developed for this study should not be considered generally applicable yet. Testing of the above modifications and their

effects on inter- and intraobserver error and reproducibility are still needed. Reproducibility seems largely reliant on experience; however, using the picture guide as a teaching tool may alleviate this. The picture guide needs to be more fully developed and then tested on a group of observers with no prior knowledge or experience to assess the general usability of the method outside of researchers already familiar with entheseal changes. Although more testing is necessary, the results of this study suggest that a fully developed, comprehensive picture guide is both a good teaching tool and a means to increase inter- and intraobserver agreement. Additional testing should include a large sample from diverse populations within a variety of health and activity constraints to reflect the variety of etiological causation of EC and to create a more comparative database.

References

Beyeler, Christine, P. Schlapbach, N. J. Gerber, J. Sturzenegger, H. Fahrer, S. J. Van der Linden, U. Burgi, W. A. Fuchs, and H. Ehrengruber. 1990. "Diffuse Idiopathic Skeletal Hyperostosis (DISH) of the Shoulder. A Cause of shoulder Pain?" *British Journal Rheumatology* 29:349–53.

Cardoso, F. Alves, and Charlotte Henderson. 2010. "Enthesopathy Formation in the Humerus: Data from Known Age-at-Death and Known Occupation Skeletal Collections." *American Journal of Physical Anthropology* 141:550–60.

Hawkey, Diane. 1988. "Use of Upper Extremity Enthesopathies to Indicate Habitual Activity Patterns."Master's Thesis, Arizona State University.

Hawkey, Diane, and Charles Merbs. 1995. "Activity-Induced Musculoskeletal Stress Markers (MSM) and Subsistence Strategy Changes among Ancient Hudson Bay Eskimos." *International Journal of Osteoarchaeology* 5:324–38.

Hawkey, Diane, and S. Street. 1992. "Activity-Induced Stress Markers in Prehistoric Human Remains from the Eastern Aleutian Islands." *American Journal of Physical Anthropology* 14:89–89.

Henderson, Charlotte, Valentina Mariotti, Doris Pany-Kucera, Geneviève Perréard-Lopreno, Sébastien Villotte, and Cynthia Wilczak. 2010. "Scoring Entheseal Changes: Proposal of a New Standardized Method for Fibrocartilaginous Entheses." Poster presented at the 18th European Meeting of the Paleopathology Association, Vienna, Austria, August 23 –26.

————. 2012. "The Effect of Age on Entheseal Changes at Some Fibrocartilaginous Entheses." *American Journal of Physical Anthropology* 147:163.

————. 2013. "Recording Specific Entheseal Changes of Fibrocartilaginous Entheses: Initial Tests Using the Coimbra Method." *International Journal of Osteoarchaeology* 23:152–62.

————. 2015. "The New 'Coimbra Method': A Biologically Appropriate Method for Recording Specific Features of Fibrocartilaginous Entheseal Changes." *International Journal of Osteoarchaeology*. Accessed Novem-

ber 5, 2015. doi: 10.1002/oa.2477.
Jacobi, Keith P., and Marie Danforth. 2002. "Analysis of Interobserver Scoring Patterns in Porotic Hyperostosis and Cribra Orbitalia." *International Journal of Osteoarchaeology* 12:248–58.
Jurmain, Robert, Francisca Alves Cardoso, Charlotte Henderson, and Sébastien Villotte. 2011. "Bioarcheology's Holy Grail: The Reconstruction of Activity." In *A Companion to Paleopathology*, edited by Anne Grauer, 557–78. Chichester: Wiley-Blackwell.
Mariotti, Valentina. 2001. "Skeletal markers of activity in the warriors from the Celtic necropolis of Casalecchio di Reno (Bologna, Italy) (IV–III c. BC)." Paper presented at Atti XIII Congresso Antropologi Italiani, Sabaudia, Italy, October 4–8.
Mariotti, Valentina, and Maria Giovanna Belcastro. 2011. "Lower Limb Entheseal Morphology in the Neandertal Krapina Population (Croatia, 130 000 BP)." *Journal of Human Evolution* 60:694–702.
Mariotti, Valentina, Fiorenzo Facchini, and Maria Giovanna Belcastro. 2007. "The Study of Entheses: Proposal of a Standardized Scoring Method for Twenty-Three Entheses of the Postcranial Skeleton." *Collegium Antropologicum* 31:291–313.
Milella, Marco, Maria Giovanna Belcastro, Christoph Zollikofer, and Valentina Mariotti. 2012. "The Effect of Age, Sex, and Physical Activity on Entheseal Morphology in a Contemporary Italian Skeletal Collection." *American Journal of Physical Anthropology* 148:379–88.
Nagy, B. L., and Diane Hawkey. 1993. "Correspondence of Osteoarthritis and Muscle Use in Reconstructing Prehistoric Activity Patterns." Paper presented at Twentieth Annual Meeting of the Paleopathology Association Toronto, Canada, April 1993.
Niinimäki, Sirpa. 2011. "What Do Muscle Marker Ruggedness Scores Actually Tell Us?" *International Journal of Osteoarchaeology* 21:292–99.
Peterson, Jane Darden. 1994. "Changes in the Sexual Division of Labor in the Prehistory of the Southern Levant." PhD diss., Arizona State University.
Schlecht, Stephen H. 2012. "Understanding Entheses: Bridging the Gap Between Clinical and Anthropological Perspectives." *The Anatomical Record* 295:1239–51.
Utsinger, P. D., Donald Resnick, and Robert Shapiro. 1976 "Diffuse Skeletal Abnormalities in Forestier Disease." *Archives of Internal Medicine* 136:763–68.
Villotte, S., and C. J. Knüsel. 2013. "Understanding Entheseal Changes: Definition and Life Course Changes." *International Journal of Osteoarchaeology* 23:135–46.