

# Don't cry over spilled ink: missing context prevents replication and creates the Rorschach effect in bone surface modification studies

Stephen R. Merritt<sup>a</sup>  
Michael C. Pante<sup>b</sup>  
Trevor L. Keevil<sup>b</sup>  
Jackson K. Njau<sup>c,d</sup>  
Robert J. Blumenschine<sup>e</sup>

## Affiliations

- a. Anthropology Department, University of Alabama at Birmingham, 1401 University Blvd., Birmingham, Alabama 35294, USA
- b. Department of Anthropology, Colorado State University 1787 Campus Delivery, Fort Collins, Colorado, 80523, USA
- c. Department of Earth and Atmospheric Sciences, Indiana University, 1001 East 10<sup>th</sup> St., Bloomington, Indiana, 47405 USA
- d. The Stone Age Institute, P.O. Box 5097, Bloomington, Indiana, 47407, USA
- e. Palaeontological Scientific Trust, O.O. Box 52379, Saxonwold 2132, Johannesburg, South Africa

## **Abstract**

The scientific replicability crisis has recently focused on bone surface modification (BSM) analysis, which underlies zooarchaeological and anthropological conclusions about the ecology and evolution of tool-assisted carcass consumption behavior. We review a recent blind test of inter-analyst correspondence in morphometric analysis of experimentally generated butchery marks that advocates algorithmic methods for diagnosing and measuring BSM in an effort to standardize methodology and minimize inter-analyst error (Domínguez-Rodrigo et al. 2017. Use and abuse of cut mark analyses: The Rorschach effect. *Journal of Archaeological Science*, 86, 14–23. <https://doi.org/10.1016/j.jas.2017.08.001>). This study overstates concern about the inaccuracy of BSM measurement and interpretation, concluding that BSM analysis is a subjective, non-scientific endeavor. Based on a minimally described sample of cut marks, it measures variables that involve inherent inaccuracy and subjectivity and overlooks how the contexts of experimental sample generation – particularly the difference between immanent and configurational processes – differentially affect cut mark morphometrics. We illustrate this discussion with experimental taphonomic examples focused on analytical context including sample construction and control over factors that affect cut mark cross-sectional size. Our analysis suggests the relationship between tool attributes and cut mark morphology is not generalizable to all experimental and archaeological butchery contexts. We show that our experimental samples capture metric variability observed in archaeological cut marks, but that intentionally incised marks and realistic defleshing marks differ in width and depth. Further, when controlling for factors that impact cut mark size including animal size class, tool type, butcher experience, and density across bone portions, overlapping cut mark widths and depths produced by phonolite and ignimbrite flakes lead to poor classification of marks according to causal flake material, which casts doubt on the ability to discriminate cut marks made by

different materials. We build datasets that include diverse experimental contexts and suggest that meta-analysis can disentangle how multiple configurational factors contribute to cut mark morphometric attributes. Ultimately, progress in BSM analysis rests on inter-analyst replicability, which must be preceded by clear discussion of all parts of the inferential loop – from the design of experiments that generate actualistic analogues, to their use in supporting archaeological arguments. Otherwise, problematic expert knowledge traditions may mask arguments from authority in sophisticated methodological language and under-reported experimental context.

### Keywords

Experimental taphonomy; cut marks; butchery; generality; realism; context; expert knowledge

### Highlights

Is morphometric analysis of bone surface modifications inherently subjective?

The Rorschach argument says so, but inappropriately measures a problematic BSM sample

Analytical context is under-reported in BSM analysis, which emphasizes new methods

Intentional incision creates wider and deeper cut marks than realistic butchery

Replication in BSM analysis requires clear contextual presentation

## **1. Introduction**

Recent analytical attention on bone surface modifications (BSM) accompanied groundbreaking zooarchaeological discoveries that may indicate pre-Oldowan tool-assisted carcass consumption (McPherron et al., 2010), an earlier human presence in southeast Asia and North America (Dambricourt Malasse et al., 2016, Holen et al., 2017), and Late Pleistocene hominin cannibalism or ritual defleshing (Bello et al., 2016). Simultaneously, innovations in morphometric analysis use increasingly sophisticated techniques to visualize and measure BSM and distinguish stone tool cut marks, terrestrial and aquatic carnivore feeding marks, and traces of other biotic agents and abiotic processes (Mate-Gonzalez et al., 2015, 2017, Harris et al., 2017, Otarola-Castillo et al., 2018, Pante et al., 2017). Despite this disciplinary progress, analytical consensus about the interaction of humans, competitors, and their prey taxa, along with the technological, ecological, and social contexts that surround carcass access, butchery behavior, and consumption is limited by divergent interpretations of BSM that align with distinct analytical or expert knowledge traditions (Domínguez-Rodrigo et al., 2010, Njau, 2012, Sahle et al., 2017).

Unstandardized measurement techniques and inconsistent descriptive terminology prevent consensus in BSM analysis, suggesting the “cycle of experimentation and interpretation has become progressively less useful” (James and Thompson, 2015:93). Although recent actualistic research developed sophisticated tools for measuring and summarizing BSM attributes and discriminating diverse taphonomic effectors, agents, and processes, (see Njau and

Blumenschine, 2006, Domínguez-Rodrigo et al., 2012, Gidna et al., 2014, 2015, Parkinson et al., 2015, Pobiner et al., 2018, Organista et al., 2016, Merritt, 2017), without replication their methodologies and conclusions may generate expert knowledge traditions that are not universally accepted and disguise non-scientific praxis (Domínguez-Rodrigo et al., 2012, 2014, Pante et al., 2015, Thompson et al., 2015, Atmanspacher and Maasen, 2016).

Domínguez-Rodrigo and colleagues (2017) addressed this issue directly with a blind test where 11 analysts trained in the same research tradition identified microscopic morphometric attributes in experimental cut marks. Poor inter-analyst correspondence led the authors to conclude that cut mark identification and measurement are subjective, and along with other recent publications, suggest an algorithmic approach including machine learning (Arriaza and Domínguez-Rodrigo, 2016, Domínguez-Rodrigo and Baquedano, 2018) and Bayesian inferential models (Harris et al., 2017, Otarola-Castillo et al., 2018) as an analytical solution that standardizes methodology and minimizes inter-analyst error when measuring and classifying BSM.

We argue that progress in BSM analysis requires inter-analyst replicability, which demands clear discussion of all parts of the inferential loop – from the design of experiments that generate actualistic analogues, to their use in archaeological arguments. We support this claim with a critical review of Domínguez-Rodrigo and colleagues' recent blind test (2017) and show that under-reported contextual detail prevents its replication and may have shaped its conclusions.

Our comments address an inflection point in archaeological interpretation – where skepticism about accurate identification and idiosyncratic expert knowledge reflect a paradigm shift toward algorithmic methods intended to automate analysis and remove human subjectivity. This approach may be fruitful, but without sufficiently reported detail, new analytical methods and the information they generate cannot be critically evaluated or replicated, and may disguise potentially faulty expert knowledge in novel quantitative methods. To support this discussion, we present experimental BSM data to highlight how analytical context, including sample construction and experimental design impact cut mark morphometrics, and illustrate the minimum contextual details required for evaluation and replication of actualistic BSM research.

## **2.1 Summary of the Rorschach argument and its under-reported context**

Domínguez-Rodrigo and colleagues (2017) conducted a blind test involving 11 analysts trained in the same analytical methods who identified 14 microscopic variables that describe cut mark orientation, morphometric attributes of the cross-section, internal microstriations, and shoulder. The testing sample of 30 cut marks was selected from a larger experimental BSM dataset representing an expert knowledge tradition that interprets cut marks made during Early Stone Age carcass consumption, and claims to successfully diagnose causal tool type or raw material from cut mark morphometrics (see Domínguez-Rodrigo, 1997, Domínguez-Rodrigo and Barba, 2005, Domínguez-Rodrigo et al., 2009, de Juana et al., 2010).

