Bones of contention: a double blind study of experts' ability to classify sheep and goat astragali from images

Ilkka M. V. Sipilä[1]*, James Steele[1], Luke Dickens[2], Louise Martin[1]

Abstract

In zooarchaeology, animal bones are normally identified using comparative macro-morphological methods, which involve visual comparison of the bone with reference materials. However, recent work has oppugned the reliability of these methods. Although previous studies applying macro-morphological methods to identify sheep and goat bones have found low error rates, these results are based on small numbers of analysts, large numbers of different bone types and do not properly account for ambiguous "sheep/goat" classifications.

We present an extensive blind study of performance and reliability for binary macro-morphological species identification using just the astragalus. Each participant made independent comparative identifications on a random subset, including repeat presentations for consistency analysis. No sheep/goat category was offered. Instead, participants reported confidence scores on each sample. The participants also reported the reference materials used and indicated their regions of attention in each image.

Findings indicate that neither the use of reference materials nor experience are good predictors of accuracy, although more experienced analysts are found to be more consistent. Forcing binary classifications leads to a more transparent analysis but indicates lower performance scores than reported elsewhere, while corresponding confidence scores positively correlate with accuracy. Qualitative analysis of reported attention regions indicate that mistakes can occur when there is an overlap in the morphologies of the two species. We conclude that overreliance on reference materials impacts performance when the morphology of reference materials is not representative

[1] UCL Institute of Archaeology, 31-34 Gordon Square, London WC1H 0PY

[2] UCL Department of Information Studies, Gower Street, London WC1E 6BT

* Corresponding author: ilkkasipila@outlook.com

of the population variance, which is especially evident when the wider bone morphology is not adequately integrated into the classification decision.

Keywords

Blind study, Zooarchaeology, Sheep and goat separation, Expert ability

Acknowledgments

We would like to thank Kone Foundation (Koneen Säätiö) for providing funding for the first author's doctoral thesis and this present article. We want to acknowledge the contributions made by researchers from the Archaeology department at University of Sheffield, National Museum Wales, and Historic England. We especially want to thank Dr Polydora Baker for her help with the zooarchaeology collection at Fort Cumberland, Jennifer Gallichan for her help with the zooarchaeology collection at National Museum Wales, and Dr Angelos Hadjikoumis and Prof Umberto Albarella for their help with the zooarchaeology reference collection at the University of Sheffield. Finally, we thank those 39 anonymous participants who completed the blind study and without whose volunteering this study would not have been possible. The presented work is an adaptation of the work presented in the first author's doctoral thesis.

Introduction

Many zooarchaeologically relevant blind studies have been published since Driver's (1992) discussion of the problems in zooarchaeological identifications. These blind studies have covered varying topics and they have varying aims, including: 1) the comparison of laboratory-based techniques and comparative methods of identification (Greenlee and Dunnell 2010; Welker et al. 2015; Pilaar Birch et al. 2019; Prendergast et al. 2019); 2) the exploration of analysts' interpretations and methodological decision making when given the same assemblage (Gobalet 2001; Atici et al. 2013; Giovas et al. 2017); 3) the assessment of published criteria (Fernandez 2001; Zeder and Lapham 2010; Zeder and Pilaar 2010; Twiss et al. 2017); 4) the reproducibility of an assemblage (Nims and Butler 2017; Lau and Whitcher Kansa 2018); 5) the reproducibility of the identification of some features (Blumenschine et al. 1996; Lloveras et al. 2014); 6) the reproducibility of applying metric measurements (Lyman and VanPool 2009); and 7) the impact of fragmentation (Pickering et al. 2006; Domínguez-Rodrigo 2012; Morin et al.

2017). Together these studies demonstrate inter- and intra-analyst variation in a wide range of zooarchaeological tasks, resulting from differences in the analysts' research backgrounds and innate abilities. The observed inter-analyst variation therefore leads to zooarchaeological identifications being subjective, even though this subjectivity is limited and controlled by a variety of reference materials including guides and manuals, both physical and virtual 3D reference specimens, and images and sketches of bones. Inter-analyst variation is commonly acknowledged and accepted in the zooarchaeological community, with for instance Twiss et al. (2017, p. 303) considering zooarchaeological identification to be an ´interpretative act', and Wolverton (2013, p.390) maintaining that it is 'important to acknowledge that morphological identification is a subjective process'. In referring to subjectivity, we mean that the results of zooarchaeological research are dependent on and limited by the analysts' research histories and abilities, and it is the inter-analyst variance of classification accuracy that is of interest. Regarding intra-analyst variance, we refer to consistency.

When the subjective nature of the zooarchaeological identification process is combined with bones and species that are difficult to separate, one should expect a high level of inconsistency and inaccuracy. It is extremely well-known within the zooarchaeological community that sheep and goats are difficult to differentiate from their skeletal remains, to the extent that Noddle (1974, p. 195) called the problem 'legendary'. However, because sheep and goats are globally important domestic species, they played a key role in the early domestication process, and they are highly important to sedentary communities (Zeder 2008; Culley et al. 2021), it is important that sheep and goat bones are reliably identified by zooarchaeologists and that zooarchaeologists can trust each other's identifications of these elements. The importance of separating sheep and goat is also reflected in the fact that one of the first major uses of ZooMS (Zooarchaeology by Mass Spectrometry) was performed to separate sheep and goats (Buckley et al. 2010).

The astragalus was chosen as the focus of this study because they tend to survive relatively intact in archaeological contexts due to their relatively high density and low nutritional value (Haruda 2017; Pöllath et al. 2018), which means that they are rarely broken intentionally, although astragali are occasionally re-purposed as gaming pieces (Gilmour 1997; Koerper and Whitney-Desautels 1999; Holmgren 2004). Furthermore, several publications have used sheep and goat astragali as the subject of their research

in showing that their methods can separate the two species (Davis 2016, 2017; Haruda 2017; Salvagno and Albarella 2017; Salvagno 2020), whereas others have argued that descriptive morphological criteria can be defined to separate sheep and goat astragali (Boessneck et al. 1964; Boessneck 1969; Prummel and Frisch 1986; Fernandez 2001; Zeder and Lapham 2010). Sheep and goat astragali are therefore ideal for testing the accuracy and consistency of zooarchaeological analysts as the community is familiar with the problem and there are many studies that provide context for the present study. Likewise, existing osteometric and comparative methods provide a point of comparison for analyst performance in this blind study, as does the prior blind study of Zeder and Lapham (2010). Zeder and Lapham's (2010) blind study is considered at length in the next section for context.

Analyst performance in identifying sheep and goat astragali

Our study builds upon previous work by Zeder and Lapham (2010), who produced an in-depth analysis of their blind study, which tested a smaller sample of analysts on a broader set of sheep and goat bones than ours. To their credit, Zeder and Lapham (2010) not only recognised the importance of accuracy and consistency requirements in zooarchaeological macro-morphological identification, but were also the first to perform an extensive blind study, and the following re-analysis was only possible due to them making the raw data readily available. Because of the influence and importance of Zeder and Lapham's (2010) work in zooarchaeology, especially their argument that macro-morphological comparisons produce excellent classification accuracies, we use their study as a relevant backdrop to highlight prevalent issues in the zooarchaeological process as whole, namely the influence of the hierarchically incompatible categories (sheep/goat versus binary classifications) on the classification accuracy. Likewise, we use Zeder and Lapham's (2010) work as an example of how macro-morphological comparisons are dependent on analyst decisions and therefore descriptive methods are not always followed consistently. We do not claim that these issues were consciously ignored, only that they happen, nor do we aim to discredit their work. Moreover, we solely focus on the astragalus classifications.

In their assessment of the postcranial elements, Zeder and Lapham (2010, Table 5) found that using descriptive criteria to classify sheep and goat astragali results in a correct classification rate of 100% for goat and 97% for sheep astragali. This finding was then

adopted by Pöllath et al. (2019, p.812) who stated that they 'consider it unlikely that our samples contained misidentified goat astragali' to bolster the readers' trust in the morphological assessment that their study was reliant upon. This is a risky stance to take considering that Zeder and Lapham (2010) reported 2.2% and 15.4% rate of sheep/goat identifications for goat and sheep astragali, respectively, which indicates a large amount of uncertainty especially around sheep astragalus identifications. Zooarchaeologists have grown accustomed to accepting sheep/goat almost as if it is yet another 'species' and such a classification is often considered a 'correct' classification (e.g. Davis 2017, p. 67), or ignored from the total number of classifications when computing accuracies as done by Zeder and Lapham (2010). As bones classified as sheep/goat present a considerable amount of uncertainty to the practitioner (Wolfhagen and Price 2017), their exclusion from the total number of classifications means that the accuracy rates for individual species may be inflated. We provide an example of this issue in **Error! Reference source not found.**. Although the sheep/goat category may also be thought of as a 'reject' class, it should not be considered a pure reject class either, since such a role is already reserved for the 'unidentified' or 'indeterminate' category. Instead, sheep/goat classification reflects a higher hierarchical and taxonomical level of classification – and is therefore closer to a 'catch-all' than a 'reject' class – which is incompatible with binary classifications such as sheep and goat identification. Moreover, even if combining the two species increases sample size and produces a general picture of caprine exploitation, doing so also obfuscates the potential differences in the management strategies for the two species (Halstead et al. 2002; Buckley et al. 2010).

In addition to reporting the accuracy for each element, Zeder and Lapham (2010) report the classification rates for each individual criteria for each element as well. For the astragalus, these criteria involve: 1) the angle of the medial articular ridge in the dorsal aspect; 2) the shape of the distal articular surface in lateral aspect; 3) the size and shape of the proximo-plantar projection in medial aspect; and 4) the prominence of the medial articular ridge in plantar aspect. In Zeder and Lapham's (2010) Table 3, in which each of the four criteria used in differentiating sheep and goat astragali were assessed individually, the error rates are reported as follows: 1) for the first criterion, the error rate is 0% for goats and 24.7% for sheep with 0% sheep/goat identifications for goat and 6.4% for sheep; 2) for the second criterion, the error rate is 4.6% for goats and 12.3%

for sheep in addition to 4.4% sheep/goat for goats and 6.4% for sheep; 3) for the third criterion, the error rate for goats is 4.6% and 2.8% for sheep, which is again accompanied with 4.4% sheep/goat for goats and 6.5% for sheep; and 4) for the final criterion, the goat astragalus error rate is 7.1% and 7.4% for sheep, and there are 8.7% sheep/goat identifications for goat and 12.8% for sheep. However, these reported error rates were computed for intact specimens and the error rates for archaeological specimens would be different as correct identifications are dependent on the survival of these four features of astragali. Moreover, although these error rates and percentages of sheep/goat identifications are seemingly low, they are not to be taken at face value as they derive from the authors' assessments of the bones even though they were aware of the ground-truth species, calling into question the validity of such a study, a point which Zeder and Lapham (2010) readily admit and attempt to correct by also reporting on results of a proper blind study. The same warning applies to the whole bone classification rates reported in Zeder and Lapham's (2010) Table 5.

Considering the aforementioned issue regarding Zeder and Lapham's (2010) assessment methodology, the fact that all relevant features may not be present in archaeological samples, and that the morphological variances of archaeological populations are likely to be different from modern samples, it is therefore not justifiable to use the identification rates from Zeder and Lapham's (2010) Tables 3 as 5 as expected accuracies in archaeological studies. Furthermore, taking the reported error rates at face value makes the assumption that all analysts would perform equally well when following Zeder and Lapham's (2010) criteria.

In fact, the blind study part of Zeder and Lapham's (2010) research demonstrates the exact opposite. In their blind study, six analysts, including the two authors, attempted the identification of ten specimens from each species for several elements, demonstrating inter-analyst variance as well as how the analysts' error rates differ when classifying sheep versus goat bones. According to Zeder and Lapham (2010), the astragali were analysed by all six participants, but from the tables included in the article and the associated Supplementary Table 4 – in which the blind study results were detailed – it appears that only five of the participants analysed the astragali. Zeder and Lapham's (2010) Supplementary Table 4 is the basis for the following two paragraphs.

Astragalus identifications from analyst decision

All analysts

	TP	TN	FN	FP	Accuracy	Precision	Recall	F_1-score
Sheep	41	49	5	0	94.74%	100.00%	89.13%	94.25%
Goat	49	41	0	5	94.74%	90.74%	100.00%	95.15%

Analysts 1 and 2

	TP	TN	FN	FP	Accuracy	Precision	Recall	F_1-score
Sheep	17	19	2	0	94.74%	100.00%	89.47%	94.44%
Goat	19	17	0	2	94.74%	90.48%	100.00%	95.00%

Analysts 4, 5 and 6

	TP	TN	FN	FP	Accuracy	Precision	Recall	F_1-score
Sheep	24	30	3	0	94.74%	100.00%	88.89%	94.12%
Goat	30	24	0	3	94.74%	90.91%	100.00%	95.24%

Table 1 *Analyst performance for all astragalus identifications in Zeder and Lapham's (2010) blind study based on the analysts' final decisions. The true positive, true negative, false negative, and false positive derive from the analysts' final decision which may differ from majority rule of the different criteria. Analysts 1 and 2 are the authors of the study*

In the blind test, the analysts had two tasks: 1) classify individual features as sheep, goat, or sheep/goat; and 2) classify the bones as sheep, goat, or sheep/goat. This then means that there are two ways in which the analyst accuracy can be computed: 1) majority-voting, in which each criterion has one vote and sheep/goat classifications are ignored; or 2) taking only the final decision into account, again ignoring sheep/goat classifications. Treating the ambiguous sheep/goat identifications as a reject category and excluding such classifications from the computations, the mean accuracy for all analysts was 94.74% (**Table 1**) for the astragalus identifications when only the final decisions are taken into account, and 89.36% (**Table 2**) when the species identifications were computed using majority-voting by criteria (Zeder and Lapham 2010). However, these accuracies are inflated by the exclusion of sheep/goat classifications from the total number of classifications. The authors (analysts 1 and 2) of the article did not fare any better than the less experienced analysts (analysts 4, 5, and 6 in **Table 1**) when just final decisions were taken into account and performed worse than the inexperienced analysts when considering each criterion as a vote (**Table 2**).

Astragalus identifications from criteria

All analysts

	TP	TN	FN	FP	Accuracy	Precision	Recall	F_1-score
Sheep	37	47	9	1	89.36%	97.37%	80.43%	88.10%
Goat	47	37	1	9	89.36%	83.93%	97.92%	90.38%

Analysts 1 and 2

	TP	TN	FN	FP	Accuracy	Precision	Recall	F_1-score
Sheep	15	17	4	1	86.49%	93.75%	78.95%	85.71%
Goat	17	15	1	4	86.49%	80.95%	94.44%	87.18%

Analysts 4, 5 and 6

	TP	TN	FN	FP	Accuracy	Precision	Recall	F_1-score
Sheep	22	30	5	0	91.23%	100.00%	81.48%	89.80%
Goat	30	22	0	5	91.23%	85.71%	100.00%	92.31%

Table 2 *Analyst performance for all astragalus identifications in Zeder and Lapham's (2010) blind study based on the majority-voting by criteria. The true positive, true negative, false negative, and false positive derive from the analysts' decision on individual criteria and were computed based on majority rule. Analysts 1 and 2 are the authors of the study*

Zeder and Lapham's (2010) Supplementary Table 4 further shows that for two astragali, analyst 1 (one of the authors) scored all four criteria individually as either goat or sheep/goat, but then decided that the bone was actually sheep. In contrast, analyst 2 (again, one of the authors) assigned one astragalus as a goat even though they thought that the bone adhered to sheep-like qualities on three of the four criteria. Furthermore, there were four other instances (all analysts) in which a bone was considered sheep/goat even though the majority of the criteria pointed to a more precise species assignation. In three of the four of such cases, the decision to provide a more precise classification would have resulted in a wrong answer.

This behaviour by the analysts to disregard the morphological criteria highlights the problem with the methodology, namely that the individual criteria do not fully capture the species variation which leads to inconsistent application of the comparative methodology. It is also a distinct possibility that size and other undescribed morphological variables affect the analysts' decision making. However, although the effect of undescribed morphological variables is unmeasurable and cannot be easily removed as a factor, bone size as a factor can be removed and doing so could help in producing a more accurate assessment of the reliability of descriptive criteria. This can be most easily achieved by using high definition images of bones, finding the outline of

the bones and cropping the image to the extents of the bone, after which all images can be scaled to a pre-determined size by padding the cropped image with varying number of pixels. To ascertain that analysts are consistent and follow the written criteria, it would be similarly helpful to collect data on the analysts' areas of attention during the classification task. Again, photographs are helpful in achieving this goal as participants would only have to trace their attention areas over the regions of the photo. Moreover, the image dataset used in the present study is suitable for future deep learning applications, meaning that the blind study acts as a benchmark of human ability for any deep learning classifiers using these images.

Aims and objectives

This study has several aims: 1) to test the impact of different types of reference materials in a zooarchaeological task; 2) analyse the participants' spatial attention during classification; and 3) as it is argued that sheep/goat classifications are not only unnecessary, but that they also hide important information about the classifications, this category is replaced by a two-step process in which the analyst is forced to make a classification and additionally report their confidence score for each classification. These self-reported confidence scores force the analysts to consider the uncertainty in their classifications and it is shown that higher confidence scores also correspond to higher accuracy overall. This study additionally provides a human benchmark for an archaeologically relevant image dataset for forthcoming deep learning algorithms. The data was gathered in an online blind study and it is analysed in a series of quantitative and qualitative analyses.