Despite being trained in the same research tradition, the test revealed significant inter-observer variability in cut mark morphological attribute identification, which the authors liken to idiosyncratic responses to Rorschach ink blots, ultimately arguing that BSM analysis is subjective. We applaud the Rorschach argument's investigation of repeatability, and agree that measurement of the cut mark variables defined by Domínguez-Rodrigo et al. (2017) is subjective. However, rather than rejecting BSM analysis as pseudo-scientific, we suggest poor inter-analyst correspondence highlights methodological problems that prohibit analytical

standardization, a viewpoint that is prominent in archaeological blind testing literature (Blumenschine et al., 1996, Wadley and Lombard, 2007, Wolverton, 2013, James and Thompson, 2015, Rots et al., 2016). Below, we explain why inter-analyst discordance is influenced by under-reporting the contexts of experimental data collection, problems with conducting and reporting the blind test, and the fact that cut mark attributes examined in the blind-test reflect a narrowly defined analytical perspective which does not appropriately describe cut mark morphometric variability.

## **2.2 Vague definition of the actualistic dataset**

Domínguez-Rodrigo et al. (2017) do not present enough contextual detail to address whether confounders of cut mark size and morphology are controlled during experimentation or sample construction. Thirty cut marks made with simple flint flakes were selected from a larger experimental dataset of 246 marks. This sample is minimally described and not examined to ensure it appropriately represents the original cut mark population (i.e., Domínguez-Rodrigo and Barba, 2005, Domínguez-Rodrigo et al., 2009). Further, the blind test cut marks were drawn from humeral, femoral, radial and tibial fragments, but without knowing which specific elements and portions (epiphyses, near-epiphyses, midshafts) contributed the marks under study, we cannot assess whether differences in bone density may have impacted cut mark size and morphology (Merritt, 2012, Braun et al., 2016).

It is unclear how many individuals acted as butchers, which tools they used, their prior experience with carcass processing, or whether they butchered goat or sheep limbs. Butchery expertise affects the number and length of cut marks (Pobiner et al., 2018), and distinct butchery actions like skinning, defleshing, and disarticulation target distinct musculoskeletal tissues and incise bone portions of different density, in turn affecting cut mark width and depth (Merritt, 2017). Butchery context may have influenced cut mark morphometric variables like shoulder effect and overlapping striae (see Domínguez-Rodrigo et al., 2017, Table 1), or affected cross-sectional size and morphology. Additionally, specimen preparation methods are not described, obscuring whether cortical surface texture, which impacts microscopic cut mark morphology, was affected by boiling, cold water maceration, or burial (James and Thompson, 2015).

## **2.3 Subjective cut mark variables and blind testing procedure**

It is impossible to completely evaluate and replicate the blind test as described. Analysts use low magnification binocular microscopes (14-40x), which is a common technique, but their experience with comparative collections of cut marks whose morphometric attributes are generated under controlled conditions is not specified.

The experimentally generated mark sample is used to represent the entire universe of stone flake cut mark morphology, but we do not know whether this sample includes a narrow distribution of character states around typical flake cut mark morphology for all 14 variables. If so, we should expect all analysts to score them similarly, but marks with unusual morphology may exaggerate analyst disagreement.

Further, the test does not address the accuracy of cut mark morphometric identification since the correct score for each mark and individual analysts' scores are not reported. Neither can we evaluate whether analyst scoring differences suggest cut marks will be misidentified as BSM produced by other agents or processes, because these marks, particularly cut mark mimics (i.e. trampling, sediment abrasion), are not examined. The magnitude of inter-observer difference is expressed only in multivariate space, blurring how each analyst scored the 14 morphological

variables during the test. Providing raw scores for each analyst or summaries of the proportions of scores assigned for each variable, would allow further exploration of these trends.

The variable microabrasion highlights these problems. These very fine striae caused by sedimentary particles are common on trampled bone but not on experimental cut marks (see Domínguez-Rodrigo et al., 2009), and should therefore be rarely identified in the blind test. Yet we see a significant difference in scoring among the 11 analysts, meaning that some analysts scored them as present whereas others did not, and this discrepancy is impossible to attribute to inclusion of marks with atypical morphology in the blind test, or false-positive identification of morphometric criteria typical of trampling marks.

Additionally, some of the morphological criteria used by Domínguez-Rodrigo et al. (2017) to describe cut marks are prone to subjective scoring or measurement error. For example, the boundary between straight, curvy, or sinuous groove trajectory is unclear (p16:Table 1). Straight marks may be rare, especially given the curvature of most long bone portions, and the fact that the slightest observable curvature would prevent them from being scored as straight. All analysts showed similarity in identifying groove trajectory, but omitting analyst “F1”, whose experience and performance is most different from consensus, reintroduces divergent perceptions of trajectory. Similarly, mark orientation, (coded as parallel, perpendicular, or oblique relative to the main axis of the specimen) is the most commonly agreed upon variable in the blind test. But because oblique orientation includes a wide range of angles, unlike parallel and perpendicular orientations that are defined by relatively restricted angular orientation (0 and 90 degrees respectively), we expect most marks displayed oblique angles and were scored as such. Again, we do not know which variable states were the most commonly reported and how frequently or by how much individual analysts’ observations differed from the true state.

Divergent scoring of these macroscopically perceptible variables is minimal compared to variables that describe cross-sectional shape, the internal groove, and shoulder at magnification. Discrete variables like internal microstriations whose presence typically distinguishes cut marks from other BSM, are identified relatively similarly (especially after excluding analyst “F1”), but describing the morphometric criteria including the trajectory, shape, and location of microstriations, along with metric attributes like whether a v-shaped groove is narrow or wide, or whether the extent of flaking comprises exactly one third of the trajectory of the shoulder or length of the groove, yield much more disagreement. As Moretti et al. (2015:268) note, “directly observing cross sections of cut marks is not readily feasible using conventional 2D microscopy...because diagnostic criteria...of the cross sections and the slopes and floor of the grooves are difficult to calculate.”

#### **2.4 Inappropriate criticism of cut mark measurement**

Building on their blind test results, Domínguez-Rodrigo et al. (2017) discuss other experimental studies, concluding that most methods for measuring cut mark morphology are inaccurate. We argue this casts unnecessary doubt on the utility of microscopic BSM research in general, and respond to their criticism of Pante et al. (2017), by showing that our low inter-observer error and precise microscopic technique offer an accurate, repeatable measurement of BSM size and shape.

Pante et al. (2017) objectively measured the 3-D shape of BSM using a white light non-contact confocal profilometer and demonstrated high inter-observer replicability and near-perfect success in quantitative discrimination of stone tool cut marks and carnivore tooth marks. Domínguez-Rodrigo claim our experimental sample “is not big enough for solid statistical

analysis,” (19) but the 80 marks created by known effectors and agents more than doubles the Rorschach blind test sample, and is appropriate for the discriminant analysis used to quantify BSM morphometrics (Klecka, 1980, Friedman, 1989).

Domínguez-Rodrigo et al. (2017) suggest the scanning methodology developed by Pante et al. (2017) measures a derived mark shape because the visualization software fills gaps between points collected during image collection. With the 3mm optical pen, our scanning resolution was 40 nm in the z-axis (or 0.00025mm) which captures an accurate 3d mark shape and is too small to affect the discrimination between BSM, which vary on the micron-scale.

Additionally, Domínguez-Rodrigo et al. (2017) incorrectly claim our “average inter-analyst error was as high as 15%” (19). Three analysts re-measured 23% of experimental cut marks and 34% of experimental tooth marks to assess inter-observer error, which included variables collected from the 3-D studiables, the deepest profile, and the central profiles, and ranged from 4.7% to 35.9%. After examining which variables most effectively discriminate cut mark and tooth mark morphometrics, we excluded the less-repeatable central profile measurements, and achieved above 97% accuracy in correctly classifying marks to their known origin. We also evaluated the accuracy and inter-observer error and showed “no statistical differences between the medians of observer samples” across all variables (Pante et al., 2017:5).

Finally, Domínguez-Rodrigo et al. (2017) incorrectly suggest our profile measurements “are highly dependent on the protocol used for scanning each mark” (19). Confocal profilometry creates a precise, complete 3-D studiable (image) of each cut mark, objectively determines its deepest point and uses it to collect profile measurements, ensuring high repeatability with an instrument calibrated to National Institute of Standards and Technology (NIST) specifications. Most importantly, Pante et al. (2017) provide detailed instructions for measuring BSM attributes that allowed high inter-analyst correspondence, introducing the possibility of independent verification.

In contrast, new photogrammetric (Mate-Gonzalez et al., 2015, 2017) and structured light scanning (Yravedra et al., 2018) BSM measurement methods developed by this expert knowledge tradition (i.e. Domínguez-Rodrigo and colleagues) claim greater precision but include minimal methodological detail about experimental context or instrument precision, prohibiting replication or comparison with other results. These methods measure a series of arbitrarily placed 2-D cross-sectional profiles of metal knife cut marks intentionally incised into defleshed bone which are not directly comparable to maximum width or depth measurements on defleshing butchery cut marks created with replicated stone tools (see below). The repeatability of the z-axis measurement is not reported, which makes assessment of error difficult, and the authors concede that photogrammetry “might not be valid for the study of inconspicuous and vaguely defined marks (e.g. trampling) where the camera may lack enough resolution to capture...fine microscopic details” (Maté-González et al. (2017:366).

## **2.5 Immanent versus configurational properties in BSM experiments**

The Rorschach argument addresses how tool attributes including type, edge morphology and lithic raw material influence cut mark morphometrics but overlooks how the experimental context of cut mark creation – specifically, intentional incision versus defleshing butchery – might impact cut mark size and shape. Immanence involves fixed process-trace relationships, unlike configuration, where immanent processes produce contextually dependent results (Wolverton and Lyman, 2000). Regarding BSM, the tool edge during a slicing event and the negative shape of displaced material are in an immanent process-trace relationship. In a highly

controlled context (i.e. intentional tool slices on pine boards) (Greenfield, 1999, 2006), tool type, raw material, tool weight, slicing pressure, etc. may clearly affect cut mark shape, but these relationships may not apply to every configurational context.

Val et al. (2017) use experimental butchery to infer the type of tool used during Middle Paleolithic carcass butchery and demonstrate morphometric similarity in cut marks produced by unmodified flakes, retouched flake tools, and cleavers. Domínguez-Rodrigo et al. (2017) appropriately criticize the small butchery sample, however thorough methodological description allows replication and critical evaluation of butchery technique, specimen preparation, and analysis. Val et al. (2017) build realistic experiments – butchering wild fauna with tools made from lithic materials that occur archaeologically. Domínguez-Rodrigo et al. (2017) speculate that cut mark morphology was similar across tool types because all tool edges were mostly straight, and suggest lithic material may have impacted cut mark attributes. We argue cut mark similarity occurs because the immanent relationship between tool type and cut mark morphology, which may allow discrimination of tool type from cut mark attributes in some experimental instances, is not a configurational relationship that is generalizable to all butchery contexts. Experimental studies that successfully discriminate cut marks produced by different tool classes often measure intentional incisions into wooden boards (Greenfield, 1999, 2006), or intentionally incised single strokes into defleshed bone (Walker and Long, 1977, Bello and Soligo, 2008, de Juana et al., 2010, Mate-Gonzalez et al., 2016, Yravedra et al., 2017). When the configurational properties of butchery behavior, (e.g., the mechanics of tool slices necessary to perform different tasks like skinning, defleshing, and disarticulation, or bone density across anatomical portions) are built into experimental design, greater variability in cut mark morphology may prohibit successful discrimination of tool attributes or raw material (Merritt 2012, 2016.) Notably, studies that successfully identify tool attributes from cut mark morphometrics often compare intentionally incised marks (i.e., immanent traces) to marks incidentally created during butchery or build samples that combine contexts (Domínguez-Rodrigo et al., 2009, de Juana et al., 2010, Mate-Gonzalez et al., 2016).

### **3. BSM case studies**

#### *3.1 Materials and methods summary*

Here, we build a sample of previously published experimental trials that control for factors that impact cut mark cross-sectional size (Merritt, 2012, Pante et al., 2017). Differences in intentionally produced and realistic defleshing butchery marks are compared to archaeological cut marks with Kruskal-Wallis tests to determine whether observed morphometric variability is represented in the experimental samples. In addition, we analyze a sub-sample of cut marks that minimizes potential confounding effects of long bone portion density, tool type, animal size, taxon, and measurement technique, to examine whether lithic raw material is related to cross-sectional size of flake cut marks. These data are presented in Supplemental Tables 1 and 2. We use the Classification Learner App in MatLab 2016b (The MathWorks, Inc., Natick, Massachusetts) to test classification of 90 phonolite and 67 ignimbrite cut marks to causal raw material using cut mark width and depth. Classification models were trained using untransformed cut mark width and depth as predictors and lithic material (phonolite or ignimbrite) as the response. We employed 5-fold cross-validation to reduce overfitting. This method partitions data into five folds, trains a model with out-of-fold data, describes in-fold classification success, and summarizes model performance (i.e., correct classification) over all

folds. This iterative approach is appropriate for small samples, because it uses the complete dataset. We trained models appropriate for binary classification and small samples including a logistic regression model and a random forest classification model (Breimen, 2001) that used Gini's diversity index to determine node splitting rules and summarized a bootstrap-aggregated ensemble of 30 weak learner trees. Supplemental files 1 and 2 provide MatLab code for these models.

The realistic butchery sample (Table 1) included a series of experiments where a pastoralist experienced in butchery used unmodified phonolite and ignimbrite flakes to deflesh domestic goat forelimbs (humerus, radioulna, metacarpal) and hindlimbs (femur, tibia, metatarsal) without retouch or sharpening (Merritt, 2012). Cut marks from midshaft portions were selected to minimize bias in cross-sectional size introduced by differential bone density per portion and combined with twelve marks from the radial midshaft of a white-tailed deer defleshed with a basalt core by SRM (included in Pante et al., 2017). This sample is compared to intentionally incised cut marks produced by TLK with a chert flake and handaxe onto defleshed cow femoral and tibial midshafts. The archaeological sample comes from three Early Stone Age butchery localities at Koobi Fora, GaJi14, FwJj14N, and FwJj14S, which include large, well-preserved butchery assemblages (Merritt, 2017). The sample of 241 cut marks was measured from 19 specimens including a variety of long bones, limb girdles, and hyoids from size 2-4 mammals (Bunn, 1983). All marks occurred on midshaft and near-epiphyseal shaft portions or dense portions like the scapular neck, iliac ramus and hyoid ramus, and Weathering stage 0-1 cortical surfaces.

Experimental specimens were prepared via boiling water maceration, but different techniques were used to measure cut mark attributes. The 3-D measurement technique described by Pante et al. (2017) was used to measure deer and cow cut marks. Maximum width (accuracy 0.005mm) and maximum depth (accuracy 0.00004mm) were collected from the deepest profile scan of each mark using Mountains® software (Digital Surf, Besancon, France). These measurements are comparable across experimental samples and have low inter-observer error (5.1% and 4.9% respectively). The goat butchery and archaeological cut mark samples were molded with polyvinyl siloxane putty, each mark was sectioned perpendicular to its long axis at the widest observable point and measured for width and depth using a binocular microscope with 0.03125 mm accuracy (Merritt, 2012). With this less precise technique, repeated measurement of 72 randomly selected mark widths and depths yielded 23.5% measurement error, but overall 93% of repeated measurements differed by less than 0.09mm, and cut mark width and depth varied on the order of millimeters.

### *3.2. Case study 1: experimental context results*

Kruskal-Wallis tests indicate that median cut mark width and depth are significantly different across all experimental and archaeological contexts (width:  $X^2=129.63$ , d.f.=1,  $p<0.001$ ; depth:  $X^2=83.02$ , d.f.=1,  $p<0.001$ ) (Table 2), and intentionally produced marks have significantly wider and deeper median values than experimental butchery marks (Figure 1). The archaeological sample is significantly different in median width and depth compared to both experimental contexts, but the overlapping range suggests the experimental sample includes metric variability observed in archaeological butchery marks.