Methods

Image dataset

The specimens included in the dataset come from National Museum Cardiff's Noddle collection, Sheffield University's zooarchaeological reference collection, and Historic England's collections stored at Fort Cumberland in Portsmouth. The list of specimens is provided in **Error! Reference source not found.**. In total, 193 astragali (100 sheep, 93 goat) were photographed from six different views. However, nine specimens were not included in the blind study as there were problems with the quality of photographs of two specimens and another seven had species information written on the bones

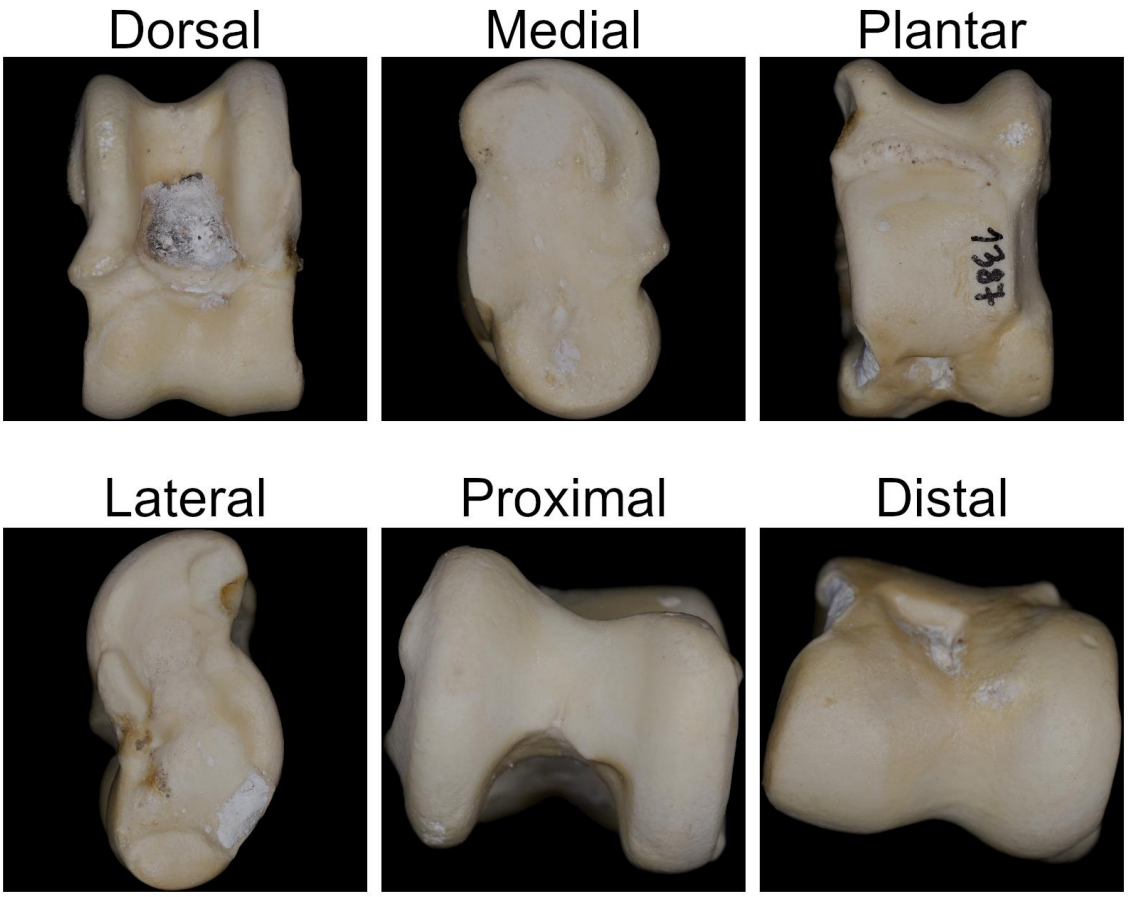

Fig. 1 *Example images for a goat astragalus*

themselves. The photography and post-processing steps are detailed in **Error! Reference source not found.** and an example of the six images (as shown to study participants) produced for each astragalus from this process is displayed in **Fig. 1**.

Data collection

The blind study was hosted at www.sheepgoat.co.uk and took place between 23 June 2020 and 31 December 2020. In total, 39 fully anonymous participants completed the entire study. Each analyst was first asked to consent to the test as required to fulfil the ethics requirements (ethics approval granted on 15 June 2020 by UCL Institute of Archaeology Ethics Committee, Reference number 2020.020), followed by a series of questions about their experience in zooarchaeology and the type of reference materials they would be using (detailed in **Error! Reference source not found.**). After the pre-test survey, the participant was taken to a page detailing the instructions for the two main tasks of the study (see **Error! Reference source not found.**). The first of these two tasks involved showing the analyst an astragalus from all six views and asking them to identify the specimen as either sheep or goat as well as give an estimate of the analyst's

confidence in that classification (a sliding scale from 'Guess' to 'Absolutely certain', or numerically from 1 to 100 with a default value of 51). The participants were given the option to indicate whether they recognized the bone from reference collections because the participant may be from the same institution as the specimen – this feature was used only once and that classification is not taken into account in any analyses. The second task involved painting the areas that the analyst thought were informative for the species assignation on top of the set of images shown on the previous page. For Analyst 29, there was a glitch in their entries and they provided 29 classifications and 31 drawings – only those drawings with corresponding classifications are included in the analysis of analyst attention. After the test, the participants were allowed to leave feedback and they were given the option to view how well they performed. The participants had the option to delete their entries throughout the test and even after it. The database entity relationship diagram is in **Error! Reference source not found.**. The analyst answers and test images are in Online Resource 11.

Because asking participants to classify all 184 bones would have been too time-consuming, a random sampling strategy was chosen. This was implemented as part of the functioning of the website, which picked 20 astragali (ten of each species) for each analyst. As it was of interest to also gauge analyst consistency, another ten astragali (five of each species) were chosen at random from those initial 20. The participants were not told that their consistency was going to be measured, but they were told that the test would involve 30 specimens. Had the analysts been told about consistency testing, they may have tried to memorise their initial identification, which could have led to artificially improved consistency scores. It was programmatically ensured that none of the repeated bones were shown immediately after its first occurrence. Because of the randomised selection of samples for all participants, one of the 184 astragali was not shown to anyone.

Measuring expertise and defining expertise groups

Qualifications, years in profession and track record are all poor predictors of test performance (Burgman et al. 2011). Previous studies have found this to be the case in other species identification tasks such as fisheries observers' ability to identify sharks (Tillett et al. 2012) and great crested newt licence holders' ability to sort images of newts to species (Austen et al. 2018). Considering the fact that these proxies for experience

are not reliable predictors of expertise, it was decided on an alternative approach to group analysts by expertise. In this study, the analysts were aggregated into expertise groups based on their answers to five questions, the exact wordings of which can be found in Online Resource 4. In general, these questions were designed to act together as a proxy for the participants' expertise in sheep and goat identification, not faunal analysis overall.

The participants were asked about: 1) their highest level of zooarchaeological qualification; 2) how many zooarchaeological assemblages they have worked on in the last five years; 3) how many hours per week on average they spent analysing zooarchaeological remains in the last five years; 4) whether they are specialised in the identification of land mammals; and 5) if their specialist experience involved sheep and goat separation. For the last two questions, the participants were given the option to select 'Not applicable'. In hindsight, this was a mistake, so these answers are interpreted in the analyses as 'No'. The first three questions were chosen because they were assumed to reflect overall experience and/or continued practice, both of which were assumed to correlate with a larger mental reference population, while continued practice is strongly correlated with skill (Ericsson and Lehmann 1996). These questions focused on the last five years (an arbitrary choice) because domain knowledge diminishes if it is not practiced (Endsley and Kiris 1995), and it could not be assumed that all participants were practicing zooarchaeologists. Furthermore, it is likely that most participants have at least some form of underlying training in mammalian zooarchaeology, although this was not asked specifically. The grouping of analysts was done with the combination of K-medoids and Principal Components Analysis (PCA), as discussed next.

PCA and K-medoids

The answers to the above questions were pre-processed by centring them around zero and scaled to unit variance due to varying number of options for each question, as recommended (Abdi and Williams 2010). This was done using Scikit-learn's StandardScaler method (Pedregosa et al. 2011) in Python 3.6. (Python Software Foundation 2016). Once the principal components were computed, those principal components that explain over 80% of the variance were used as input to K-medoids clustering, which aggregated analysts into groups of similar skill-level. K-medoids is more

robust to noise and outliers than K-means clustering as the cluster centres in K-medoids are the most centrally located objects, whereas in K-means the cluster centre can be between objects (Zhang and Couloigner 2005). The number of analyst groups created was based on the combined evaluation of sum of squared distances (the elbow method), Calinski-Harabasz index, and silhouette score.

Analysing the impact of reference materials

In addition to asking about the participants' use of reference texts (one of Boessneck 1969 or Boessneck et al. 1964, Zeder and Lapham (2010), Prummel and Frisch (1986), Other, and None), the analysts were asked to provide information on whether they were planning on using physical specimens, photographs, sketches, or 3D models of either or both species as reference aides during the test. This information was important in measuring the impact of different reference material types on the analysts' performances. Although it is acknowledged that it would be better to have a control group and separate sessions for all participants so that the impact of different reference materials was more directly measurable, this was not possible within the research timeframe and it was thought improbable that large enough cohort of participants could have been found for a such a multi-stage study.

Generalized linear mixed effects model

Instead of approaching the impact of reference materials in an unfeasible longitudinal study, generalized linear mixed effects models (GLMM) are used to model the relationship between the different types of reference materials and the analysts' classifications. GLMM is a statistical modelling method that merges the properties of linear mixed models (LMM) and generalized linear models (GLM; Bolker et al., 2009). Like GLM, GLMM involves the use of link functions to model non-normal data, whereas GLMM resembles LMM in its use of random effects (Bolker et al. 2009; Stroup 2013). Thus, GLMM is preferrable over GLM and LMM when the response variable is non-normal and one can expect variance beyond that explained by the measured variables (e.g. type of reference material). Additionally, GLMM can be effectively applied to unbalanced and zero-inflated experimental designs, which is the case here (Bolker et al. 2009; Moscatelli et al. 2012). GLMM is suitable for the present problem because the response variable is a binary variable encoding the correctness of a classification (1 =

'Correct', 0 = 'Incorrect') and random variation can be modelled for the subjects (analysts) and the test items (bones). The present study additionally uses the species of the animal as a source of variance for each analyst.

The variables modelling random variance (i.e. heterogeneity between clusters) are collectively called random effects and the fixed effects are those factors whose levels are determined by the experiment and are the main interest of the study (Bolker et al. 2009; Moscatelli et al. 2012). However, note that both random and fixed effects act as explanatory variables and their parameters are estimated through maximum likelihood estimation (Bolker et al. 2009). Models that implement random effects aim to incorporate variance between clusters – in the present study the clusters are crossed so that each specimen may have been seen by multiple analysts, but each analyst's ability also varies by the species of specimen. In practice, this model definition is achieved by allowing each analyst and specimen to have their own intercepts, and the slope for each analyst is defined by species. In other words, the random effects in the models discussed in the results section account for variance between analysts (be it due to skill or otherwise), between specimens (due to differences in morphology, quality of the photograph, or any other reason), and between species within the analyst (because the analyst may be biased towards one species). These variances have to be estimated as measuring them directly is very difficult. Fixed effects on the other hand are the factors of interest, namely the different reference material types and analyst expertise groups.

The following example model lends heavily from Barr et al. (2013) who provides a step-by-step explanation of defining a linear mixed model and which is here adapted for GLMM. In this example, the binary response variable Y_{si} is modelled by a fixed effect *X* (e.g. species) and two random effects (subject, *S*, and item, *I*) with a slope of *X* varying within subject:

$$g^{-1}(E(Y_{si}|S_{0s}, S_{1s}, I_{0i})) = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s})X_i + e_{si}.$$

As the response variable is a binary variable in the present study, a logit link function is used to relate the observations and the predictors. In the above equation, $g^{-1}(\cdot)$ is the inverse of $logit(\cdot)$ link function, $E(Y_{si}|S_{0s}, S_{1s}, I_{0i})$ is the expected value of the *sth* subject for the *ith* item when given by-subject random intercept S_{0s}, by-subject random slope S_{1s}, and by-item random intercept I_{0i}. β_0 is the overall intercept, β_1 is the overall slope, and X_i is the fixed effect predictor dummy variable for the *ith* item (e.g. a level of

species such that 0 = 'Goat'; 1 = 'Sheep'). The term e_{si} models the residual error for the *ith* item and *sth* subject.

$\beta_0, \beta_1, S_{0s}, S_{1s}$ and I_{0i} are the parameters estimated through maximum likelihood estimation. The distribution of variance parameters (S_{0s}, S_{1s}, I_{0i}) is usually assumed to be Gaussian with a mean of zero, although non-Gaussian random effects have been suggested (Lee and Nelder 1996, 2006). The variance distributions for the example model are

$$(S_{0s}, S_{1s}) \sim \mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_{00}^2 & \rho\tau_{00}\tau_{11}\\ \rho\tau_{00}\tau_{11} & \tau_{11}^2\end{bmatrix}\right),$$

$$I_{0i} \sim N(0, \omega_{00}^2),$$

$$e_{si} \sim N(0, \sigma^2).$$

Here, τ_{00}^2 is the random intercept variance for subjects, τ_{11}^2 is the random slope variance for subjects, $\rho\tau_{00}^2\tau_{11}^2$ is the intercept-slope covariance for subjects, and ω_{00}^2 is the random intercept variance for items. These parameters then tell us if the inclusion of S_{0s}, S_{1s} and I_{0i} random effects in the model is of importance, because if they are associated with zero or very low variance it would indicate that there is little to no effect in the probability of correct answer between analysts and/or specimens. For a deeper explanation of GLMM, see Agresti (2002) or Stroup (2013), and for tutorials on implementing GLMM in R, see Baayen et al. (2008) and Bolker et al. (2009).

Finding the best fit GLMM

The process of finding best fit GLMM involves confirming that the inclusion of random effects is sensible. This can be done by fitting a baseline GLM and comparing a series of GLMMs that each have varying combinations of random effect structures to the baseline GLM. At this stage, the GLMMs and the GLM do not have fixed effects. The comparison of the baseline model and GLMM with random effects is tested through Likelihood Ratio Tests (LRT), and the best random effects structures are chosen based on the lowest Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and negative log likelihood. The best random effects structure is not the final model and it is extended by adding fixed effects in varying combinations. The models fitted with fixed effects and the model with the best random effects structure are further compared using LRT to show that the fixed effects add useful information to the best random effects model.

With many fixed effects, multicollinearity may become an issue, so models presenting such characteristics are ignored.

Software and packages

All models are created using the *lme4* (Bates et al. 2015) package in R version 4.1.1. (R Core Team 2021). Laplace approximation is used to estimate the model parameters with the help of the BOBYQA (Bound Optimization BY Quadratic Approximation) optimizer. Regarding multicollinearity, models with Variance Inflation Factor of above 5 are considered above the critical threshold, although higher and lower thresholds are also common (see Zuur et al., 2010; Thompson et al., 2017). Over- and underdispersion are unidentifiable for binary response values, so they are not evaluated for the models (Kain et al. 2015).

Auxiliary measurements

All classifications were timed by starting a timer as the page loaded and stopped as the participants progressed to the drawing task. As the classification page contained multiple tasks, the measured response speed is not a direct measure of response speed and this metric therefore has a potentially low signal-to-noise ratio. For this reason analysis of response speed is not included as part of the main text, but it is available in **Error! Reference source not found.**.

In addition, the width of the browser window was measured because the study was designed to be shown in a browser window with a minimum width of 1,586 pixels. Analysis regarding the impact of browser window size shows that those analysts with browser window widths less than 1,586 pixels did not perform markedly worse than those with wider browser windows and all analysts are therefore included in all analyses (**Error! Reference source not found.**).

Results

All statistics are performed in Python 3.6 using SciPy statistics package (Virtanen et al. 2020), apart from the GLMM analyses of reference materials for which the *lme4* package in R has been established as one of the standard packages.

Analyst expertise groups

A summary table of analyst expertise is presented in **Table 3**. Majority (34) of the analysts have a master's or a doctorate degree with a zooarchaeological component, but only three of the participants perform zooarchaeological identifications full-time. Majority of the participants (32) consider themselves as land mammal specialists and 25 participants also say they have experience of separating sheep and goat bones. Using the information in **Table 4** as input to PCA, the first three principal components of the PCA explains approximately 82.95% of variance (**Fig. 2**A). The first three components are then used as the inputs to K-medoids. Evaluating the K-medoids output through Calinski-Harabasz index and sum of squared distances (elbow method) results in the division of the analysts into four groups (**Fig. 2**B and **Fig. 2**C). Although the Silhouette coefficient results in two clusters (**Fig. 2**D), the difference between the coefficients of two and four clusters is small. On this basis, the analysts are divided into four groups, with each analyst's group membership indicated in **Table 4**.