### *3.3. Case study 2: lithic raw material results*

Phonolite produced a significantly wider ( $X^2=9.98$ , d.f.=1,  $p=0.002$ ) and deeper ( $X^2=10.76$ , d.f.=1,  $p=0.001$ ) sample of cut marks than ignimbrite during goat defleshing trials according to the Kruskal-Wallis test (Figure 2). However, because ignimbrite cut mark depth varied little compared to the more dispersed sample of phonolite cut mark depth, the test returned a significant difference despite an equivalent median value.

A scatterplot shows overlapping mark width and depth across lithic materials (Figure 3). Classification of marks to causal tool material was poor (Table 3). A logistic regression model correctly classified 67.5% of marks to their known tool material based on mark width and depth and explained only a small amount of the deviance of tool material accounted for by these measurements (adjusted  $R^2=0.0794$ ). Random forest classification (Figure 4) successfully identified 58.6% of marks to known tool type.

### *3.4. Discussion of case studies*

We built an experimental cut mark sample created by different stone tool classes and raw materials on wild and domesticated animals of different size to explore how cut mark width and depth vary across butchery contexts. These results suggest that intentionally incised marks are significantly wider and deeper than realistic butchery marks, and casts doubt on experimental BSM studies that combine samples of intentionally created and realistic butchery marks to describe cut mark morphometrics. However, we note that tool raw material, animal size, and taxon are not evenly distributed across our butchery contexts, and future work with larger samples will explore how these potential confounding factors interact.

A sub-sample of realistic butchery marks on goat midshafts controls for the influence of animal size, tool type, and bone density, and demonstrates that ignimbrite and phonolite cut mark samples differ significantly in median width and depth, but overlapping mark size leads to poor classification of individual marks to raw material. This conclusion is based on a sample of marks that were measured with the same technique and analyzed with non-parametric methods that are appropriate for non-normally distributed data. Although we examine different tool materials and BSM variables than other studies that infer causal tool type or raw material from cut mark attributes, our results suggest that measurement subjectivity and poor experimental control during mark creation will confound inferences about causal raw material.

We suggest archaeological diagnosis of BSM first experimentally assess immanent properties, subsequently study process-trace relationships in a variety of configurational contexts, and ensure repeatability with blind testing on modern control samples. Meta-analysis on large samples can disentangle how multiple configurational factors contribute to cut mark morphometrics, but is contingent on clear contextual presentation.

## **4.1 Discussion**

As we have argued previously when identifying the agent responsible for fossil BSM and inferring the behavioral, ecological, and taphonomic contexts that influenced their creation, “diagnosis based solely on the micromorphology of isolated marks should be avoided. Rather contextual information is vital to the correct identification of past activities” (Njau, 2012:47). Here we extend this argument to include the context in which experimental samples are constructed.

Models approximate reality and their predictions include a trade-off between realism, precision, and generality (Levins, 1966, Capaldo, 1997, 1998). We appreciate the call for experimental design that minimizes equifinality, but criticize the Rorschach argument’s

reductionist discussion of analogy (Wylie, 2002:147). In fact, Bunge, who Domínguez-Rodrigo et al. (2017) cite extensively, recognizes the flexibility of analogies, noting that “the fruitfulness of analogical inference depends essentially upon the nature of the case, i.e., upon the substantive knowledge and imagination of the user” (Bunge, 1981:223). Because the archaeological variability of BSM traces and the contexts that created them are unknown, analysts should carefully evaluate the realism of their experimental design, publish clear and replicable methods and results which address the precision of model predictions and their generally applicability to archaeological inferences in a variety of temporal, geographic, behavioral, ecological, and social contexts.

#### **4.2 Expert knowledge, automation, and the argument from authority**

Amidst a crisis of reproducibility, many scientific disciplines recognize that standardization and critical examination of methodology underlie analytical consensus (MacLeod et al., 2010, Baker, 2016). Actualistic BSM models currently produce divergent evidence-based interpretations of the role of human behavior in the formation of the earliest potential butchery traces at Dikika and large Oldowan archaeofaunal assemblages like FLK *Zinjanthropus* (James and Thompson, 2015, Sahle et al., 2017). At their core, these arguments require accurate diagnosis of the causal processes and the behavior of taphonomic agents that created assemblages of BSM bearing specimens. Here the field finds itself at a crossroads where an expert knowledge approach – a set of skills for identifying and interpreting BSM developed by studying reference collections generated under controlled circumstances and passed between scholars in an academic tradition (e.g., Blumenshine et al., 1996), is being abandoned for methods that promise to eliminate subjectivity in BSM analysis. Machine learning, touted as the “dawn of artificial intelligence in taphonomy” (Domínguez-Rodrigo et al., 2017:22), identifies patterns in training datasets and classifies experimentally generated BSM to known causal contexts in order to establish morphometric predictive models (Arriaza and Domínguez-Rodrigo, 2016). Although explanatory in nature, Bayesian inferential models have also been recently used to classify experimental BSM based on morphometrics (Harris et al., 2017, Otarolla-Castillo et al. 2018, Shmueli, 2010). These algorithmic methods reflect different predictive and explanatory approaches, and both may provide powerful solutions for classification in BSM datasets that include many observations, multiple variables, and different scales of measurement. Still, these sophisticated quantitative methods are applied to data collected by researchers that directly observe BSM attributes and make analytical decisions when automating measurement. Algorithmic classification is not a new approach to reducing the uncertainty introduced by human subjective observation in large, complex datasets. In fact, the expert systems approach that foreshadowed machine learning (Lagrange and Renaud, 1985, Wilcock, 1985, Doran, 1988, Garson, 1990, van der Maaten et al. 2006) recognized that “information does not exist in the world waiting to be extracted by a robot, but, rather, it should be situated in meaningful contexts” (Barcelo, 2009:106). Therefore, we argue that actualistic BSM research requires control over experimental confounders, discussion of measurement accuracy and repeatability, and when possible, published primary data. This contextual presentation will allow critical review of analytical design (i.e., preproducibility (Stark, 2018)), and promote replication.

We suggest the Rorschach blind test suffers from these analytical problems because it measures subjective attributes on a BSM sample generated without much experimental control, and rather than revise morphometric BSM definitions or measurement techniques in light of inter-analyst discordance, it searches idealistically for objective classification while de-

emphasizing the impact of the human decisions about sample construction and algorithmic methods (Domínguez-Rodrigo and Baquedano, 2018).

Moreover, the Rorschach argument's vague actualistic description and lack of primary data prohibits replication. Our results do not settle the methodological debate about BSM measurement but suggest presentation of experimental context is required for disciplinary progress. We are currently applying our precise and repeatable scanning technique to build a reference sample of over 700 "bone surface modifications inflicted by [diverse] actors and effectors in controlled settings" including multiple mammalian carnivore and crocodile tooth marks, large ungulate trampling marks, and cut and percussion marks from different raw materials and tool types (Pante et al. 2017:10). We continue to advocate extensive study of controlled reference collections as the primary method for training analysts and developing new analytical tools, as well as for making initial diagnoses of marks on fossil assemblages that can be compared to and potentially integrated into automated algorithm-based methods. Requests to study our BSM imagery or specimens should be directed to the primary author.

## **5. Conclusion**

Models perform simulations of reality, often employing sophisticated analyses and visualizations, and gain credibility from discursive networks where language and rhetoric operate to build confidence. Therefore "careful attention should be paid to the representations of certainty, uncertainty, and ignorance in such communications" (Hulme, 2013:37). Here, the major flaws of the Rorschach effect argument serve as cautionary notes for future BSM research. The lack of clear context, including the experimental design of butchery trials and analytical methods used to explore their results prevents the scientific community from evaluating the veracity of the work, and commits the fallacy of argument from authority. Further, the reductionist discussion of perfect experimental analogues casts unnecessary doubt on the utility of actualistic work in general. Experimental analogues can never be perfect re-creations of the past, but they should be clearly described so that the realism, generality, and precision of model predictions and their applicability to archaeological cases can be assessed (Levins, 1966, Capaldo, 1997, 1998).

We conclude that BSM analysis is not a subjective art – it is a science whose inferences about the operation of ancient systems are compromised when vague contextual presentation hides poorly controlled data collection and inappropriate analytical methods. Given the difficulties of building large actualistic BSM datasets, meta-analysis may be a productive way to integrate results from diverse experimental studies, but ultimately, transposing knowledge from an authoritative expert tradition to a community of scholars who can reach scientific consensus requires clear presentation (Ehm, 2016, Lakens et al., 2016).

## **6. Acknowledgments**

MCP thanks the Department of Anthropology and College of Liberal Arts, Colorado State University for funding part of this research. We are also grateful to the anonymous reviewers whose comments strengthened this manuscript.

## **Figure and Table captions**

Figure 1. Cut mark width (left) and depth (right) for intentionally incised, experimental butchery, and archaeological samples. The notch shows the 95% confidence interval around the median line, and is a visual representation of the Kruskal-Wallis test.

Figure 2. Cut mark width (left) and depth (right) for phonolite and ignimbrite flake butchery samples. The notch shows the 95% confidence interval around the median line, and is a visual representation of the Kruskal-Wallis test.

Figure 3. Cut mark width versus depth (both measured in millimeters) for phonolite and ignimbrite flake butchery sample.

Figure 4. Classification tree results built from 30 weak learner trees in a random forest model using Gini's diversity index to determine node splitting rules. Width (sw) and depth are measured in millimeters. This model only classified 58.6% of marks to the correct tool type.

Table 1. Experimental butchery trials examined for cut mark width and depth

Table 2. Summary of cut mark width and depth according to experimental and archaeological context

Table 3. Confusion matrices for cut mark classification

Supplemental Table 1. Experimental cut mark data

Supplemental Table 2. Archaeological cut mark data from Koobi Fora, Kenya

Supplemental File 1. Merritt\_tool\_material\_logistic\_regression\_classifier.m

Supplemental File 2. Merritt\_tool\_material\_bagged\_tree\_classifier.m

## **References**

- Arriaza, M. C., & Domínguez-Rodrigo, M. (2016). When felids and hominins ruled at Olduvai Gorge: A machine learning analysis of the skeletal profiles of the non-anthropogenic Bed I sites. *Quaternary Science Reviews*, *139*, 43–52.  
<https://doi.org/10.1016/j.quascirev.2016.03.005>
- Atmanspacher, H., & Maasen, S. (2016). *Reproducibility: Principles, Practices, and Prospects*. Hoboken: Wiley.
- Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452–454.
- Barceló, J. A. (2009). The birth and historical development of computational intelligence applications in archaeology. *Archeologia E Calcolatori*, *20*, 95–109.
- Bello, S. M., Wallduck, R., Parfitt, S. A., & Stringer, C. B. (2017). An Upper Palaeolithic engraved human bone associated with ritualistic cannibalism. *PLoS ONE*, *12*(8), e0182127.  
<https://doi.org/https://doi.org/10.1371/journal.pone.0182127>
- Bello, S. M., & Soligo, C. (2008). A new method for the quantitative analysis of cutmark micromorphology. *Journal of Archaeological Science*, *35*(6), 1542–1552.  
<https://doi.org/10.1016/j.jas.2007.10.018>
- Blumenschine, R. J., Marean, C. W., & Capaldo, S. D. (1996). Blind Tests of Inter-analyst Correspondence and Accuracy in the Identification of Cut Marks, Percussion Marks, and Carnivore Tooth Marks on Bone Surfaces. *Journal of Archaeological Science*, *23*(4), 493–507. <https://doi.org/10.1006/jasc.1996.0047>
- Braun, D. R., Pante, M., & Archer, W. (2016). Cut marks on bone surfaces: influences on variation in the form of traces of ancient behaviour. *Interface Focus*, *6*, 20160006.  
<https://doi.org/10.1098/rsfs.2016.0006>
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.  
<https://doi.org/10.1023/A:1010933404324>
- Bunge, M. (1981). Analogy between systems. *International Journal of General Systems*, *7*.
- Bunn, H. T. (1983). Comparative analysis of modern bone assemblages from a San hunter-gatherer camp in the Kalahari Desert, Botswana, and from a spotted hyena den near Nairobi, Kenya. In J. Clutton-Brock & C. Grigson (Eds.), *Animals and Archaeology, Hunters and Their Prey, vol. 1, British Archaeological Reports International Series 163* (pp. 143–148). London.

- Capaldo, S. D. (1997). Experimental determinations of carcass processing by Plio-Pleistocene hominids and carnivores at FLK 22 (Zinjanthropus). Olduvai Gorge, Tanzania. *Journal of Human Evolution*, 33(5), 555–97. <https://doi.org/10.1006/jhev.1997.0150>
- Capaldo, S. D. (1998). Simulating the Formation of Dual-Patterned Archaeofaunal Assemblages with Experimental Control Samples. *Journal of Archaeological Science*, 25(4), 311–330. <https://doi.org/10.1006/jasc.1997.0238>
- Dambricourt Malassé, A., Moigne, A. M., Singh, M., Calligaro, T., Karir, B., Gaillard, C., ... Garcia Sanz, M. (2016). Intentional cut marks on bovid from the Quranwala zone, 2.6 Ma, Siwalik Frontal Range, northwestern India. *Comptes Rendus - Palevol*, 15(3–4), 317–339. <https://doi.org/10.1016/j.crpv.2015.09.019>
- de Juana, S., Galán, A. B., & Domínguez-Rodrigo, M. (2010). Taphonomic identification of cut marks made with lithic handaxes: An experimental study. *Journal of Archaeological Science*, 37(8), 1841–1850. <https://doi.org/10.1016/j.jas.2010.02.002>
- Domínguez-Rodrigo, M. (1997). Meat-eating by early hominids at the FLK 22 Zinjanthropus site, Olduvai Gorge (Tanzania): an experimental approach using cut-mark data. *Journal of Human Evolution*, 33(6), 669–90. <https://doi.org/10.1006/jhev.1997.0161>
- Domínguez-Rodrigo, M., & Barba, R. (2005). A study of cut marks on small-sized carcasses and its application to the study of cut-marked bones from small mammals at the FLK Zinj site. *Journal of Taphonomy*, 3(3), 121–134.
- Domínguez-Rodrigo, M., de Juana, S., Galán, A. B., & Rodríguez, M. (2009). A new protocol to differentiate trampling marks from butchery cut marks. *Journal of Archaeological Science*, 36(12), 2643–2654. <https://doi.org/10.1016/j.jas.2009.07.017>
- Domínguez-Rodrigo, M., Pickering, T. R., & Bunn, H. T. (2010). Configurational approach to identifying the earliest hominin butchers. *Proceedings of the National Academy of Sciences of the United States of America*, 107(49), 20929–20934. <https://doi.org/10.1073/pnas.1013711107>
- Domínguez-Rodrigo, M., Pickering, T. R., & Bunn, H. T. (2012). Experimental study of cut marks made with rocks unmodified by human flaking and its bearing on claims of ~3.4-million-year-old butchery evidence from Dikika, Ethiopia. *Journal of Archaeological Science*, 39(2), 205–214. <https://doi.org/10.1016/j.jas.2011.03.010>
- Domínguez-Rodrigo, M., Bunn, H. T., & Yravedra, J. (2014). A critical re-evaluation of bone surface modification models for inferring fossil hominin and carnivore interactions through a multivariate approach: Application to the FLK Zinj archaeofaunal assemblage (Olduvai Gorge, Tanzania). *Quaternary International*, 322–323, 32–43. <https://doi.org/10.1016/j.quaint.2013.09.042>