The first (Group 1) of these four groups is interpreted as being formed of professionals (e.g. commercial zooarchaeologists) because it mainly includes full-time or near full-time workers who also have analysed many different assemblages in the past five years. Group 2 are relative novices at classifying land mammals and especially sheep and goat bones, while they also have not generally worked on many assemblages nor do they spend a lot of time acting as zooarchaeologists. Group 3 members are classed as postgraduates due to all of them holding master's qualifications. Group 3 members also all expressed having experience of separating sheep and goats, but they do not tend to spend as much time identifying bones as members in Group 1. Finally, participants in Group 4 are classed as doctorates because they all have PhD qualifications, and these analysts also have experience of separating sheep and goat bones. Group 4 analysts are similar to Group 3 analysts in that they do not spend a lot of time identifying bones. The expected order of performance for these groups from worst to best is: Group 2, Group 3, Group 4, Group 1. The Group 1 analysts have been placed ahead of Group 4 analysts in our expectations on the basis that the professionals are more active in using their skills, which we assume to be more important for performance than the highest level of zooarchaeological qualification. We also expect the doctorates (Group 4) to outperform the postgraduates (Group 3) based on their higher level of education, while novices

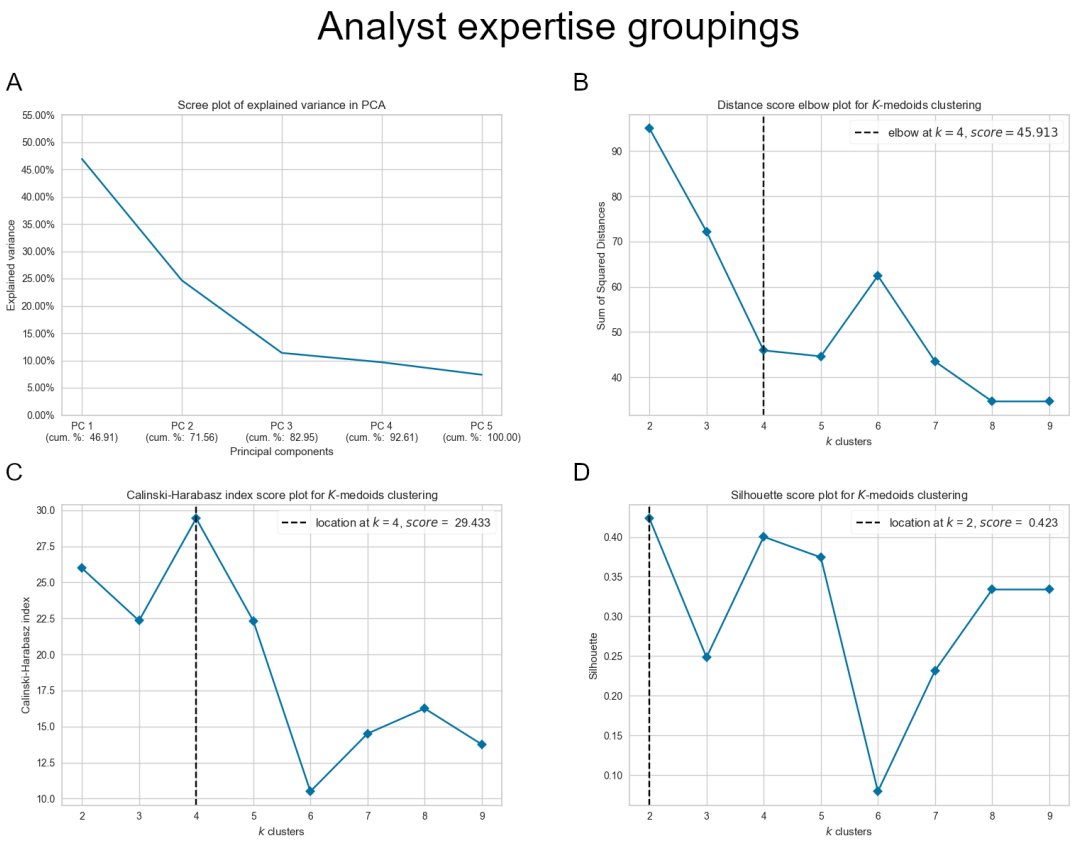

Fig. 2 *A) PCA scree plot showing cumulative explained variance. B) Sum of squared distance score for K-medoids. C) Calinski-Harabasz index score for K-medoids. D) Silhouette score for K-medoids. The vertical dashed line in figures B, C, and D indicates the suggested number of cluster. Note the closeness in scores between two and four clusters in figure D*

(Group 2) are expected to be the worst performing group based on their lack of familiarity with sheep and goat differentiation and land mammal specialisation.

Analyst performance

The four most accurate analysts (three of which are in Group 4) managed to correctly classify 29 of the 30 sets of astragali, giving them an accuracy of 96.67% (Table 5). Their accuracy is far better than the average accuracy for all analysts, which is 81.15%. The least accurate participant is Analyst 68 in Group 1, whose accuracy (53.33%) is indistinguishable from chance. The mean accuracy for analysts in Group 4 (87.27%) is ten percentage points higher than for analysts in Groups 1 (77.08%) and 2 (77.13%), with Group 3 analysts (82.82%) also faring better than Group 1 and Group 2 analysts (Table 6). However, median accuracy for analysts in Group 1 (80.00%) is better than for analysts in Group 2 (73.33%) and much closer to Group 3 median (81.67%). The median performance in Group 4 is 93.33%, which shows that the distribution of analyst accuracies is skewed left, just like it is for Group 1, whereas Group 2 is skewed right and Group 3 is only negligibly skewed left. It is therefore inferred that Groups 1, 2 and 4 may

Count of analysts by levels of expertise

	Highest degree			
	None	BA/BSc	MA/MSc	PhD
No of analysts	1	4	17	17

	# Assemblages					
	1-10	10-20	20-30	30-40	40-50	>50
No of analysts	23	10	3	1	1	1

	Hours per week			
	<10	10-20	20-30	Full time
No of analysts	23	9	4	3

	Land mammal specialist		Sheep/goat experience	
	Yes	No	Yes	No
No of analysts	32	7	25	14

Table 3 *Analysts' expertise summarised*

contain outlying analysts. The boxplots in **Fig. 3** confirm this suspicion and it is further noted that Group 3 has one outlying analyst who performed far better than most for that group of analysts.

The analysts generally performed better when identifying sheep (84.74% accuracy) than goat (77.56%) astragali (Table 7). The median for both species is slightly higher than the mean, with a median of 86.67% for sheep and 80.00% for goat. Group 4 analysts (mean: 93.33%, median: 100%) are the best at identifying sheep astragali, but Group 3 had the highest mean accuracy for goat astragali (mean: 83.09%, median: 83.34%) whilst also being the only group of analysts that has similar accuracies for both species. However, Group 4 analysts had the highest median accuracy for goat astragali (mean: 81.21, median: 86.67%). As the sheep and goat accuracies for all analysts were found to violate normal distribution in Shapiro-Wilk test (Table 8), Mann-Whitney U test was performed and it was found that analysts are overall more accurate in classifying sheep than goat astragali ($U = 965.0$, $p = 0.0393$, $N = 39$). Cohen's d effect size (0.4892) indicates a moderate effect as Cohen's d values of 0.15, 0.36, and 0.65 represent the thresholds (based on empirical evidence) for small, medium, and large effect sizes, respectively (Lovakov and Agadullina 2021).

Analyst expertise groups

	Analyst	Group	Qual.	# Assemblages	Hours per week	Land mammal specialist	Sheep/goat experience
Professionals	3	1	MA/MSc	20-30	10-20	Yes	Yes
	10	1	PhD	20-30	10-20	Yes	Yes
	51	1	PhD	40-50	20-30	Yes	Yes
	61	1	PhD	10-20	20-30	Yes	Yes
	67	1	MA/MSc	>50	20-30	Yes	Yes
	68	1	MA/MSc	30-40	Full time	Yes	No
	71	1	PhD	10-20	Full time	Yes	Yes
	84	1	MA/MSc	10-20	Full time	Yes	Yes
Novices	1	2	BA/BSc	1-10	<10	No	No
	11	2	MA/MSc	1-10	<10	Yes	No
	26	2	BA/BSc	1-10	<10	No	No
	29	2	BA/BSc	1-10	10-20	Yes	No
	44	2	MA/MSc	1-10	<10	Yes	No
	47	2	PhD	1-10	<10	No	No
	56	2	PhD	1-10	<10	No	No
	58	2	BA/BSc	1-10	<10	No	No
	60	2	None	1-10	<10	No	No
	65	2	MA/MSc	1-10	<10	Yes	No
	66	2	MA/MSc	1-10	<10	Yes	No
	96	2	MA/MSc	1-10	20-30	No	No
Postgraduates	2	3	MA/MSc	1-10	<10	Yes	Yes
	4	3	MA/MSc	1-10	10-20	Yes	Yes
	69	3	MA/MSc	1-10	10-20	Yes	Yes
	82	3	MA/MSc	10-20	<10	Yes	Yes
	95	3	MA/MSc	1-10	10-20	Yes	Yes
	99	3	MA/MSc	1-10	10-20	Yes	Yes
	100	3	MA/MSc	1-10	<10	Yes	Yes
	104	3	MA/MSc	20-30	<10	Yes	Yes
Doctorates	18	4	PhD	10-20	10-20	Yes	Yes
	36	4	PhD	10-20	<10	Yes	No
	37	4	PhD	1-10	<10	Yes	Yes
	43	4	PhD	10-20	<10	Yes	Yes
	46	4	PhD	1-10	<10	Yes	Yes
	48	4	PhD	10-20	<10	Yes	Yes
	53	4	PhD	10-20	<10	Yes	Yes
	62	4	PhD	1-10	<10	Yes	Yes
	81	4	PhD	1-10	10-20	Yes	Yes
	83	4	PhD	1-10	<10	Yes	Yes
	103	4	PhD	10-20	<10	Yes	Yes

Table 4 *Itemised answers and analyst group membership*

Analyst performances

		Both species (mean)				Sheep				Goat			
	Analyst	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.	F_1
Professionals	3	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%
	10	83.33%	87.50%	83.33%	82.86%	100.00%	75.00%	100.00%	85.71%	66.67%	100.00%	66.67%	80.00%
	51	76.67%	77.78%	76.67%	76.43%	66.67%	83.33%	66.67%	74.07%	86.67%	72.22%	86.67%	78.79%
	61	80.00%	85.71%	80.00%	79.17%	100.00%	71.43%	100.00%	83.33%	60.00%	100.00%	60.00%	75.00%
	67	90.00%	90.18%	90.00%	89.99%	93.33%	87.50%	93.33%	90.32%	86.67%	92.86%	86.67%	89.66%
	68	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%	53.33%
	71	80.00%	80.54%	80.00%	79.91%	86.67%	76.47%	86.67%	81.25%	73.33%	84.62%	73.33%	78.57%
	84	86.67%	87.33%	86.67%	86.61%	80.00%	92.31%	80.00%	85.71%	93.33%	82.35%	93.33%	87.50%
Novices	1	86.67%	87.33%	86.67%	86.61%	93.33%	82.35%	93.33%	87.50%	80.00%	92.31%	80.00%	85.71%
	11	70.00%	75.57%	70.00%	68.27%	93.33%	63.64%	93.33%	75.68%	46.67%	87.50%	46.67%	60.87%
	26	70.00%	70.09%	70.00%	69.97%	66.67%	71.43%	66.67%	68.97%	73.33%	68.75%	73.33%	70.97%
	29	65.52%	65.87%	65.71%	65.48%	71.43%	62.50%	71.43%	66.67%	60.00%	69.23%	60.00%	64.29%
	44	86.67%	87.33%	86.67%	86.61%	80.00%	92.31%	80.00%	85.71%	93.33%	82.35%	93.33%	87.50%
	47	73.33%	73.76%	73.33%	73.21%	66.67%	76.92%	66.67%	71.43%	80.00%	70.59%	80.00%	75.00%
	56	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%
	58	70.00%	72.50%	70.00%	69.14%	86.67%	65.00%	86.67%	74.29%	53.33%	80.00%	53.33%	64.00%
	60	73.33%	73.76%	73.33%	73.21%	80.00%	70.59%	80.00%	75.00%	66.67%	76.92%	66.67%	71.43%
	65	76.67%	77.78%	76.67%	76.43%	86.67%	72.22%	86.67%	78.79%	66.67%	83.33%	66.67%	74.07%
	66	90.00%	91.67%	90.00%	89.90%	80.00%	100.00%	80.00%	88.89%	100.00%	83.33%	100.00%	90.91%
	96	70.00%	70.09%	70.00%	69.97%	73.33%	68.75%	73.33%	70.97%	66.67%	71.43%	66.67%	68.97%
Postgraduates	2	96.67%	96.88%	96.67%	96.66%	100.00%	93.75%	100.00%	96.77%	93.33%	100.00%	93.33%	96.55%
	4	86.67%	87.33%	86.67%	86.61%	80.00%	92.31%	80.00%	85.71%	93.33%	82.35%	93.33%	87.50%
	69	83.33%	83.48%	83.33%	83.31%	86.67%	81.25%	86.67%	83.87%	80.00%	85.71%	80.00%	82.76%
	82	80.00%	80.54%	80.00%	79.91%	73.33%	84.62%	73.33%	78.57%	86.67%	76.47%	86.67%	81.25%
	95	75.86%	75.96%	75.71%	75.75%	80.00%	75.00%	80.00%	77.42%	71.43%	76.92%	71.43%	74.07%
	99	76.67%	76.79%	76.67%	76.64%	73.33%	78.57%	73.33%	75.86%	80.00%	75.00%	80.00%	77.42%
	100	83.33%	83.48%	83.33%	83.31%	80.00%	85.71%	80.00%	82.76%	86.67%	81.25%	86.67%	83.87%
	104	80.00%	80.54%	80.00%	79.91%	86.67%	76.47%	86.67%	81.25%	73.33%	84.62%	73.33%	78.57%
Doctorates	18	93.33%	94.12%	93.33%	93.30%	100.00%	88.24%	100.00%	93.75%	86.67%	100.00%	86.67%	92.86%
	36	96.67%	96.88%	96.67%	96.66%	100.00%	93.75%	100.00%	96.77%	93.33%	100.00%	93.33%	96.55%
	37	96.67%	96.88%	96.67%	96.66%	93.33%	100.00%	93.33%	96.55%	100.00%	93.75%	100.00%	96.77%
	43	86.67%	89.47%	86.67%	86.43%	100.00%	78.95%	100.00%	88.24%	73.33%	100.00%	73.33%	84.62%
	46	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%	93.33%
	48	83.33%	84.72%	83.33%	83.16%	93.33%	77.78%	93.33%	84.85%	73.33%	91.67%	73.33%	81.48%
	53	93.33%	94.12%	93.33%	93.30%	100.00%	88.24%	100.00%	93.75%	86.67%	100.00%	86.67%	92.86%
	62	63.33%	78.85%	63.33%	57.64%	100.00%	57.69%	100.00%	73.17%	26.67%	100.00%	26.67%	42.11%
	81	66.67%	67.94%	66.67%	66.06%	53.33%	72.73%	53.33%	61.54%	80.00%	63.16%	80.00%	70.59%
	83	90.00%	90.18%	90.00%	89.99%	93.33%	87.50%	93.33%	90.32%	86.67%	92.86%	86.67%	89.66%
	103	96.67%	96.88%	96.67%	96.66%	100.00%	93.75%	100.00%	96.77%	93.33%	100.00%	93.33%	96.55%

Table 5 *Analyst performance scores for both species and separately. Note that Precision, Recall, and F_1-scores in both species section are computed as the means of the two species*

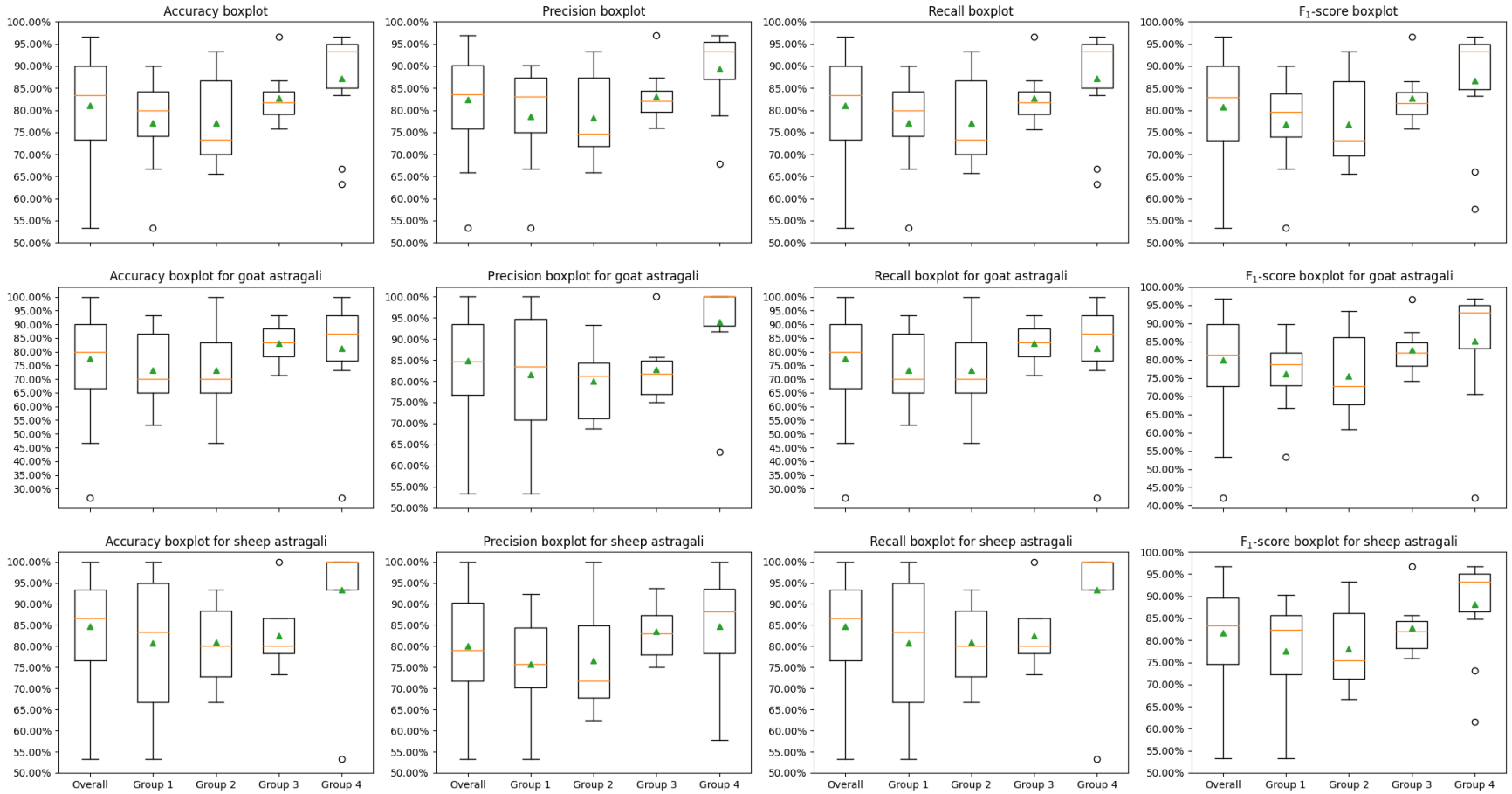

Fig. 3 *Boxplots of accuracy, precision, recall and F_1-scor for the analysts. First row shows the metrics for both species, whereas the second and third row show the metrics for the two species separately. The green triangle is the group (or overall) mean and the orange line is the median. Circles are outlier analysts*

Means and medians by expertise group for both species

Overall

	Accuracy	Precision	Recall	F_1-score
Mean (± SD)	81.15% (± 10.80%)	82.47% (± 10.34%)	81.15% (± 10.79%)	80.83% (± 11.17%)
Median	83.33%	83.48%	83.33%	82.86%

Group 1

	Accuracy	Precision	Recall	F_1-score
Mean (± SD)	77.08% (± 11.88%)	78.63% (± 12.68%)	77.08% (± 11.88%)	76.87% (± 11.81%)
Median	80.00%	83.13%	80.00%	79.54%

Group 2

	Accuracy	Precision	Recall	F_1-score
Mean (± SD)	77.13% (± 9.43%)	78.26% (± 9.24%)	77.14% (± 9.40%)	76.84% (± 9.61%)
Median	73.33%	74.67%	73.33%	73.21%

Group 3

	Accuracy	Precision	Recall	F_1-score
Mean (± SD)	82.82% (± 6.65%)	83.12% (± 6.68%)	82.80% (± 6.67%)	82.76% (± 6.67%)
Median	81.67%	82.01%	81.67%	81.61%

Group 4

	Accuracy	Precision	Recall	F_1-score
Mean (± SD)	87.27% (± 11.82%)	89.40% (± 9.07%)	87.27% (± 11.82%)	86.65% (± 13.12%)
Median	93.33%	93.33%	93.33%	93.30%

Table 6 *Means and medians by expertise groups*

The analyst performances for overall, sheep, and goat accuracies were further subjected to statistical tests to verify the observed between-group patterns. As Group 4 scores did not satisfy the assumption of normality in Shapiro-Wilk test (Table 8), the group-wise scores were transformed using Box-Cox power transformation, but the group-wise scores did not satisfy the equality of variance assumption in Levene's test (W = 8.2307, p = 0.0003). Using other power transformations did not help with transforming the data to satisfy the assumption of normality, and therefore Kruskal-Wallis H test (Kruskal and Wallis 1952) was undertaken on the untransformed accuracies to compare if at least one group's ranks dominate at least one other group. Kruskal-Wallis H test is followed by Dunn's test with Bonferroni correction to understand exactly which groups are different, as recommended as the post-hoc test (Dunn 1964; Dinno 2015).