- Domínguez-Rodrigo, M., Saladi, P., Aceres, I. C., Huguet, R., Yravedra, J. E., Rodríguez-Hidalgo, A., ... Cobo-Sánchez, L. (2017). Use and abuse of cut mark analyses: The Rorschach effect. *Journal of Archaeological Science*, 86, 14–23. <https://doi.org/10.1016/j.jas.2017.08.001>
- Domínguez-Rodrigo, M., & Baquedano, E. (2018). Distinguishing butchery cut marks from crocodile bite marks through machine learning methods. *Scientific Reports*, 8(1), 1–8. <https://doi.org/10.1038/s41598-018-24071-1>
- Doran, J. (1988). Expert systems and archaeology: what lies ahead? *Computer and Quantitative Methods in Archaeology 1987 (BAR International Series 393)*, 237–241.
- Ehm, W. (2016). Reproducibility from the Perspective of Meta-Analysis. In H. Atmanspacher & S. Maasen (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects* (pp. 141–167). Wiley. Hoboken.
- Friedman, J. H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84, 165–175.
- Garson, G. D. (1990). Expert Systems: An Overview for Social Scientists. *Social Science Computer Review*, 8(3), 387–410. <https://doi.org/10.1177/089443939000800304>
- Gidna, A. O., Kisui, B., Mabulla, A., Musiba, C., & Domínguez-Rodrigo, M. (2014). An ecological neo-taphonomic study of carcass consumption by lions in Tarangire National Park (Tanzania) and its relevance for human evolutionary biology. *Quaternary International*, 322–323, 167–180. <https://doi.org/10.1016/j.quaint.2013.08.059>
- Gidna, A., Domínguez-Rodrigo, M., & Pickering, T. R. (2015). Patterns of bovid long limb bone modification created by wild and captive leopards and their relevance to the elaboration of referential frameworks for paleoanthropology. *Journal of Archaeological Science: Reports*, 2, 302–309. <https://doi.org/10.1016/j.jasrep.2015.03.003>
- Gifford-Gonzalez, D. (1991). Bones are not enough: Analogues, knowledge, and interpretive strategies in zooarchaeology. *Journal of Anthropological Archaeology*, 10(3), 215–254. [https://doi.org/10.1016/0278-4165\(91\)90014-O](https://doi.org/10.1016/0278-4165(91)90014-O)
- Greenfield, H. J. (2006). Slicing cut marks on animal bones: diagnostics for identifying stone tool type and raw material. *Journal of Field Archaeology*, 31(2), 147–163.
- Greenfield, H. J. (1999). The Origins of Metallurgy: Distinguishing Stone from Metal Cut-marks on Bones from Archaeological Sites. *Journal of Archaeological Science*, 26, 797–808. <https://doi.org/10.1006/jasc.1998.0348>
- Harris, J. A., Marean, C. W., Ogle, K., & Thompson, J. (2017). The trajectory of bone surface modification studies in paleoanthropology and a new Bayesian solution to the identification

controversy. *Journal of Human Evolution*, 110, 69–81.  
<https://doi.org/10.1016/j.jhevol.2017.06.011>

Holen, S. R., Deméré, T. A., Fisher, D. C., Fullagar, R., Paces, J. B., Jefferson, G. T., ... Holen, K. A. (2017). A 130,000-year-old archaeological site in southern California, USA. *Nature*, 544(27 April 2017), 479–483. <https://doi.org/10.1038/nature22065>

Hulme, M. (2013). How climate models gain and exercise authority. In K. Hastrup & M. Skrydstrup (Eds.), *The Social life of climate change models: anticipating nature* (pp. 30–44). New York: Routledge.

James, E. C., & Thompson, J. C. (2015). On bad terms: Problems and solutions within zooarchaeological bone surface modification studies. *Environmental Archaeology*, 20(1), 89–103. <https://doi.org/10.1179/1749631414Y.0000000023>

Klecka, W.R. (1980). *Discriminant Analysis*. Beverly Hills, CA: Sage Publications, Inc.

Lagrange, M. S., & Renaud, M. (1985). Intelligent knowledge-based systems in archaeology: A computerized simulation of reasoning by means of an expert system. *Computers and the Humanities*, 19(1), 37–52. <https://doi.org/10.1007/BF02259616>

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4(24), 1–10. <https://doi.org/10.1186/s40359-016-0126-3>

Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist*.

MacLeod, N., Benfield, M., & Culverhouse, P. (2010). Time to automate identification. *Nature*, 467(7312), 154–155. <https://doi.org/10.1038/467154a>

Mate Gonzalez, M. A., Yravedra, J., Gonzalez-Aguilera, D., Palomeque-Gonzalez, J. F., & Domínguez-Rodrigo, M. (2015). Micro-photogrammetric characterization of cut marks on bones. *Journal of Archaeological Science*, 62, 128–142.  
<https://doi.org/10.1016/j.jas.2015.08.006>

Maté-González, M. Á., Palomeque-González, J. F., Yravedra, J., González-Aguilera, D., & Domínguez-Rodrigo, M. (2016). Micro-photogrammetric and morphometric differentiation of cut marks on bones using metal knives, quartzite, and flint flakes. *Archaeological and Anthropological Sciences*. <https://doi.org/10.1007/s12520-016-0401-5>

Mate-Gonzalez, M., Aramendi, J., Yravedra, J., Rosell, J., Gonzalez-Aguilera, D., & Domínguez-Rodrigo, M. (2017). Assessment of statistical agreement of three techniques for the study of cut marks: 3D digital microscope, laser scanning confocal microscopy and micro-photogrammetry. *Journal of Microscopy*, 00(0), 1–15.  
<https://doi.org/10.1111/jmi.12575>

MatLab 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.  
<https://www.mathworks.com>.

McPherron, S. P., Alemseged, Z., Marean, C. W., Wynn, J. G., Reed, D., Geraads, D., ... Béarat, H. A. (2010). Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. *Nature*, 466(7308), 857–860.  
<https://doi.org/10.1038/nature09248>

Merritt, S. R. (2011). *Controlled butchery observations as a means for interpreting Okote Member hominin carnivory at Koobi Fora, Kenya*. Rutgers University. Retrieved from <http://search.proquest.com/docview/859254836>

Merritt, S. R. (2012). Factors affecting Early Stone Age cut mark cross-sectional size: Implications from actualistic butchery trials. *Journal of Archaeological Science*, 39(9), 2984–2994. <https://doi.org/10.1016/j.jas.2012.04.036>

Merritt, S. R. (2016). Cut Mark Cluster Geometry and Equifinality in Replicated Early Stone Age Butchery. *International Journal of Osteoarchaeology*, 26(4), 585–598.  
<https://doi.org/10.1002/oa.2448>

Merritt, S. R. (2017). Investigating hominin carnivory in the Okote Member of Koobi Fora, Kenya with an actualistic model of carcass consumption and traces of butchery on the elbow. *Journal of Human Evolution*, 112, 105–133.  
<https://doi.org/10.1016/j.jhevol.2017.08.004>

Moretti, E., Arrighi, S., Boschin, F., Crezzini, J., Aureli, D., & Ronchitelli, A. (2015). Using 3D microscopy to analyze experimental cut marks on animal bones produced with different stone tools. *Ethnobiology Letters*, 6(2), 267–275. <https://doi.org/10.14237/eb1.6.2.2015.349>

Mountains® surface imaging & metrology software ‘Digital Surf: Mountain- sMap®Premium ©2015’, <http://www.digitalsurf.fr/en/mntkey.html>.

Muttart, Njau, J. K. (2012). Reading Pliocene bones. *Science*, 336(April), 46–48.  
<https://doi.org/10.1126/science.1216221>

Njau, J. K., & Blumenschine, R. J. (2006). A diagnosis of crocodile feeding traces on larger mammal bone, with fossil examples from the Plio-Pleistocene Olduvai Basin, Tanzania. *Journal of Human Evolution*, 50(2), 142–162. <https://doi.org/10.1016/j.jhevol.2005.08.008>

Organista, E., Pernas-Hernandez, M., Gidna, A., Yravedra, J., & Domínguez-Rodrigo, M. (2016). An experimental lion-to-hammerstone model and its relevance to understand hominin-carnivore interactions in the archeological record. *Journal of Archaeological Research*, 66, 69–77. <https://doi.org/10.1016/j.jas.2015.12.004>

Otárola-Castillo, E., Torquato, M. G., Hawkins, H. C., James, E., Harris, J. A., Marean, C. W., ... Thompson, J. C. (2018). Differentiating between cutting actions on bone using 3D