This analysis demonstrates that there is no evidence in favour of any one group of analysts having higher overall accuracy than at least one other group (df = 3, H = 7.0575, p = 0.0701). Likewise, the test did not provide evidence for group-wise differences in goat astragalus accuracy (df = 3, H = 4.4727, p = 0.2147), but there does appear to be a group-wise difference in the analysts' ability to classify sheep astragali (df = 3, H =

9.7208, $p = 0.0211$). Thus, performing Dunn's test to discover the group-wise differences in sheep astragalus accuracies, it is shown that there is a significant (when $\alpha < 0.05$) difference between Group 2 and Group 4 analysts (Group 2 and Group 4: $p = 0.0269$), but not between any other two groups (Group 1 and Group 4: $p = 0.1876$; Group 3 and Group 4: $p = 0.1296$; for all other group-wise comparisons: $p = 1$).

Although the only between-groups difference identified in statistical testing was found between the analyst groups 2 and 4 regarding sheep astragalus accuracy, **Fig. 3** additionally implies that Group 4 outperforms the other groups in overall and sheep accuracy, but not in goat accuracy. It may be that the number of participants is not large enough to produce statistical significance in a rank-based statistical testing, even though graphical evaluation demonstrated a clear difference between groups.

Analyst consistency

Overall analyst consistency is 86.41%, which indicates a good, but not excellent consistency among the participants (**Table 9**). Groups 1 and 4 are more consistent than Groups 2 and 3, suggesting that more experienced analysts are more consistent. Considering that Shapiro-Wilk test for normality (**Table 10**) shows that Group 4 consistency scores do not satisfy the assumption of normality – and Box-Cox power transformation was unsuccessful – Kruskall-Wallis H test was again applied on the untransformed values. The outcome of this test does not support the hypothesis that any group of analysts is more consistent than any other group, however ($df = 3$, $H = 4.4148$, $p = 0.22$).

The Bland-Altman plot (**Fig. 4**) shows the difference in accuracies between the test and re-test samples (N = 10) against the mean accuracy of these same samples and therefore demonstrates the relationship between accuracy and consistency for all groups – Group 4 is the most consistent and the most accurate, whereas Group 1 analysts are more consistent but not more accurate than Group 2 and Group 3 analysts. Thus, although statistical testing did not demonstrate significant differences in consistency across the analyst groups, the graphical analysis suggests that the more experienced analysts are more consistent. It may be that group size is again a factor in the statistical tests, same as for the analyses of group-wise differences in accuracy.

Means and medians by expertise groups for each species

Overall

	Sheep accuracy	Goat accuracy	Sheep precision	Goat precision	Sheep recall	Goat recall	Sheep F_1-score	Goat F_1-score
Mean (± SD)	84.74% (± 13.29%)	77.56% (± 15.95%)	80.12% (± 11.75%)	84.83% (± 12.27%)	84.74% (± 13.29%)	77.56% (± 15.95%)	81.77% (± 10.52%)	79.90% (± 12.60%)
Median	86.67%	80.00%	78.95%	84.62%	86.67%	80.00%	83.33%	81.25%

Group 1

	Sheep accuracy	Goat accuracy	Sheep precision	Goat precision	Sheep recall	Goat recall	Sheep F_1-score	Goat F_1-score
Mean (± SD)	80.83% (± 17.25%)	73.33% (± 14.25%)	75.76% (± 12.39%)	81.51% (± 16.58%)	80.83% (± 17.25%)	73.33% (± 14.25%)	77.55% (± 12.31%)	76.19% (± 11.66%)
Median	83.34%	70.00%	75.74%	83.49%	83.34%	70.00%	82.29%	78.68%

Group 2

	Sheep accuracy	Goat accuracy	Sheep precision	Goat precision	Sheep recall	Goat recall	Sheep F_1-score	Goat F_1-score
Mean (± SD)	80.95% (± 9.93%)	73.33% (± 16.57%)	76.59% (± 12.62%)	79.92% (± 8.67%)	80.95% (± 9.93%)	73.33% (± 16.57%)	78.10% (± 8.71%)	75.59% (± 11.11%)
Median	80.00%	70.00%	71.83%	81.18%	80.00%	70.00%	75.34%	72.75%

Group 3

	Sheep accuracy	Goat accuracy	Sheep precision	Goat precision	Sheep recall	Goat recall	Sheep F_1-score	Goat F_1-score
Mean (± SD)	82.50% (± 8.68%)	83.09% (± 8.33%)	83.46% (± 6.97%)	82.79% (± 7.98%)	82.50% (± 8.68%)	83.09% (± 8.33%)	82.78% (± 6.57%)	82.75% (± 6.95%)
Median	80.00%	83.34%	82.94%	81.80%	80.00%	83.34%	82.01%	82.01%

Group 4

	Sheep accuracy	Goat accuracy	Sheep precision	Goat precision	Sheep recall	Goat recall	Sheep F_1-score	Goat F_1-score
Mean (± SD)	93.33% (± 13.66%)	81.21% (± 19.96%)	84.72% (± 12.08%)	94.07% (± 10.84%)	93.33% (± 13.66%)	81.21% (± 19.96%)	88.09% (± 11.21%)	85.22% (± 16.37%)
Median	100.00%	86.67%	88.24%	100.00%	100.00%	86.67%	93.33%	92.86%

Table 7 *Means and medians for sheep and goat astragali by expertise groups*

Shapiro-Wilk test for normality for accuracy

		Accuracy		Sheep accuracy		Goat accuracy	
	N	**W**	***p***	**W**	***p***	**W**	***p***
Overall	39	0.95911	0.1665	0.90581	0.0033	0.92088	0.0093
Group 1	8	0.89195	0.2440	0.92196	0.4459	0.93753	0.5870
Group 2	12	0.86348	0.0541	0.90530	0.1856	0.96024	0.7872
Group 3	8	0.88284	0.2005	0.87754	0.1784	0.91860	0.4186
Group 4	11	0.77146	0.0040	0.52396	0.0000	0.73547	0.0013

Table 8 *Shapiro-Wilk tests for normality for analysts' accuracy overall, and for sheep and goats separately. Only Group 4 analysts have a non-normal distribution*

Consistency		
Group	**Mean**	**SD**
Overall	86.41%	14.78%
Group 1	90.00%	7.56%
Group 2	80.00%	20.00%
Group 3	82.50%	14.88%
Group 4	93.64%	8.09%

Table 9 *Mean consistency based on the ten repeated specimens*

Shapiro-Wilk test for normality for consistency			
Group	**N**	**W**	***p***
Group 1	8	0.84891	0.0929
Group 2	12	0.88632	0.1057
Group 3	8	0.91981	0.4283
Group 4	11	0.75439	0.0024

Table 10 *Shapiro-Wilk test for normality for analysts' consistency. Again, only Group 4 analysts have a non-normal distribution*

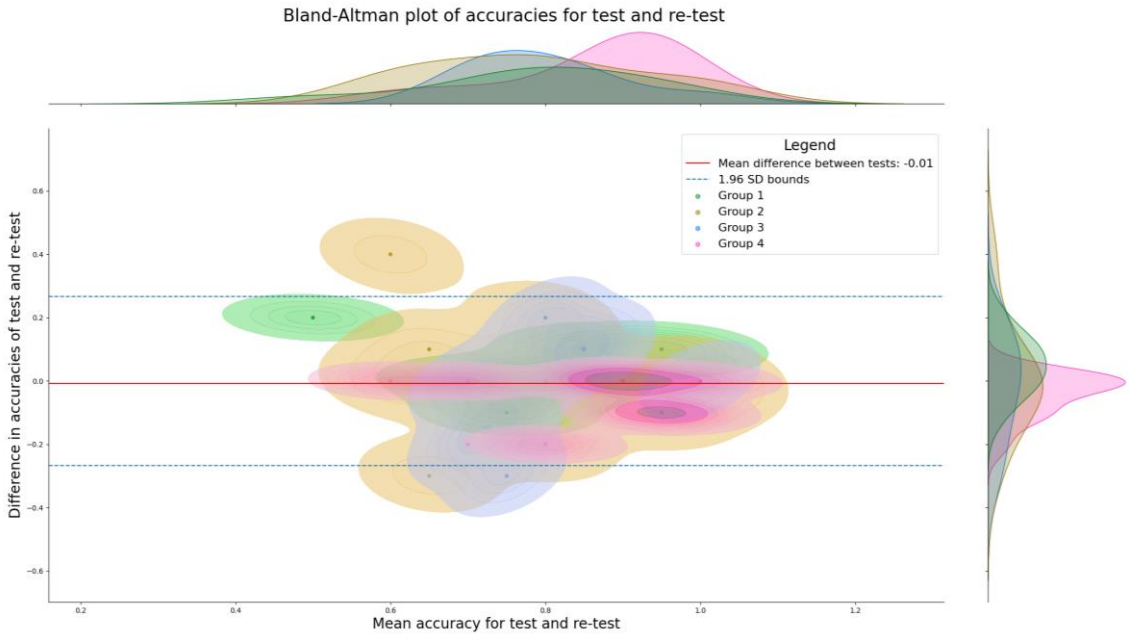

Fig. 4 *Bland-Altman plot of accuracies for test and re-test samples (N=10). The group densities are the kernel density estimates with the bandwidth smoothing adjusted down from 1 to 0.6. Best viewed in colour, available online*

Reference materials

The analysts were allowed to use any reference materials they thought would be helpful. Indeed, when prompted about the usefulness of reference materials after the main part of the study, only three analysts mentioned that reference materials were not useful, whereas 30 analysts found their reference materials having been useful and six analysts did not answer the question. The participants' reference material usage is itemised in **Table 11**, which shows that none of the analysts relied solely on goat astragalus in any of the media, but seven analysts considered it appropriate to rely on a physical sheep astragalus. Two of those seven analysts used reference images of sheep as well. The counts of different reference material types are shown in **Table 12**.

In addition, there appear to be group-wise trends as well, which are summarised in **Table 13** and **Table 14**. In short, Group 1 analysts use physical reference specimens more frequently than analysts in other groups, Group 2 and Group 3 analysts prefer using images and sketches, and Group 4 analysts refrain from using reference materials apart from reference texts. Regarding the analysts' use of reference texts, Groups 1, 2 and 4 prefer Zeder and Lapham's (2010) publication and Group 3 prefer Boessneck et al. (1964) or Boessneck (1969). Furthermore, **Fig. 5** shows that Group 3 analysts are more likely to use many different reference sources, whereas Group 4 analysts tend to use just one.

GLMM for reference materials

In the first instance, four different random effects structures are compared to a baseline GLM through LRT (**Table 15**). LRT is a suitable test for mixed-effects models when the model only contains random effects (Bolker et al. 2009). When applying LR testing to fixed effects models, the model parameters should be estimated through Maximum Likelihood rather than Restricted Maximum Likelihood and the sample size should be large (Bolker et al. 2009). As there is no clear definition of what constitutes a large sample size (we consider our sample size of 1,168 classification to be adequate, although these come from 39 participants) and because Bolker et al. (2009, p.132) "would recommend against using the LR test for fixed effects unless the total sample size and numbers of blocks are very large" for using LRT, we also report AIC scores. It was found that the best combination of random effects is one where Species ('Goat' = 0, 'Sheep' = 1) acts as random slope within Analyst random effect, while both Analyst and Specimen are the random intercepts. Species is also used as a fixed effect in this

Reference material use by analyst

Analyst	Group	Reference specimen	Reference images	Reference sketches	Reference model	Reference texts
3	1	Sheep	Both	No	No	Zeder & Lapham
10	1	No	No	No	No	None
51	1	No	No	No	No	Other
61	1	Sheep	Both	Both	Both	Zeder & Lapham
67	1	Sheep	No	No	No	Zeder & Lapham
68	1	Sheep	Sheep	Both	No	Boessneck
71	1	Sheep	No	No	No	Zeder & Lapham
84	1	No	No	No	No	None
1	2	No	Both	Both	No	Zeder & Lapham
11	2	No	Both	No	No	Zeder & Lapham
26	2	No	No	Both	No	Zeder & Lapham
29	2	Sheep	Sheep	No	No	Zeder & Lapham
44	2	No	Both	No	No	None
47	2	No	No	Both	No	Zeder & Lapham
56	2	No	No	Both	No	Boessneck
58	2	No	Both	No	No	None
60	2	No	No	Both	No	Zeder & Lapham
65	2	Sheep	No	No	No	Zeder & Lapham
66	2	No	No	No	No	Zeder & Lapham
96	2	No	No	No	No	Zeder & Lapham
2	3	No	Both	Both	No	Boessneck
4	3	No	Both	Both	No	Boessneck
69	3	Both	Both	Both	No	Other
82	3	No	No	No	No	Zeder & Lapham
95	3	No	Both	Both	No	Other
99	3	No	Both	Both	Both	Zeder & Lapham
100	3	No	No	Both	No	Boessneck
104	3	Both	Both	Both	No	Boessneck
18	4	No	No	No	No	Zeder & Lapham
36	4	No	No	No	No	Boessneck
37	4	No	No	Both	No	Zeder & Lapham
43	4	No	No	No	No	Zeder & Lapham
46	4	No	No	No	No	None
48	4	No	No	No	No	Zeder & Lapham
53	4	No	No	No	No	Other
62	4	No	Both	No	No	None
81	4	No	No	No	No	Zeder & Lapham
83	4	No	No	No	No	None
103	4	No	No	No	No	Boessneck

Table 11 *Reference materials used by the analysts*

configuration. The best random effects model is Null 4 to which the different types of reference materials are added as fixed effects.

As reference specimens, images, and sketches were not utilised uniformly (see **Table 11**), their levels were re-formatted to binary ('No' = 0, 'Yes' = 1) levels to indicate whether the analyst used a given reference material. Regarding reference text usage, Prummel and Frisch (1986) was not used by anyone and therefore reference text fixed

Count of analysts by reference material level

Reference specimen

	Sheep	Goat	Both	None
No of analysts	7	0	2	30

Reference images

	Sheep	Goat	Both	None
No of analysts	2	0	13	24

Reference sketches

	Sheep	Goat	Both	None
No of analysts	0	0	15	24

Reference model

	Sheep	Goat	Both	None
No of analysts	0	0	2	37

Reference texts

	Zeder & Lapham	Boessneck	Prummel & Frisch	Other	None
No of analysts	20	8	0	4	7

Table 12 *Counts of analysts for each level of all reference materials*

Number of analysts using reference materials

	Group 1	Group 2	Group 3	Group 4
Reference specimen	5	2	2	0
Reference images	3	5	6	1
Reference sketches	2	5	7	1
Reference model	1	0	1	0
Reference texts	6	10	8	8

Percentage of analysts (within group)

	Group 1	Group 2	Group 3	Group 4
Reference specimen	62.50%	16.67%	25.00%	0.00%
Reference images	37.50%	41.67%	75.00%	9.09%
Reference sketches	25.00%	41.67%	87.50%	9.09%
Reference model	12.50%	0.00%	12.50%	0.00%
Reference texts	75.00%	83.33%	100.00%	72.73%

Table 13 *Summary of groups' use of reference materials*

effect has four levels (‘None’ = 0, ‘Boessneck’ = 1, ‘Other’ = 2, ‘Zeder and Lapham’ = 3). 3D model usage was too infrequent to be of use, so it is not taken into consideration. The response variable is a binary variable corresponding to whether the classification was correct or not.

Number of analysts using reference texts

	Zeder & Lapham	Boessneck	Prummel & Frisch	Other	None
Group 1	4	1	0	1	2
Group 2	9	1	0	0	2
Group 3	2	4	0	2	0
Group 4	5	2	0	1	3

Percentage of analysts (within group)

	Zeder & Lapham	Boessneck	Prummel & Frisch	Other	None
Group 1	50.00%	12.50%	0.00%	12.50%	25.00%
Group 2	75.00%	8.33%	0.00%	0.00%	16.67%
Group 3	25.00%	50.00%	0.00%	25.00%	0.00%
Group 4	45.45%	18.18%	0.00%	9.09%	27.27%

Table 14 *Summary of groups' use of reference texts*

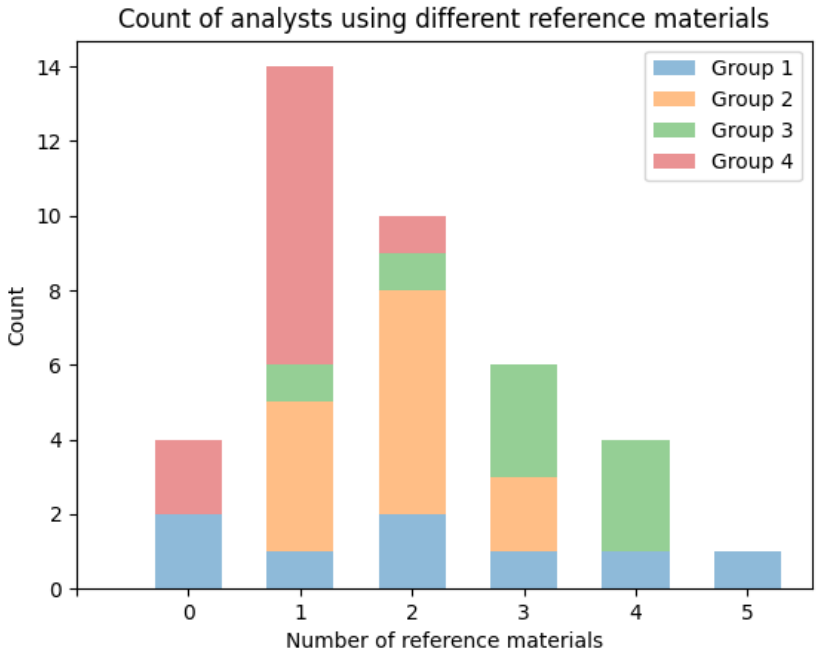

Fig. 5 *A bar plot of analysts' number of used reference resources. Best viewed in colour, available online*

Using each reference material type individually as the fixed effect, it was found that only reference specimens (Ref 3 model) and reference images (Ref 4 model) added more information to the Null 4 model (Table 16) in LRT and AIC. Furthermore, it was found that adding all reference materials (Ref 5 model) and analyst grouping (Ref 6 model) as fixed effects is also justified given the LRT and AIC results in Table 16. Incorporating interactions of the different reference materials as well as the analyst groupings, however, resulted in an overfit model and this model is therefore not reported here. Although the Likelihood Ratio Test did not show a significant improvement for Ref 6 model over Ref 5 (χ^2 = 7.2753, Df = 3, Pr(> χ^2) = 0.0636), the focus is nonetheless on Ref

Random effects structures (null model selection)

Model	Effects	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Baseline	None (GLM model)	1132.2	1137.3	-565.11	1130.2			
Null 1	Analyst (intercept)	1116.9	1127	-556.45	1112.9	17.324	1	3.15e-05
Null 2	Specimen (intercept)	1060	1070.1	-527.99	1056	74.253	1	< 2.2e-16
Null 3	Analyst (intercept), Specimen (intercept)	1045.5	1060.7	-519.75	1039.5	90.731	2	< 2.2e-16
Null 4	Analyst (intercept), Species (slope within Analyst), Specimen (intercept)	1023.6	1054.0	-505.81	1011.6	118.61	5	< 2.2e-16

Table 15 *Random effects structure selection. Each null model was compared to the baseline GLM. The best random effects structure is highlighted in grey*

Fixed effect models compared to Null 4

Model	Fixed effects	Comparison	AIC	BIC	logLik	deviance	χ^2	Df	Pr(> χ^2)
Ref 1	Species + Reference texts	Null 4	1027.2	1072.8	-504.60	1009.2	2.426	3	0.4888
Ref 2	Species + Reference sketches	Null 4	1024.2	1059.6	-505.09	1010.2	1.4386	1	0.2304
Ref 3	Species + Reference specimen	Null 4	1021.0	1056.5	-503.51	1007.0	4.5982	1	0.0320
Ref 4	Species + Reference images	Null 4	1019.5	1055	-502.77	1005.5	6.0832	1	0.0137
Ref 5	Species + Reference texts + Reference sketches + Reference specimen + Reference images	Null 4	1021.9	1082.7	-498.97	997.94	13.677	6	0.0335
Ref 6	Species + Reference texts + Reference sketches + Reference specimen + Reference images + Group	Null 4	1020.7	1096.6	-495.33	990.66	20.953	9	0.0129

Table 16 *Likelihood ratio tests showing that additional reference materials contribute to the model. Null 4 is the best fit null model from* ***Table 15****. Fixed effects models that are better fit than Null 4 are highlighted in grey*

Maximum likelihood estimates for Ref 5 model

Fixed effects

Parameter	Effect	Coefficient	Std. Error	z-value	Pr(>|z|)
β_0	Intercept	2.2611	0.4348	5.201	1.99e-07
β_1	Species: Sheep	0.7915	0.4350	1.819	0.0689
β_2	Boessneck	0.5909	0.5247	1.126	0.2601
β_3	Other text	0.2904	0.5624	0.516	0.6056
β_4	Zeder and Lapham	-0.2205	0.4136	-0.533	0.5939
β_5	Reference specimen	-0.4497	0.3330	-1.351	0.1768
β_6	Reference images	-0.6148	0.3238	-1.898	0.0577
β_7	Reference sketches	-0.3342	0.3320	-1.007	0.3140

Random effects

Parameter	Effect	Variance	Std. Dev.	Corr.
I_{0i}	Specimen (Intercept)	2.5299	1.5906	
S_{0s}	Analyst (Intercept)	0.7184	0.8476	
S_{1s}	Species: Sheep (Slope within Analyst)	2.6757	1.6357	-0.84

Table 17 *Maximum likelihood estimates for Ref 5 model. Significant coefficients highlighted in grey*

Multicollinearity for fixed effects in Ref 6

Fixed effect	VIF	Increased SE	Tolerance
Species	1.01	1.00	0.99
Group	3.13	1.77	0.32
Reference texts	2.27	1.51	0.44
Reference specimen	1.93	1.39	0.52
Reference images	1.75	1.32	0.57
Reference sketches	1.78	1.33	0.56

Table 18 *Multicollinearity measures for fixed effects included in Ref 6 model*

6 model because we are interested in the Group level effect and the AIC score is lower. Any question of whether Simpson's Paradox has occurred is negated by the fact that none of the Ref 5 model coefficients for reference materials are significant Table 17, nor are they hugely different from Ref 6 coefficients. In other words, whether the data is split by Group (as in Ref 6 model) or all analysts were considered as part of a single group (as in Ref 5 model) our interpretation of the results remains the same. The VIF and tolerance values for species, expertise groups, reference texts, sketches, specimen, and images indicate that these fixed effects express negligible levels of multicollinearity in Ref 6 model (**Error! Reference source not found.**). The formal definition of Ref 6 model is provided in Online Resource 9.

Maximum likelihood estimates for Ref 6 model

Fixed effects

Parameter	Effect	Coefficient	Std. Error	z-value	Pr(>|z|)
β_0	Intercept	1.8266	0.4802	3.8040	0.0001
β_1	Species: Sheep	0.7852	0.4477	1.7540	0.0794
β_2	Boessneck	0.1842	0.5147	0.3580	0.7204
β_3	Other text	-0.0512	0.5435	-0.0940	0.9249
β_4	Zeder and Lapham	-0.3337	0.3993	-0.8360	0.4033
β_5	Reference specimen	-0.0842	0.3664	-0.2300	0.8183
β_6	Reference images	-0.5471	0.3186	-1.7170	0.0860
β_7	Reference sketches	-0.2191	0.3213	-0.6820	0.4953
β_8	Group 2	0.1985	0.3795	0.5230	0.6010
β_9	Group 3	0.6382	0.4537	1.4060	0.1596
β_{10}	Group 4	1.0188	0.4339	2.3480	0.0189

Random effects

Parameter	Effect	Variance	Std. Dev.	Corr.
I_{0i}	Specimen (Intercept)	2.6887	1.6397	
S_{0s}	Analyst (Intercept)	0.8164	0.9036	
S_{1s}	Species: Sheep (Slope within Analyst)	2.9412	1.7150	-0.93

Table 19 *Maximum likelihood estimates for Ref 6 model. Significant coefficients highlighted in grey*

Ref 6 model produces log odds estimates of a correct answer given the species of the specimen, analyst expertise group membership, and the combination of reference materials used by the analyst, conditional on specimen and analyst random effects. Observing the coefficients for this model (Error! Reference source not found.), it is notable that none of the individual reference materials have a significant and positive effect on the probability of a correct answer as is the case for Ref 5 model. The only variable with a significant effect on the log odds of a correct answer is whether the analyst is in Group 4.

Self-reported confidence

Self-reported confidence was measured on the scale from 1 to 100 with a default value of 51. Out of the total of 1,168 classifications made by the participants, the default confidence value was not changed in 225 instances. Thus, this section involves the

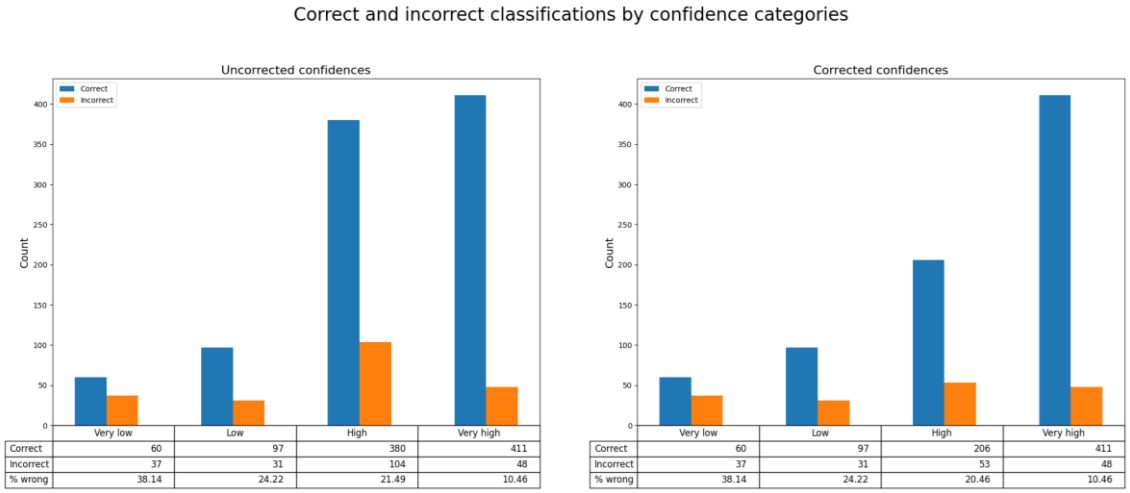

Fig. 6 *Correct and incorrect classifications by confidence categories for the corrected and uncorrected sets. Best viewed in colour, available online*

evaluation of both corrected (N=943) and uncorrected (N=1,168) confidence values in separate tests, where the corrected subset of classifications simply means the removal of classifications with a confidence value of 51. Furthermore, although the self-reported confidence values represent interval data, it can be argued that they should be represented through Likert-like scale due to obtaining the values using a sliding scale. Thus, the classifications are placed into four confidence bins: 1) confidence values [1-25] are considered 'Very low'; 2) confidence values of (25-50] are 'Low'; 3) (50-75] are 'High'; and 4) confidence scores (75-100] are labelled 'Very high' confidence. Here, parentheses indicate exclusivity and square brackets inclusivity.

For both corrected and uncorrected sets of classifications, χ^2-tests were performed to see if there are statistically significant differences in the expected and observed frequencies of correct and incorrect answers for the four different confidence categories. The results of this test are shown in Table 20 – the χ^2-tests were statistically significant in both cases when $\alpha < 0.05$, which means that the number of correct and incorrect classifications are dependent on the confidence category. Although the association between the confidence category and the classification correctness is not particularly strong according to Cramér's V measure, the relationship between confidence categories and correct answers is quite clear in **Fig. 6**, especially for corrected confidence scores. This plot shows that as analysts are more confident about their classification, the ratio of incorrect to correct answers is reduced, although the error rate is still 10.46% in the highest confidence category. Thus, replacing sheep/goat identification with strict species identifications accompanied by self-assessed

χ^2-test for confidence categories and correctness of answer

All analysts

	Uncorrected confidences				Corrected confidences			
	Observed		χ^2 expected		Observed		χ^2 expected	
Category	Corr.	Incorr.	Corr.	Incorr.	Corr.	Incorr.	Corr.	Incorr.
Very low	60	37	78.73	18.27	60	37	79.62	17.38
Low	97	31	103.89	24.11	97	31	105.06	22.94
High	380	104	392.84	91.16	206	53	212.58	46.42
Very high	411	48	372.55	86.46	411	48	376.74	82.26

	χ^2	*p*	df	N	χ^2	*p*	df	N
Statistics	49.383	1.10e-10	3	1,168	48.94	1.30e-10	3	943

Effect sizes

Corrected test	Cramér's V	Uncorrected test	Cramér's V
All analysts	0.2056	All analysts	0.2278

Table 20 *χ^2-test for confidence categories and correctness of answer. The test is significant for both uncorrected and corrected set of confidence scores*

confidence scores for each bone could increase the overall accuracy if one only takes those classifications with the highest confidence category into account. Reduced error rate would have the effect of reducing noise in subsequent analyses and therefore enable statistically more powerful zooarchaeological research.

Applying different thresholds to the confidence scores (x-axis) and plotting them against the mean accuracy for the classifications with a confidence score above that threshold (y-axis), it can be shown (Fig. **7**A and Fig**. 7**C) that the self-reported confidence threshold that maximizes the classification accuracy in the present task is 96, when the mean accuracy reaches 95.11%. However, the mean accuracy begins to plateau around the 85 mark, when the mean accuracy is 93.33%. In addition, there is a large difference in the number of classifications that are taken into account at these two different thresholds – when the threshold is set at 96 (inclusive), we are only including 15.16% of the classifications in the corrected set of classifications, whereas at 85 (inclusive) level we include 31.81% of classifications. To include at least 50% of the classifications, the threshold has to be set to 75 (inclusive), when the mean classification accuracy is 88.77%. To include at least 75% of the classifications, the threshold must be lowered to 57 (inclusive), when the mean classification accuracy is 86.08%.

In Fig. **7**B and Fig**. 7**D, the mean accuracy for classifications below the given confidence threshold is presented for the uncorrected and corrected datasets, respectively. These

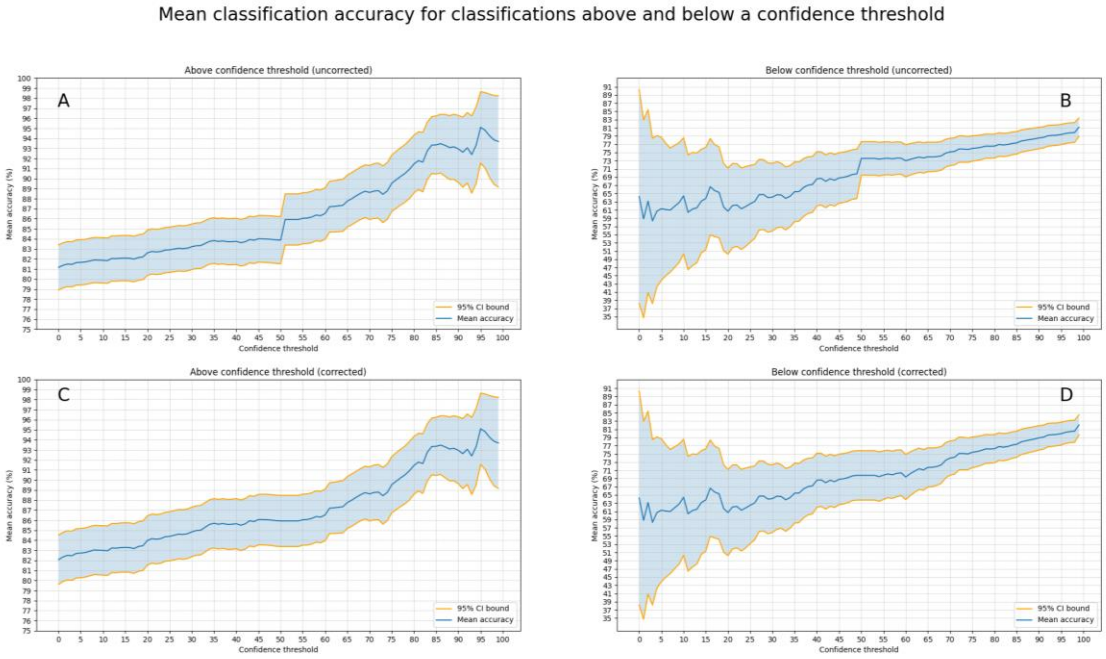

Fig. 7 *A) Mean accuracy for all classifications above the confidence threshold, without correction for confidence scores. B) Mean accuracy for all classifications below the confidence threshold, without correction for confidence scores. C) Mean accuracy for all classifications above the confidence threshold, with correction for confidence scores. D) Mean accuracy for all classifications below the confidence threshold, with correction for confidence scores. Best viewed in colour, available online*

figures demonstrate that lower confidence scores are less informative than higher confidence scores because of the large confidence intervals – there are far fewer classifications with low confidence scores, but classifications with low confidence scores cannot immediately be said to be misclassified. Instead, Fig. **7**B and Fig**. 7**D mainly reflect the methodology for obtaining the self-reported confidence scores in that the lower end of the sliding scale was labelled 'Guess', so the expected accuracy for a classification with a confidence score of 1 is 50%. The trend in these figures is that the analysts' accuracy is slightly above the expected accuracy, but the expected accuracy is within the 95% confidence interval of the empirical data.