- geometric morphometrics and Bayesian analyses with implications to human evolution. *Journal of Archaeological Science*, 89, 56–67. <https://doi.org/10.1016/j.jas.2017.10.004>
- Pante, M. C., Scott, R. S., Blumenshine, R. J., & Capaldo, S. D. (2015). Revalidation of bone surface modification models for inferring fossil hominin and carnivore feeding interactions. *Quaternary International*, 355, 164–168. <https://doi.org/10.1016/j.quaint.2014.09.007>
- Pante, M. C., Muttart, M. V., Keevil, T. L., Blumenshine, R. J., Njau, J. K., & Merritt, S. R. (2017). A new high-resolution 3-D quantitative method for identifying bone surface modifications with implications for the Early Stone Age archaeological record. *Journal of Human Evolution*, 102, 1–11. <https://doi.org/10.1016/j.jhevol.2016.10.002>
- Parkinson, J. A., Plummer, T., & Hartstone-Rose, A. (2015). Characterizing felid tooth marking and gross bone damage patterns using GIS image analysis: An experimental feeding study with large felids. *Journal of Human Evolution*, 80, 114–134. <https://doi.org/10.1016/j.jhevol.2014.10.011>
- Pobiner, B. L., Higson, C. P., Kovarovic, K., Kaplan, R. S., Rogers, J., & Schindler, W. (2018). Experimental butchery study investigating the influence of timing of access and butcher expertise on cut mark variables. *International Journal of Osteoarchaeology*, (March), 1–11. <https://doi.org/10.1002/oa.2661>
- Rots, V., Hayes, E., Cnuts, D., Lepers, C., & Fullagar, R. (2016). Making sense of residues on flaked stone artefacts: Learning from blind tests. *PLoS ONE*, 11(3), e0150437. <https://doi.org/10.1371/journal.pone.0178311>
- Sahle, Y., El Zaatari, S., & White, T. D. (2017). Hominid butchers and biting crocodiles in the African Plio–Pleistocene. *Proceedings of the National Academy of Sciences*, 1–6. <https://doi.org/10.1073/pnas.1716317114>
- Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
- Stark, P. B. (2018). Before reproducibility must come preproducibility. *Nature*, 557(7707), 613. <https://doi.org/10.1038/d41586-018-05256-0>
- Thompson, J. C., McPherron, S. P., Bobe, R., Reed, D., Barr, W. A., Wynn, J. G., ... Alemseged, Z. (2015). Taphonomy of fossils from the hominin-bearing deposits at Dikika. *Journal of Human Evolution*, 86, 112–135. <https://doi.org/10.1016/j.jhevol.2015.06.013>
- Val, A., Costamagno, S., Discamps, E., Chong, S., Claud, E., Deschamps, M., ... Thiébaud, C. (2017). Testing the influence of stone tool type on microscopic morphology of cut-marks: Experimental approach and application to the archaeological record with a case study from the Middle Palaeolithic site of Noisetier Cave (Fréchet-Aure, Hautes-Pyrénées, France). *Journal of Archaeological Science: Reports*, 11, 17–28. <https://doi.org/10.1016/j.jasrep.2016.11.028>

- van der Maaten, L., Boon, P., Lange, G., Pajmans, H., & Postma, E. (2006). Computer vision and machine learning for archaeology. *Proceedings of Computer Applications and Quantitative Methods in Archaeology*, 1–7. Retrieved from [http://proceedings.caaconference.org/files/2006/33\\_Maaten\\_et\\_al\\_CAA2006.pdf](http://proceedings.caaconference.org/files/2006/33_Maaten_et_al_CAA2006.pdf)
- Wadley, L., & Lombard, M. (2007). Small things in perspective: the contribution of our blind tests to micro-residue studies on archaeological stone tools. *Journal of Archaeological Science*, 34(6), 1001–1010. <https://doi.org/10.1016/j.jas.2006.09.016>
- Walker, P. L., & Long, J. C. (1977). An Experimental Study of the Morphological Characteristics of Tool Marks. *American Antiquity*, 42(4), 605–616.
- Wilcock, J. (1985). A review of expert systems: their shortcomings and possible applications in archaeology. *Computer Applications in Archaeology*, 13, 139–144.
- Wolverton, S. (2013). Data Quality in Zooarchaeological Faunal Identification. *Journal of Archaeological Method and Theory*, 20, 381–396. <https://doi.org/10.1007/s10816-012-9161-4>
- Wolverton, S., & Lyman, R. L. (2000). Immanence and Configuration In Analogical Reasoning. *North American Archaeologist*, 21(3), 233–247. <https://doi.org/10.2190/QYCW-11QX-THNV-5E6R>
- Wylie, A. (2002). *Thinking from things: essays in the philosophy of archaeology*. Berkley: University of California Press.
- Yravedra, J., Maté-González, M. Á., Palomeque-González, J. F., Aramendi, J., Estaca-Gómez, V., San Juan Blazquez, M., ... Domínguez-Rodrigo, M. (2017). A new approach to raw material use in the exploitation of animal carcasses at BK (Upper Bed II, Olduvai Gorge, Tanzania): a micro-photogrammetric and geometric morphometric analysis of fossil cut marks. *Boreas*. <https://doi.org/10.1111/bor.12224>
- Yravedra, J., Aramendi, J., Maté-González, M. Á., Courtenay, L. A., & González-Aguilera, D. (2018). Differentiating percussion pits and carnivore tooth pits using 3D reconstructions and geometric morphometrics. *PLoS ONE*, 13(3), 1–18. <https://doi.org/10.1371/journal.pone.0194324>

Table 1. Experimental butchery trials examined for cut mark width and depth

Trial	Context	Taxon	Side	Limb	Tool Type	Tool	Material	Cut Mark location	Cut Mark Count	Original reference
Tr3	defleshing butchery	goat	left	forelimb	flake	Tr3	phonolite	long bone midshafts	5	Merritt, 2012
Tr3	defleshing butchery	goat	left	hindlimb	flake	Tr3	phonolite	long bone midshafts	11	Merritt, 2012
IB10	defleshing butchery	goat	left	forelimb	flake	F92	phonolite	long bone midshafts	27	Merritt, 2012
IB10	defleshing butchery	goat	left	hindlimb	flake	F92	phonolite	long bone midshafts	47	Merritt, 2012
IB9	defleshing butchery	goat	right	forelimb	flake	F71	ignimbrite	long bone midshafts	23	Merritt, 2012
IB9	defleshing butchery	goat	right	hindlimb	flake	F72	ignimbrite	long bone midshafts	44	Merritt, 2012
D1	defleshing butchery	deer	???	forelimb	core	D1	basalt	radius midshaft	12	Pante et al., 2017
C1	intentional mark creation	cow	???	hindlimb	flake	C1	chert	femur midshaft	19	Pante et al., 2017
C2	intentional mark creation	cow	???	hindlimb	handaxe	C2	chert	femur midshaft	20	Pante et al., 2017

Table 2. Summary of cut mark width and depth according to experimental and archaeological context

context	animal	tool	raw material	mark count		width (mm)	depth (mm)
butchery	goat	flake	phonolite	90	median	0.156	0.031
					maximum	0.406	0.125
					minimum	0.094	0.031
					Lilliefors test <sup>a</sup>	<b>p&lt;0.001, k=0.2534</b>	<b>p&lt;0.001, k=0.4206</b>
butchery	goat	flake	ignimbrite	67	median	0.125	0.031
					maximum	0.500	0.063
					minimum	0.063	0.031
					Lilliefors test <sup>a</sup>	<b>p&lt;0.001, k=0.2456</b>	<b>p&lt;0.001, k=0.5327</b>
butchery	deer	core	basalt	12	median	0.225	0.065
					maximum	0.465	0.089
					minimum	0.157	0.038
					Lilliefors test <sup>a</sup>	p=0.238, k=0.194	p=0.196, k=0.201
intentional	cow	flake	chert	19	median	0.290	0.082
					maximum	0.740	0.188
					minimum	0.095	0.031
					Lilliefors test <sup>a</sup>	p=0.269, k=0.154	p=0.500, k=0.126
intentional	cow	handaxe	chert	20	median	0.443	0.077
					maximum	1.140	0.174
					minimum	0.100	0.026
					Lilliefors test <sup>a</sup>	p=0.278, k=0.149	p=0.221, k=0.156
Butchery mark sample				169	median	0.125	0.031
					maximum	0.500	0.125
					minimum	0.063	0.031
					Lilliefors test <sup>a</sup>	<b>p&lt;0.001, k=0.211</b>	<b>p&lt;0.001, k=0.437</b>
Intentional mark sample				39	median	0.340	0.082
					maximum	1.140	0.188
					minimum	0.095	0.026
					Lilliefors test <sup>a</sup>	<b>p=0.016, k=0.157</b>	p=0.248, k=0.112
Archaeological mark sample				241	median	0.219	0.063
					maximum	1.500	0.500
					minimum	0.938	0.031
					Lilliefors test <sup>a</sup>	<b>p&lt;0.001, k=0.240</b>	<b>p&lt;0.001, k=0.327</b>

a. The Lilliefors test returns a significant result when samples are not normally distributed. Non-normal samples are highlighted in bold.