Between-group differences in confidence scores

In addition to the correct and incorrect answers being dependent on the confidence scores for all answers, there may exist between-group differences in this relationship. The bar plots in **Fig. 8** show that the error rates for analysts in Groups 3 and 4 are much lower for the 'Very high' confidence category than in the other three confidence categories. Although the overall pattern is similar for Groups 1 and 2, the effect is much smaller. χ^2-tests were not suitable for finding out the relationship within analyst groups

as the expected values were less than five in more than 20% of the cells for all but Group 2, which is a commonly cited minimum threshold for using χ^2-tests for independence

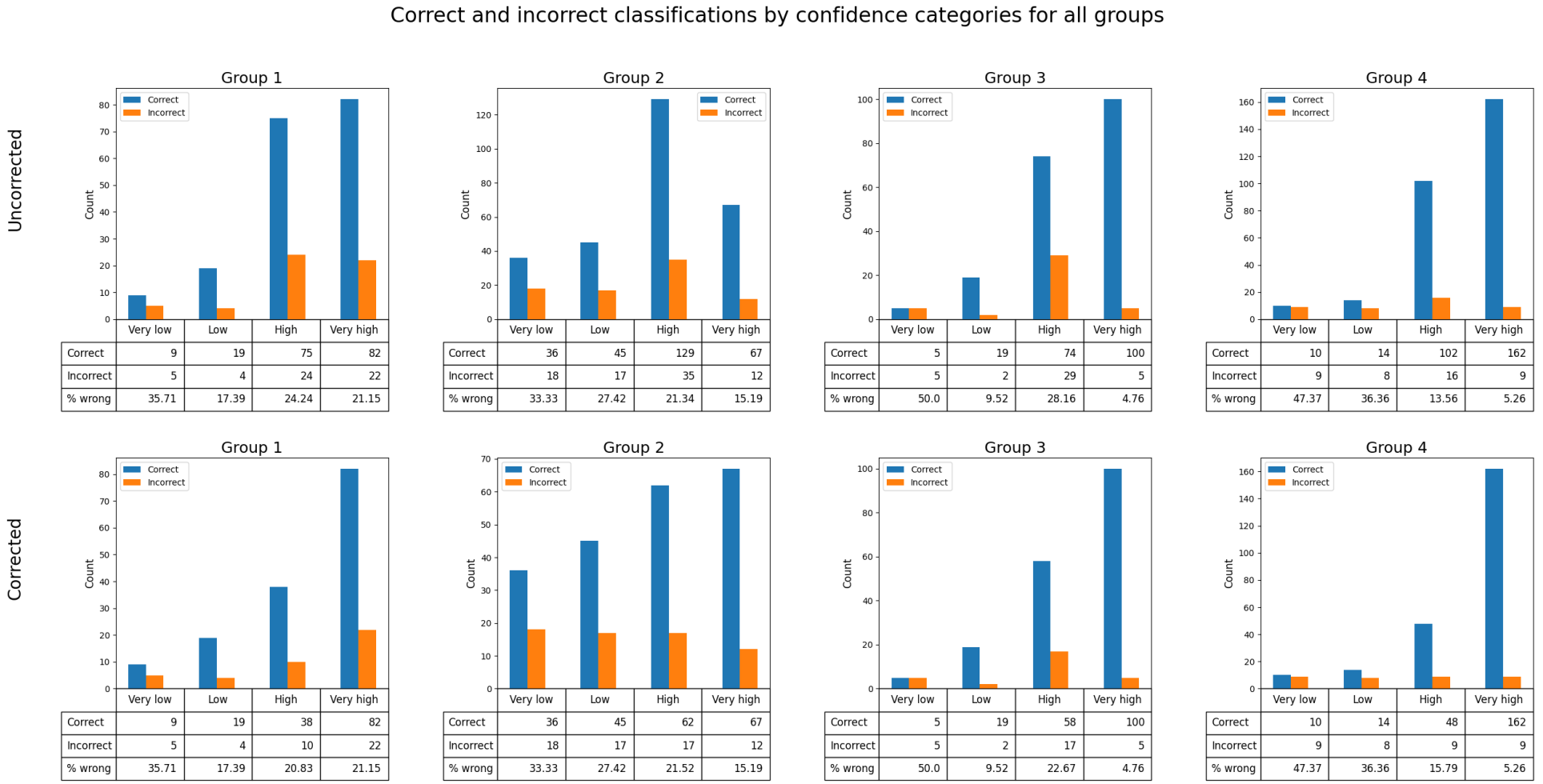

Fig. 8 *Correct and incorrect classifications by corrected confidence categories for all analyst groups. Best viewed in colour, available online*

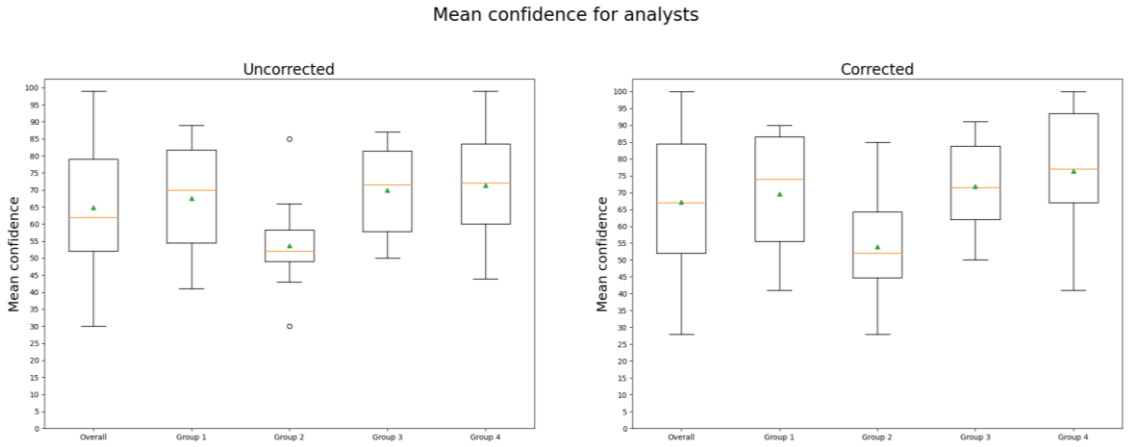

Fig. 9 *Boxplot of mean confidences for the analysts. The green triangle indicates mean value, while the orange line reflects median. Circles are outlying analysts*

(Bewick et al. 2004). Thus, it was opted to use the continuous confidence scores rather than the categorical values in the comparison of analyst expertise groups.

These continuous confidence values did not satisfy the assumption of normal distribution (Table 22) and therefore Kruskal-Wallis H test (Table 23) followed by Dunn's test (Table 21) with Bonferroni correction was performed. These tests demonstrate that Group 2 confidence scores are significantly different from the other groups when α = 0.05, and Group 3 confidence scores are additionally significantly different from Group 4 confidence scores when only taking the corrected set of confidence values into account. This result implies that less experienced analysts are less confident about their decisions, as would be expected. However, taking the analysts' mean confidence scores and plotting them with boxplots (**Error! Reference source not found.**), it can be shown that Group 3 analysts' mean confidences are not much different from Group 4 analysts' mean confidences. This difference in the conclusions drawn from the boxplot and Dunn's test can be explained by the fact that raw confidence scores of individual answers were used in Dunn's test, whereas the boxplots display analysts' mean confidences. Thus, while analysts in Group 3 are not, on average, less confident than analysts in Group 4, it may be that Group 3 participants were shown astragali that were, by chance, morphologically more ambiguous than those shown to Group 4 analysts.

Between-species differences in confidence scores

Next, it is aimed to answer the question of which species the analysts are more confident about classifying. This part of the analysis was devised due to seven analysts opting to use a physical sheep astragalus as a reference and it was of interest to find out

Shapiro-Wilk's test for normality for analyst mean confidence (group-wise)

		Uncorrected		Corrected	
	N	W	*p*	W	*p*
Group 1	8	0.93308	5.56e-09	0.87640	2.45e-11
Group 2	12	0.96900	6.27e-07	0.93930	3.42e-09
Group 3	8	0.94836	1.68e-07	0.91187	7.18e-10
Group 4	11	0.90987	3.81e-13	0.83412	2.77e-16

Table 21 *Shapiro-Wilk's test for normality for analysts' mean confidence scores for all groups*

Kruskal-Wallis H test for confidence scores across analyst groups

	df	H	*p*
Corrected continuous	3	107.14	4.52e-23
Uncorrected continuous	3	108.02	2.93e-23

Table 23 *Kruskal-Wallis H test for confidence scores across analyst experience groups*

Dunn's test for continuous confidence scores across analyst groups

	Uncorrected				Corrected			
	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
Group 1	1.0000	1.13e-10	1.0000	0.3414	1.0000	5.50e-12	1.0000	0.5056
Group 2	1.13e-10	1.0000	1.81e-13	1.85e-20	5.50e-12	1.0000	4.62e-10	8.45e-22
Group 3	1.0000	1.81e-13	1.0000	1.0000	1.0000	4.62e-10	1.0000	0.0489
Group 4	0.3414	1.85e-20	1.0000	1.0000	0.5056	8.45e-22	0.0489	1.0000

Table 22 *Dunn's test for continuous confidence scores across analyst groups. Statistically significant results highlighted in grey. Values are Bonferroni corrected p-values at* $\alpha = 0.05$

whether the analysts were more confident about classifying sheep astragali. This hypothesis was also informed by the general observation that suitable goat astragali were harder to find for inclusion in the present study and that the literature on goat bone development is not as extensive as for sheep, which leads us to believe that there may be a bias in the zooarchaeological community towards a higher confidence in identifying sheep astragali.

First, it was tested whether the confidence scores for sheep and goat bones followed normal distributions. This test was performed using Shapiro-Wilk test, which shows that the confidence scores for sheep (uncorrected: W = 0.9429, *p* = 3.39e-14; corrected: W = 0.8918, *p* = 1.05e-17) and goat (uncorrected: W = 0.951, *p* = 5.25e-13; corrected: W = 0.9062, *p* = 1.81e-16) violates the normality assumption, and thus, Mann-Whitney U test was performed (Table 25). This test supports the hypothesis that analysts are less confident about their decision when classifying goats, although the effect size (Cohen's

Mann-Whitney U test for confidences between sheep and goats

	Median (sheep)	Median (goat)	Mean (sheep)	Mean (goat)	N (sheep)	N (goat)	U	*p*	Cohen's *d* (effect size)
Uncorrected	70	66	66.46	63.29	584	584	182491	0.0372	0.1251
Corrected	76	73	70.16	66.21	471	472	121386	0.0143	0.1459

Table 25 *Mann-Whitney U test for confidences between sheep and goats*

Mann-Whitney U test for sheep and goat confidences, group-wise

Group		Median (sheep)	Median (goat)	Mean (sheep)	Mean (goat)	N (sheep)	N (goat)	U	*p* adj.	Cohen's d (effect size)
Uncorrected	1	70	70	69.59	65.68	120	120	7779.5	1	0.1571
	2	51	51	55.56	51.75	179	180	17696.0	0.4179	0.1548
	3	72.5	70	71.44	68.56	120	119	7709.0	1	0.1327
	4	76	77	72.36	70.35	165	165	13986.5	1	0.0811
Corrected	1	82	75	75.25	69.16	92	97	5223.5	0.1666	0.2320
	2	67.5	61	57.01	51.98	136	138	10513.0	0.341	0.1786
	3	79	73	74.14	70.91	106	105	6113.0	0.8669	0.1480
	4	83	82.5	76.73	75.19	137	132	9014.5	1	0.0609

Table 24 *Mann-Whitney U test for sheep and goat confidence scores for each analyst group. Bonferroni correction is used to compute the adjusted p-value*

d) suggests that the species of the bone has only a small effect on confidence (see Lovakov and Agadullina 2021). Furthermore, even though the median and mean confidence scores for sheep tend to be larger than for goat, within-group Mann-Whitney U tests did not produce any statistically significant results for uncorrected or corrected sets (Table 24).

Specimen difficulty

Next, we want to identify difficult and easy specimens, which is achieved by computing the Difficulty Index (also known as Facility Index) for each specimen. Difficulty Index simply measures the proportion of answers that were correct for a given specimen, meaning that the lower the index, the harder the specimen. This index is first used in comparing the levels of difficulty between the two species because we observed that species impacted the analysts' confidence scores and accuracies. We then explore the between-group differences in Difficulty Index as some groups may have been shown easier specimens by chance. Finally, we perform a correlation test to see if difficulty

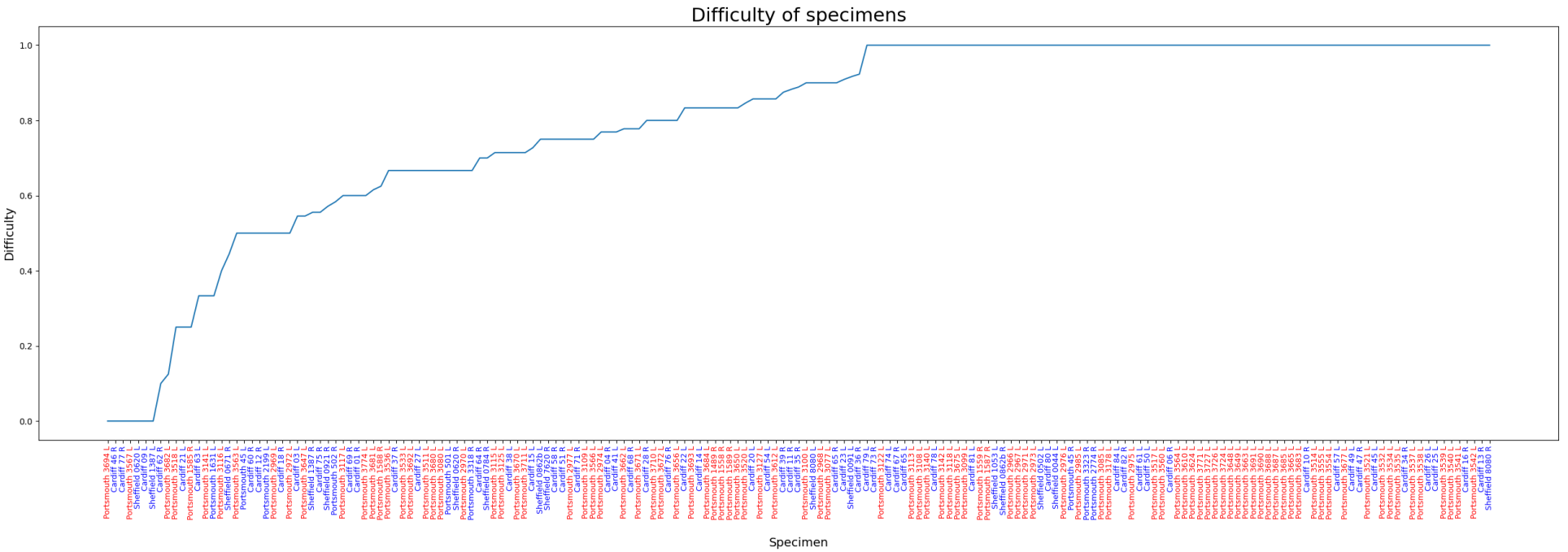

Fig. 10 *Difficulty Index for each specimen. The specimen names are constructed as follows: Portsmouth refers to Historic England collection, Sheffield to Sheffield University, and Cardiff to National Museum Wales. The number in the name refers to their catalogue ID and L and R indicate the side of the animal. Red labels are sheep and blue labels are goats. Use these codes to identify the animal in* ***Error! Reference source not found.*** *– for specimens from Cardiff, the last two digits of the ID are the same as here. Best viewed in colour, available online*

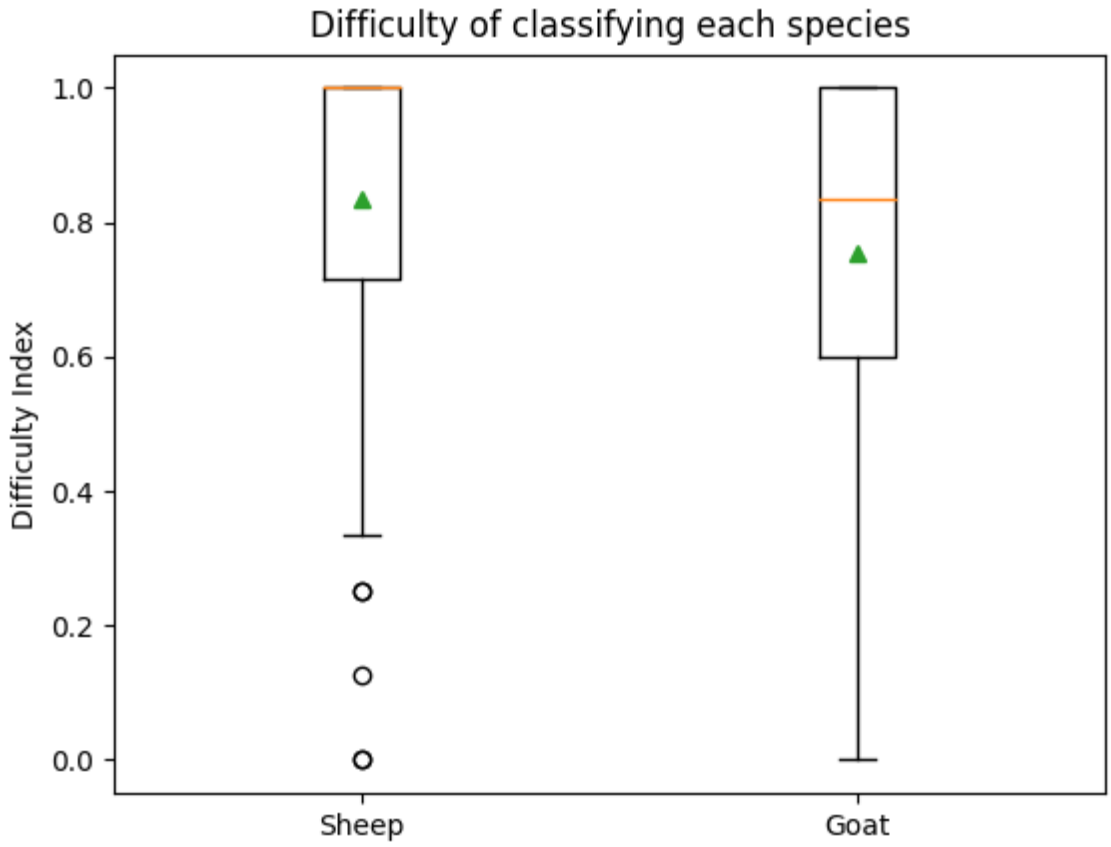

Fig. 11 *Boxplot of the Difficulty Index. Outliers are circles, the green triangle is the within group mean, and the orange line is the within group median*

correlates with analysts' mean confidence. However, note that specimens with low numbers of classifications may over- or under-estimate the difficulty of the specimen.