Table 3. Confusion matrices for cut mark classification<sup>a</sup>

Logistic Regression Model			
		Predicted class	
		Ignimbrite	Phonolite
True class	Ignimbrite	66% (44)	34% (23)
	Phonolite	31% (28)	69% (62)
Random Forest, Weak Learner Model			
		Predicted class	
		Ignimbrite	Phonolite
True class	Ignimbrite	66% (44)	34% (23)
	Phonolite	47% (42)	53% (48)

a. Cell values list the percentage of marks classified in each category, along with the number of marks in parentheses.

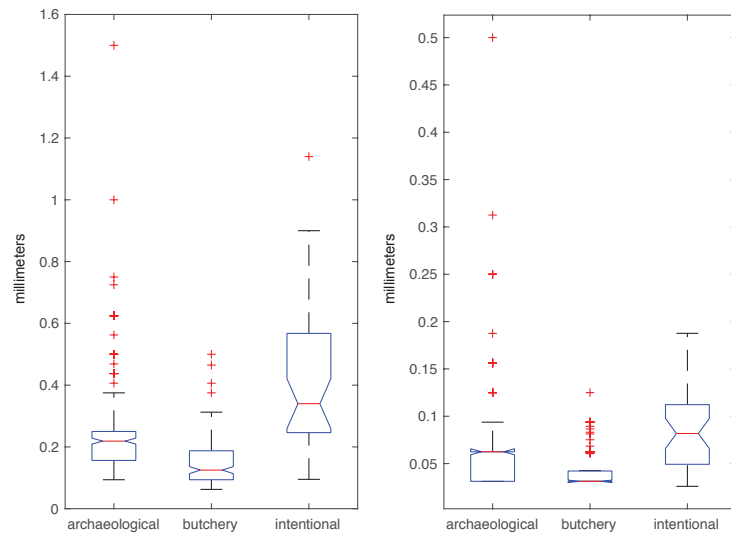


Figure 1. Cut mark width (left) and depth (right) for intentionally incised, experimental butchery, and archaeological samples. The notch shows the 95% confidence interval around the median line, and is a visual representation of the Kruskal-Wallis test.

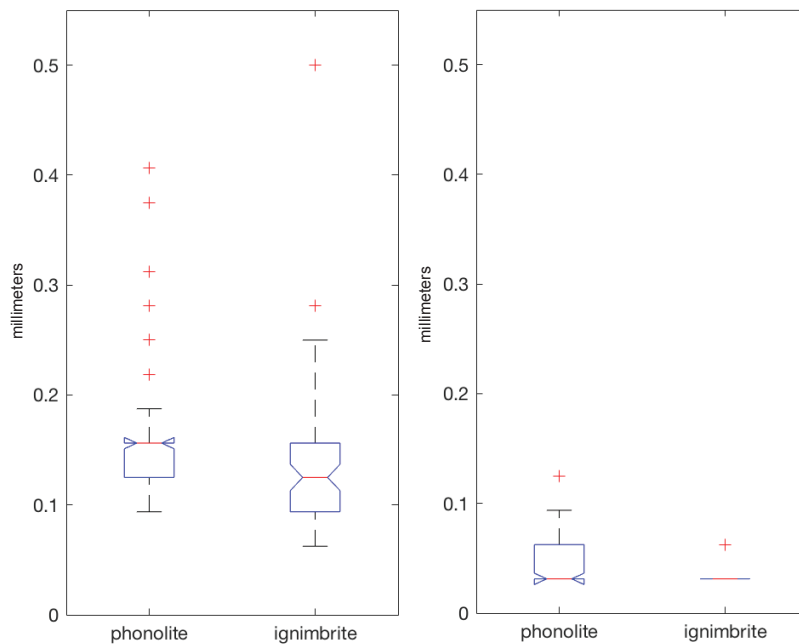


Figure 2. Cut mark width (left) and depth (right) for phonolite and ignimbrite flake butchery samples. The notch shows the 95% confidence interval around the median line, and is a visual representation of the Kruskal-Wallis test.

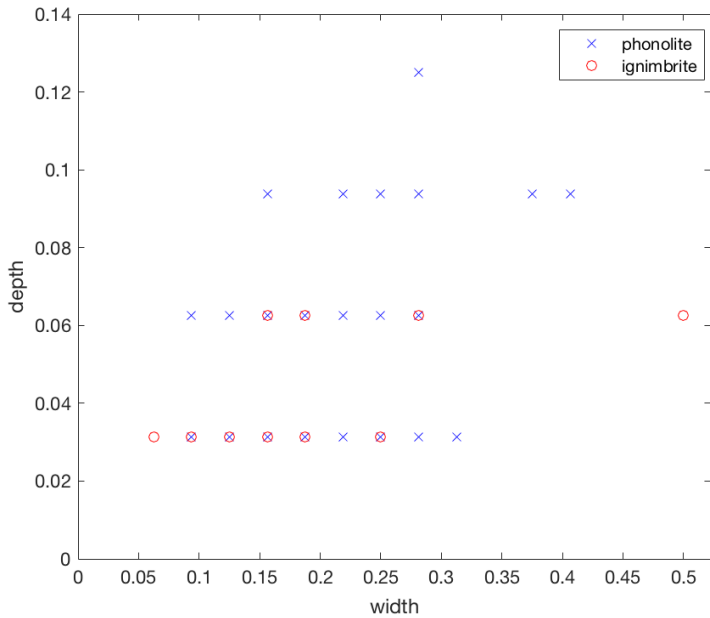


Figure 3. Cut mark width versus depth (both measured in millimeters) for phonolite and ignimbrite flake butchery sample.

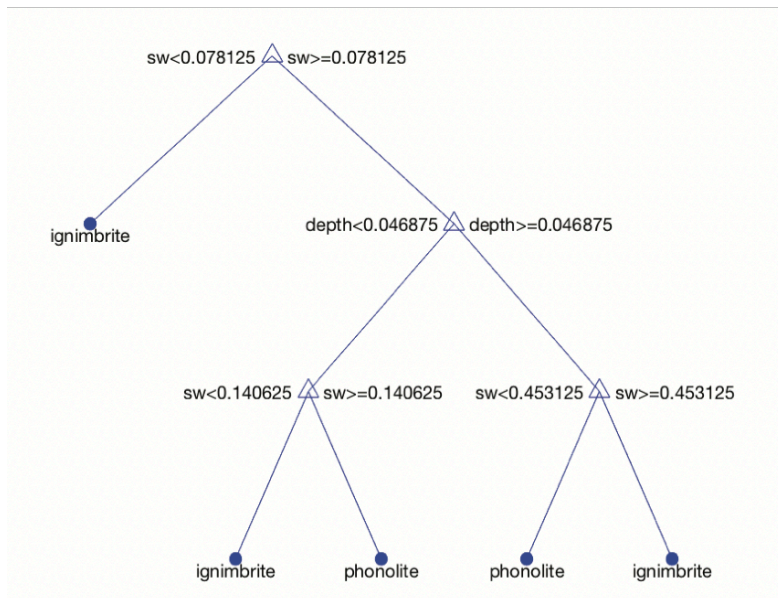


Figure 4. Classification tree results built from 30 weak learner trees in a random forest model using Gini's diversity index to determine node splitting rules. Width (sw) and depth are measured in millimeters. This model only classified 58.6% of marks to the correct tool type.