Between-species comparison

The Difficulty Index is shown for each specimen in Fig. **10**, demonstrating that a large proportion of specimens are very easy and only a small number are very difficult. The specimens with red labels in the x-axis represent sheep and the blue labels represent goat astragali. Difficulty Index violates the assumption of normality in Shapiro-Wilk test for both species (Sheep: N = 99, W = 0.741, p = 6.33e-12; Goat: N = 84, W = 0.8161, p = 7.6e-09). Thus, Mann-Whitney U test was performed, showing that sheep are significantly (at $\alpha < 0.05$) easier than goats for the analysts to classify (N_{sheep} = 99, N_{goat} = 84, $Mean_{sheep}$ = 0.8344, $Mean_{goat}$ = 0.7546, U = 4896, p = 0.03). The Cohen's d effect size is small to moderate (d = 0.3093) and Fig. **11** indicates that the harder sheep astragali are outliers. The spread of difficulties is also shown as a kernel density estimate plot in Fig. **12** for all specimens.

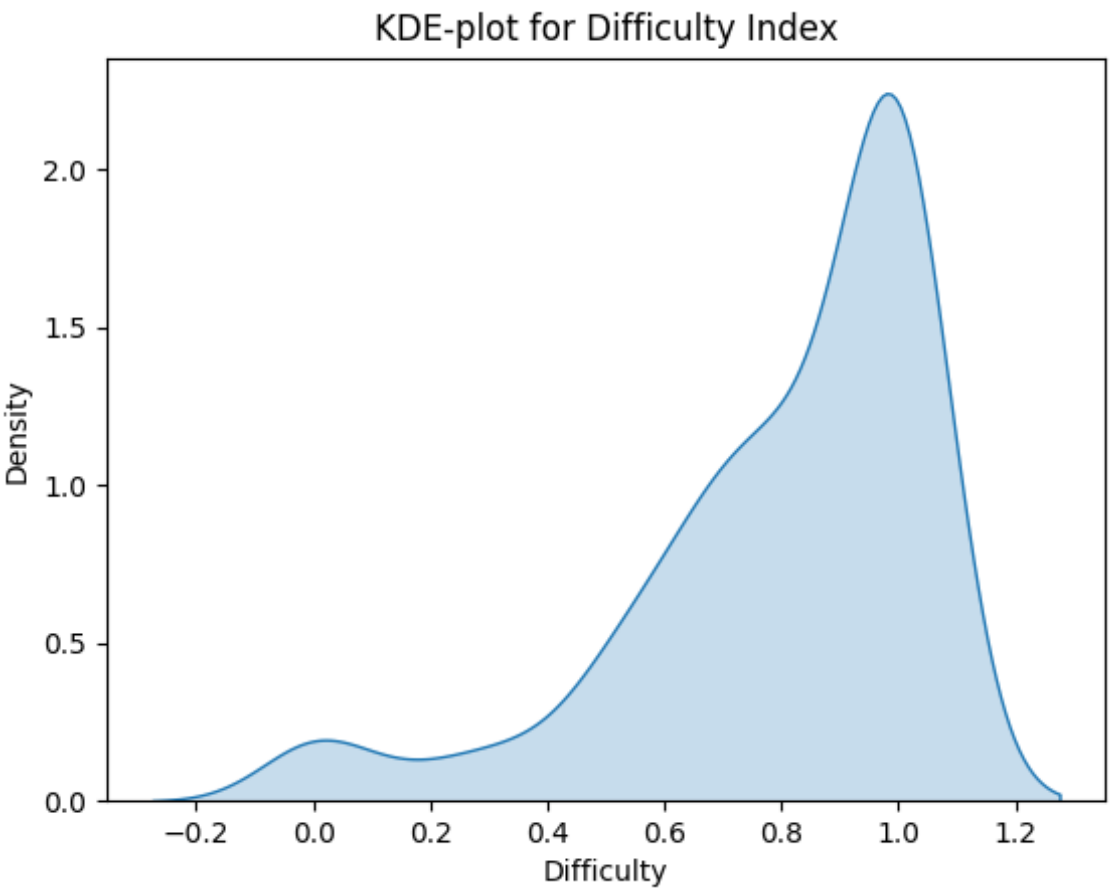

Fig. 12 *Kernel Density Estimate for Difficulty Index for all specimens (N=183)*

Between-group comparison

As it was observed that there are differences in analyst performances and confidence scores between expertise groups (albeit not statistically significant), it is of interest to verify that the difficulty of the specimens seen by analysts in different groups does not vary significantly between groups. After computing the mean difficulty of the specimens seen by each analyst, the within-group difficulties passed the Shapiro-Wilk tests for normality for all groups (Group 1: W = 0.9358, p = 0.5698; Group 2: W = 0.9150, p = 0.2473; Group 3: W = 0.9702, p = 0.8999; Group 4: W = 0.9506, p = 0.6515) and Levene's test for equality of variance (F = 0.5304, p = 0.6644), and thus, one-way ANOVA is performed. The result of the ANOVA test is that the specimen difficulties between analyst expertise groups are not significantly different (F = 0.2410, p = 0.8672), and this conclusion is supported by the kernel density estimate plot for group-wise item difficulty (**Fig. 13**). We can therefore be confident in the assumption that the random selection of specimens for each analyst did not significantly affect any of the analyses.

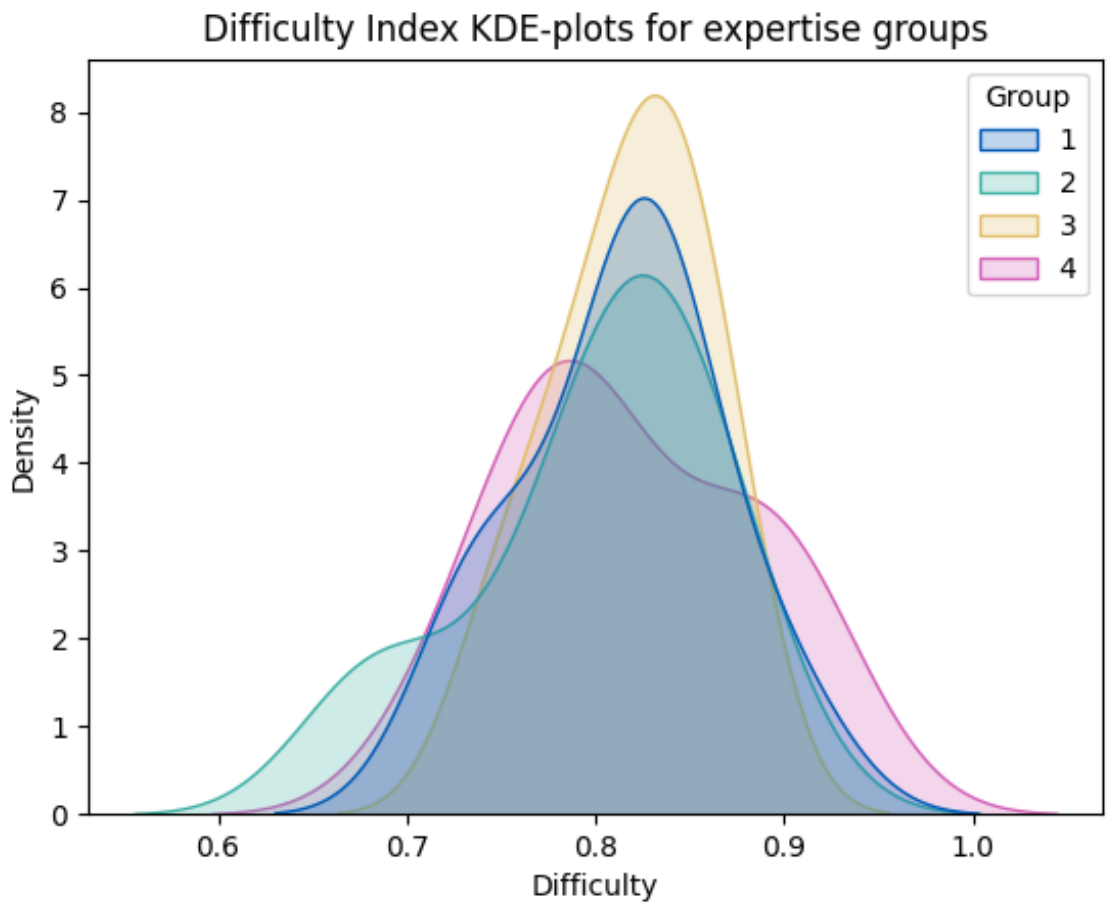

Fig. 13 *Difficulty Index KDE-plot for expertise groups. The mean specimen difficulties were computed for each analyst based on all 30 test items and these are represented by group in this figure. Best viewed in colour, available online*

Correlation tests

The specimen difficulty is expected to correlate with the analysts' mean confidence. Because we are using the corrected set of confidence scores here, the tests are performed on a reduced number of specimens. As the specimen mean confidences violate the assumption of normality in Shapiro-Wilk test (W = 0.9731, *p* = 0.0016, N = 179) as does the specimen Difficulty Index (W = 0.7390, *p* = 1.71e-16, N = 179), Spearman's rank correlation test is performed, which demonstrates a positive correlation (r = 0.3778, *p* = 1.85e-07, N = 179). Thus, as specimens have a higher Difficulty Index (i.e. they are easier), the analysts are more confident about their classification, although the correlation is not very strong.

The strength of the correlation between mean confidence and item difficulty is not explained by the number of responses for a given specimen, as shown by Spearman's correlation test between the number of responses and difficulty index (r = -0.0186, *p* = 0.803, N = 183). This is further demonstrated in **Error! Reference source not found.**, where the mean Difficulty Index for specimens (triangles in boxplots) are plotted against

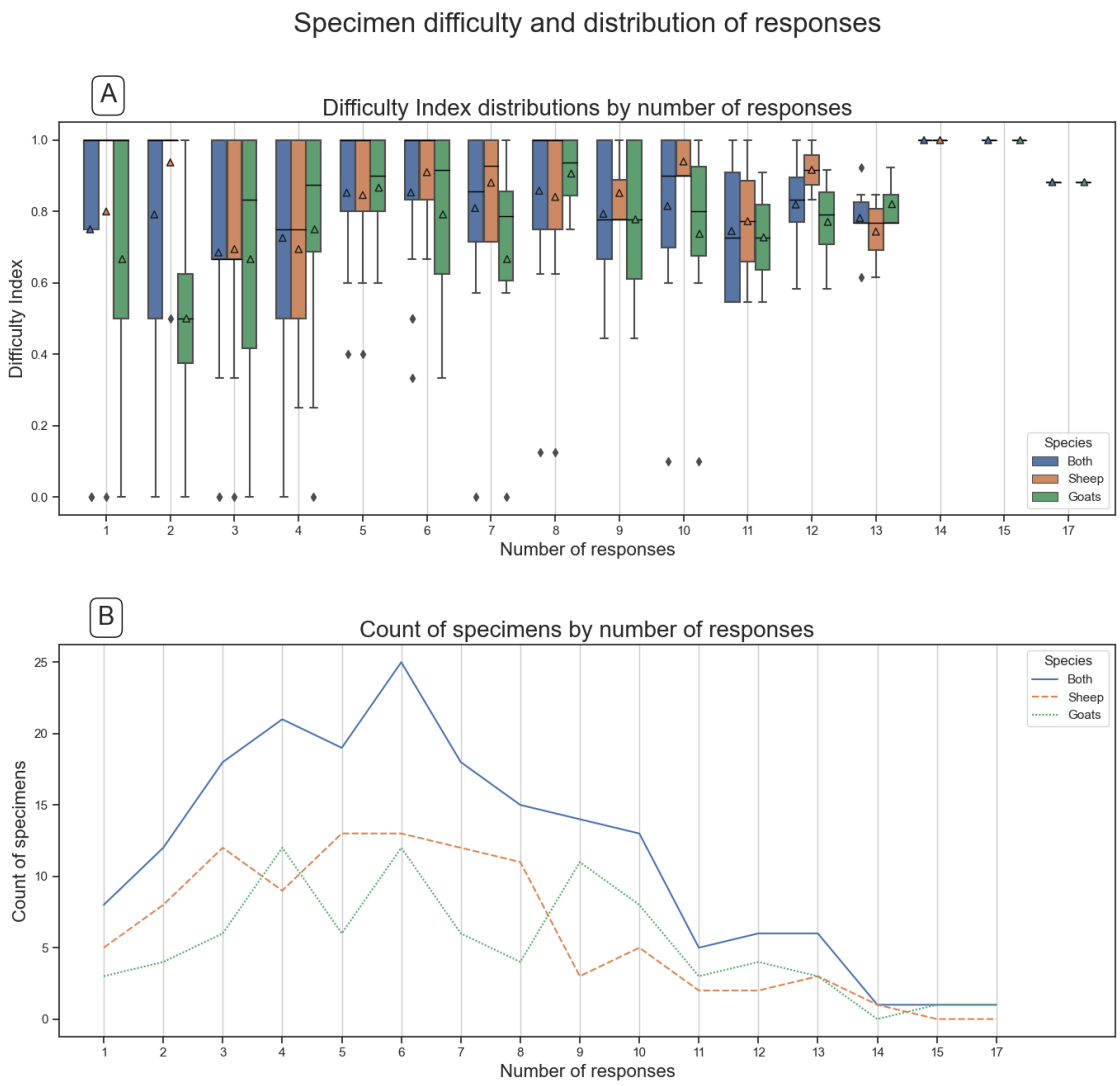

Fig. 14 *A) Relationship between the number of responses and the Difficulty Index. The coloured triangles are means associated with the boxplots and the diamonds are the outlier specimens in terms of difficulty. Note how the means (triangles) do not vary significantly despite the change in the number of responses. B) The count of specimens by their number of responses. The number of responses contains both original answers and all consistency tests. Best viewed in colour, available online*

the number of responses by analysts (x-axis). This figure shows how the mean (triangles) Difficulty Index remains relatively stable between 0.7 and 0.9 for all items with 13 or fewer responses while the number of specimens ranges from five to 25 in the same interval. The three specimens with 14 or more responses appear to have been very easy for the analysts, as two of them have a Difficulty Index of 1 and the specimen with 17 responses has a Difficulty Index of just below 0.9. **Error! Reference source not found.** additionally shows this phenomenon for both species, demonstrating how the mean Difficulty Index (triangles) is more often higher for sheep when specimens are grouped by the number of times they were attempted.

Qualitative analysis of the analysts' areas of attention

As part of the study, the participants were asked to paint on top of the images of the bones the features that they used in their classifications. This resulted in hundreds of drawings for both species and these drawings are used here in a qualitative manner with respect to the zooarchaeological descriptions typically used for differentiating the sheep and goat astragali. First, the overall patterns are explored and then we turn our attention to the easiest and hardest specimens of each species.

Overall analyst attention

The analysts' drawings of the regions of the bones pertinent to the classification (**Error! Reference source not found.**A) follow very closely the regions included in the descriptive criteria defined by Zeder and Lapham (2010) and Boessneck (1969; Boessneck et al. 1964) – the medial articular ridge in dorsal and plantar views, the distal articular surface in lateral view, and the proximo-plantar projection in medial view are the most often drawn on regions for both species. The number of drawings for each view is very similar for the two species (see counts in **Error! Reference source not found.**A), so there are no large differences in the execution of the drawing task between species.

To obtain a more nuanced understanding of the analysts' behaviour, the drawings of correct (**Error! Reference source not found.**B) and incorrect (**Error! Reference source not found.**C) answers are compared. As would be expected, when the classifications are correct, the analysts' drawings and the descriptive criteria are in synchrony – the features drawn on sheep and goat astragali correspond to the descriptive criteria associated with the respective species. However, when the analysts are incorrect in their classifications, the drawn areas of goat specimens now correspond to sheep-like qualities in descriptive criteria and the sheep drawings correspond to goat-like qualities. Thus, the analysts make their decisions by following descriptive criteria very closely, but as the criteria do not fully encapsulate the variances of the morphologies of the two species, the analysts make mistakes. This pattern is easiest to discern in the drawings for the dorsal and medial views of the correct and incorrect answers. In the dorsal view, the medial articular ridge is described as being generally oblique for goats and horizontal

for sheep, but the opposite is true for incorrect classifications. Concerning the medial view, the proximo-plantar projection is usually described as being more rounded for sheep and pointed for goats, but again, the opposite is true for the drawings of the incorrect answers.

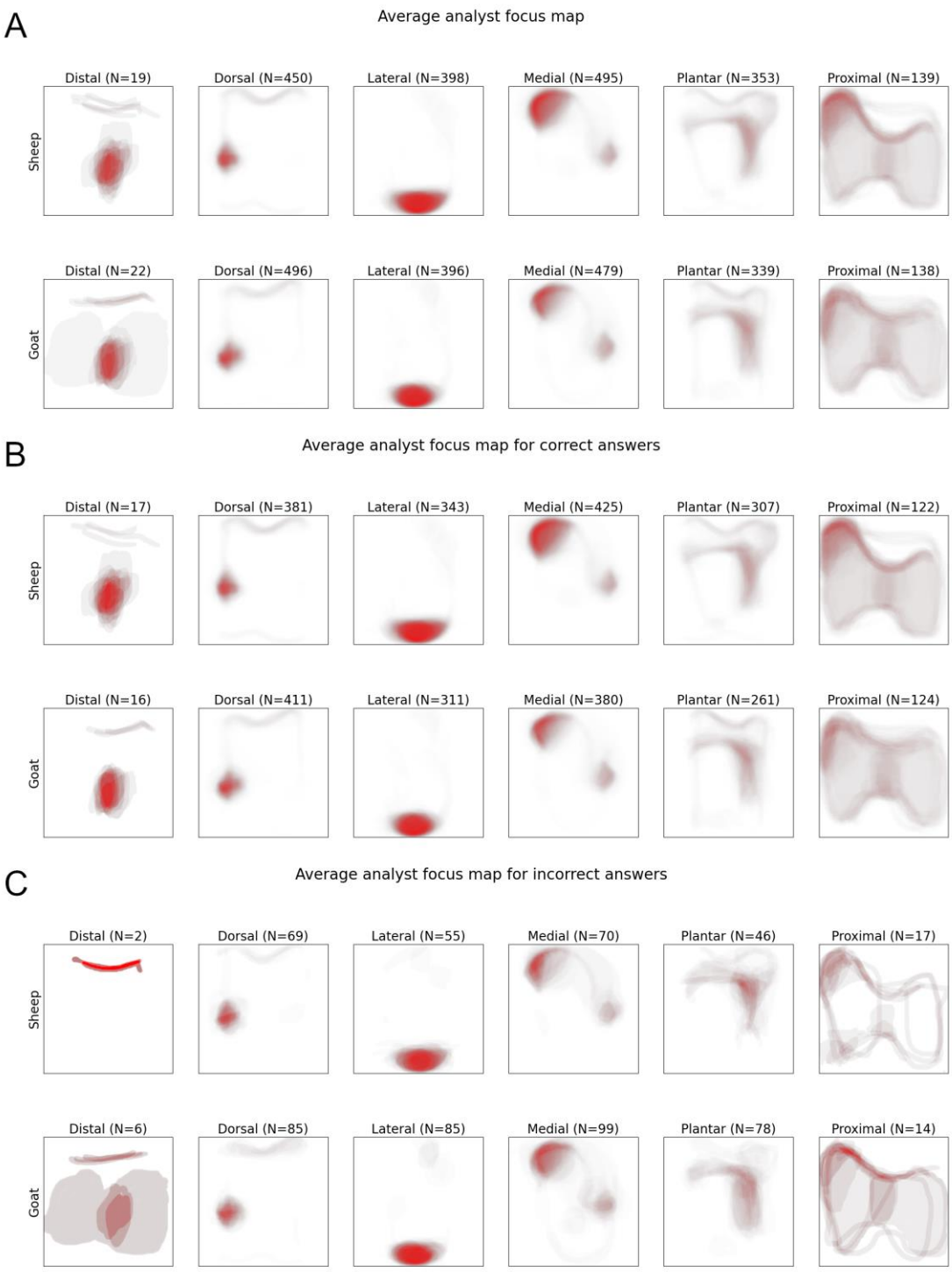

Fig. 15 *Average focus maps by view and species. Each view is averaged by the number of times (N) it was drawn by analysts. The true label is on the left. A) Average analyst focus map for the specimens. These classifications include correct and incorrect answers for both species. B) Average analyst focus maps for the correctly classified specimens. C) Average analyst focus maps for the incorrectly classified specimens*

The easiest and the hardest specimens

To underline this overall pattern, we turn to the easiest and the hardest specimens as examples of cases when the morphological shape of the bone does not fit within the variation covered by descriptive criteria. The easiest specimens are chosen on the basis of which bones had the highest Difficulty Index and the most classification attempts,

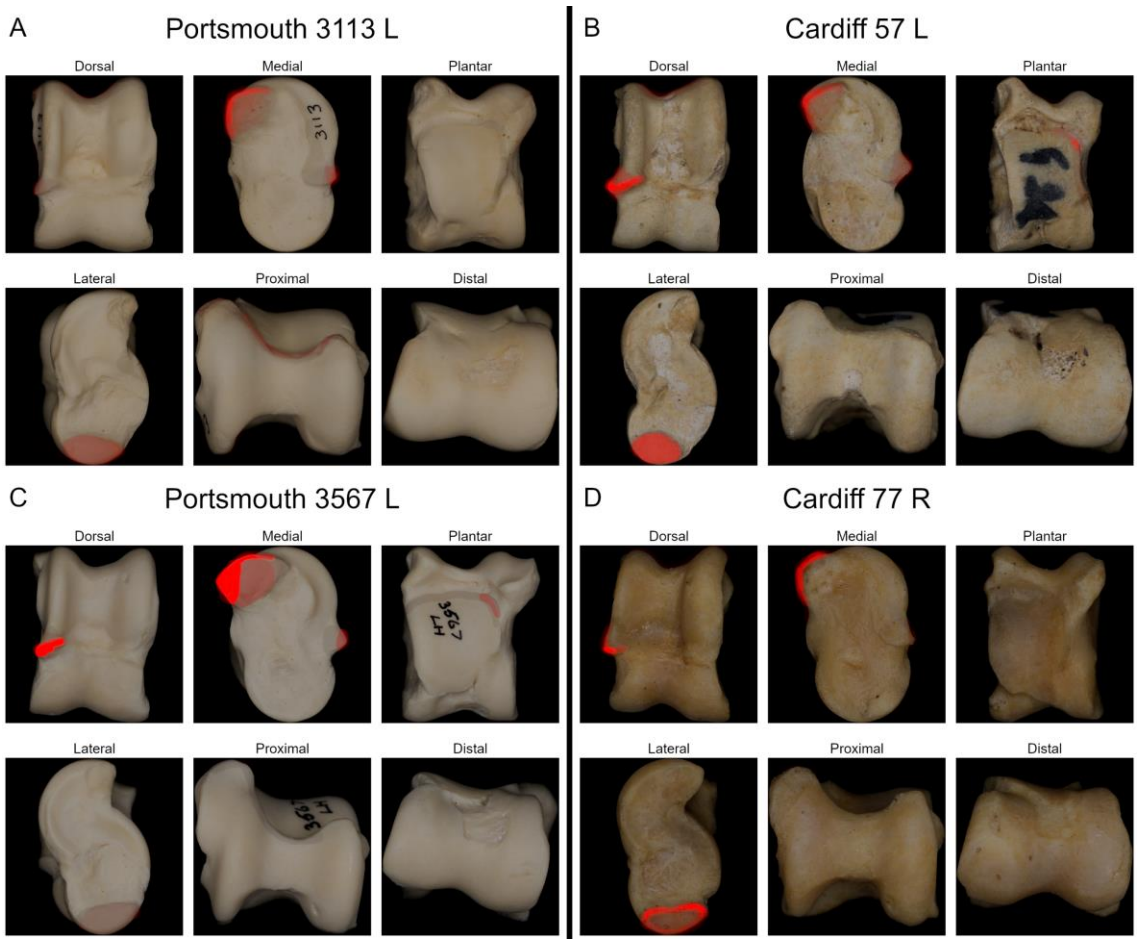

Fig. 16 *The easiest (top row) and hardest (bottom row) sheep and goat specimens. The sheep are on the left and goats on the right side. R in the specimen name refers to right side and L to the left side of the animal. A) The easiest sheep astragalus - Portsmouth 3113 L. B) The easiest goat astragalus - Cardiff 57 L. C) The hardest sheep astragalus - Portsmouth 3567 L. D) The hardest goat astragalus - Cardiff 77 R. Note the curved nature of the bone in dorsal view. This curvature may cause the medial articular ridge to appear more horizontal in dorsal view. Best viewed in colour, available online*

whereas the hardest specimens are chosen based on the lowest Difficulty Index and highest number of classification attempts. The easiest sheep is Portsmouth 3113 L (correctly classified in all 14 classifications, **Fig. 16**A) and the easiest goat is Cardiff 57 L (correctly classified in all 15 attempts, **Fig. 16**B). The hardest sheep is Portsmouth 3567 L (incorrectly classified in all three attempts, **Fig. 16**C) and the hardest goat specimen is Cardiff 77 R (incorrectly classified in all seven cases, **Fig. 16**D). L and R indicate the side of the animal the specimen comes from. The red areas in these images correspond to all drawings made by the analysts.

To begin with, the proximo-plantar projection and medial articular ridge in medial view are very similar for the easiest sheep (**Fig. 16**A) and the hardest goat specimens (**Fig. 16**D), showing the overlapping morphological variances of the two species. Furthermore, the medial articular ridge for the hardest goat specimen, Cardiff 77 R, is nearly horizontal with respect to the longest axis of the bone in dorsal view, which would

normally be associated with sheep (Boessneck 1969; Zeder and Lapham 2010). The angle of the medial articular ridge for Cardiff 77 R specimen is even more horizontal than for the easiest sheep, Portsmouth 3113 L. It is therefore quite understandable that this bone was problematic for analysts if they followed the instructions in descriptive criteria. However, Cardiff 77 R has morphological qualities that may make it vary from the expected. For instance, Cardiff 77 R has a general curvature that can be seen in the dorsal view, which may result in the medial articular ridge becoming more horizontal as well as reduce its prominence in medial view. In the plantar view of this specimen, the plantar articular surface is also distinctly shifted towards the lateral side, which may be a further symptom of the bending of the astragalus. This particular specimen can therefore be likened to a spring that is pressed unevenly such that as one side is compressed, the opposite side bulges outwards. Even though the morphology of Cardiff 77 R does not conform to the expected morphology, it is not possible to say definitively that this bone is abnormal because we simply do not know the population variance. In other words, although Cardiff 77 R is just one sample, bones with morphologies beyond the descriptive criteria and other comparative methods may be more common than currently thought given the vastness of the archaeological record. Current comparative methods (especially those depending on written and/or drawn descriptions) therefore underestimate the importance of morphological variation as the variation is reduced to single points in a multidimensional space.

The hardest sheep and the easiest goat specimens, too, have shared morphological qualities. The hardest sheep – Portsmouth 3567 L (**Fig. 16**C) – has an oblique medial articular ridge in the dorsal view that is quite prominent in the medial view and the proximo-plantar projection in medial view is very pointed, all of which are features normally associated with goat-like morphologies (Boessneck 1969; Zeder and Lapham 2010) and are found in Cardiff 57 L (**Fig. 16**B). Yet, the distal articular surface in the lateral view of Portsmouth 3567 L is not tear-drop shaped (a feature typically considered goat-like) as it runs across the entire lateral face of the bone, which is as expected from a sheep (Zeder and Lapham 2010).

To summarise, the medial articular ridge and the pointiness of the proximo-plantar projection are the most likely sources of much of the confusion for human analysts due to the analysts' dependence of these features in both dorsal and medial views. The

medial articular ridge obliqueness in dorsal view is not very different between the hard and the easy specimens (apart from Cardiff 77 R, for which we gave a possible explanation), whereas its prominence in medial view is very similar between the easy goat and the hard sheep as well as between the easy sheep and the hard goat. Likewise, the proximo-plantar projection is pointed for hard sheep and easy goat, but less so for hard goat and easy sheep. These examples therefore underline the argument that morphological descriptions do not apply to individuals whose morphologies vary from the morphology of those samples that were used in defining the original descriptive criteria. Thus, even though these features are good individual features to separate a large proportion of sheep and goat astragali, the overlapping morphological variance is such that zooarchaeology as a discipline could benefit from a holistic approach to classifying bones to species. In other words, instead of typifying bone morphologies to simple rules and specific features, the whole bone morphology should be considered for more reliable classifications in the future, especially as the features of a given bone are dependent on all other features of that same bone as well as the articulating bones. Bones should therefore be thought of as continuous geometries, not as a series of discrete features. Of course, the observations made here on a limited number of bones require further quantitative testing, but this is beyond the scope of this article.

Discussion

The results of the blind study are complex, but we have achieved our three main aims. First, we have tested the impact of different types of reference materials in the classification of sheep and goat astragali. Second, we have analysed the participants' spatial attention. Third, we have shown that self-reported confidence scores can be used in place of the ambiguous sheep/goat category. The additional benefit of our study is that it can be used as a benchmark for an archaeologically relevant image dataset for forthcoming deep learning models.

However, we acknowledge that the results may be affected by the method of acquiring the participant data since zooarchaeologists do not normally classify bones only from photographs and instead zooarchaeologists aim to inspect the bones carefully in person. The images used also did not incorporate relevant size information to help the analysts to differentiate the species, which may affect the results. The removal of size information was necessary because it is not usually part of the described morphological

criteria, although osteometric analyses demonstrate size differences between sheep and goat astragali (Davis 2017; Haruda 2017; Salvagno and Albarella 2017).

Although it has previously been found that there is little agreement in the identifications of pottery and lithic artefacts when they are based on digital photographs and when the identifications are performed in a laboratory (Heilen and Altschul 2013), Heilen and Altschul's (2013) study failed to account for the fact that two different analysts were involved in the classification of the materials, so their observed discrepancy may simply reflect the two analysts' capabilities. We therefore argue that even though the analysts' performances could be better if performed in person, there is no firm evidence that using digital images is the cause of lower classification accuracies. In fact, only one participant (Analyst 104) raised any concerns about the photographs, saying that couple of the photographs were "blurred at the proximal end", "truncated", and "speckled or very light/white and it was not clear where the surfaces were." Analyst 104 also mentioned that lighting and angle could be adjusted when analysing bones in person. Finally, we would further counter the argument that using images rather than physical specimens somehow impacted the results by arguing that identifications of bones from images are not infrequent, and they typically occur when an analyst is faced with a difficult morphology and they want to consult with colleagues elsewhere in the world.

Analyst performance

The blind study demonstrates that the consistency of human experts is good, but their overall accuracy (81.15%) is lower than the accuracy reported (89.36%-94.74%, see **Table 1** and **Table 2**) by Zeder and Lapham (2010). This difference may be explained by the differing methodologies and the number of participants. Furthermore, it was observed that human error rates for sheep (15.26%) is lower than for goat astragali (22.44%) and that the analysts were also more confident about classifying sheep astragali. The noted differences regarding accuracy and confidence are reflected in how difficult the bones of the two species were for the analysts overall, with sheep bones generally being easier to identify. This observed difference between species means that using comparative methods of identifying archaeological remains of sheep and goat astragali are likely to lead to a bias towards sheep astragalus identifications. Our study therefore raises the question of how large of an impact this bias has for sheep to goat

ratios. We discuss the impact of confidence and especially in relation to sheep:goat ratio in more detail in a separate section below.

Group-level differences

Although statistical tests were unable to find clear differences between the analyst groups apart from sheep astragalus accuracy between Group 2 and Group 4, we consider these expertise groups to be valid and display different levels of abilities based on the performance boxplots in Fig. **3** and consistency plot in Fig. **4**. However, the expertise groups themselves were not found to match expectations; prior to the analysis of the group performances, we set the expectation that Group 1 analysts (the professionals) would outperform other groups due to zooarchaeological identification tasks being part of their daily routine, followed by Group 4 (doctorates), Group 3 (postgraduates), and finally Group 2 (novices). However, our study shows that Group 4 analysts were the most accurate and consistent group even though they tended to have worked on fewer number of assemblages and spent less time on identification tasks per week than Group 1 participants. Moreover, Group 4 outperformed other groups despite their lack of reliance on reference materials. Taking all of the evidence regarding Group 4 (i.e. high accuracy, high consistency, high confidence, infrequent use of reference materials, PhD level education, and relatively few hours spent in identification tasks) into consideration, we can be confident in stating that they were the true experts. Thus, the group performance results are at odds with our expected order.

One possible explanation for the discrepancy between our expectations and the results may be that our study design does not truly capture expertise in sheep and goat astragalus differentiation. We do not think this is the case, since the deviation of Group 1 performance from the expectation (i.e. observed rank 3 vs expected rank 1) is enough to explain the overall deviation from the expected order of the group performances. We can also clearly see that novices (Group 2) are worse than postgraduates (Group 3) who are worse than doctorates (Group 4), demonstrating a progression in performances as the analysts become more highly qualified and better acquainted with sheep and goat separation. Note, however, that the highest level of qualification is not the defining feature of the expertise groupings and other questions also play a role in the creation of the groups. If the study design had been flawed, it is more likely that the rank order of the analyst groups would be closer to random and we would not have been able to place

the analysts into such clearly defined groups as the analysts would have been assigned to groups randomly.

Instead, we argue that the hours worked in zooarchaeological identification tasks and how many assemblages the analysts had worked on in the last five years do not have a large impact on analysts' ability to separate sheep and goat astragali. We argue this to be the case on the basis that Group 1 analysts were not more accurate and only somewhat more consistent than analysts in Group 2 (relative novices) and those in Group 3 (postgraduates). This result further conforms to the more general assessment of the impact of experience on performance by Ericsson and Lehmann (1996), who state that increased amount of knowledge gained through experience in a domain does not always lead to superior performance compared to the less-experienced individuals. In other words, sheep and goat astragalus classification is not necessarily a task where analyst expertise can be measured via time spent in zooarchaeological tasks or the variety of tasks. Instead, it is argued that latent ability to identify shapes, continuous self-improvement, or even a teaching role are more likely to lead to better performance, but this cannot be deduced from the collected data and requires further research.

Inference on reference materials

Surprisingly, we did not find any evidence of reference materials being helpful in the classification task. In the GLMM analysis, using Boessneck (1969; Boessneck et al. 1964) as a reference text was the only type of reference material that was found to have a positive coefficient, while all the other reference materials had a negative coefficient. This result means that not using any type of reference material results in a higher log odds of correctly identifying the species than when relying on reference materials, apart from when using Boessneck (1969; Boessneck et al. 1964) and even then the difference is not statistically significant. The only important factor in having a higher log odds of a correct answer is membership in Group 4.

This result may seem flawed to many, since most of us believe that having reference materials are helpful (e.g. Bochenski 2008) – this was certainly the authors' belief as well. However, we argue that reference materials are likely to be most useful when a bone morphology is encountered for the first time or when the morphology is very uncommon, as the analyst has not yet formed a mental frame of reference for that morphology. As one builds this mental frame of reference, the usefulness of reference

materials wains. This is evident in the fact that most zooarchaeologists are capable of identifying a large variety of bones accurately without relying on any reference materials, but they are quick to ask for help when encountering an unusual specimen. Likewise, reference materials are often used when trying to differentiate between two very similar morphologies, but as noted, our results do not support such a usage. As Group 4 analysts were much less likely to rely on reference materials, they have the highest accuracy, they are the most consistent, and they were the most confident, we put forward the argument that these analysts have built a mental frame of reference for the morphological differences between sheep and goat astragali and we again reiterate our argument that they are likely to be the true experts at this task. This result cannot be explained by the selection of bones that Group 4 analysts analysed, since the distribution of specimen difficulties were not found to differ between groups.

Additionally, the result that reference materials were not found to increase the log odds of a correct answers could reflect the fact that the population variances are not captured by single reference items (be it a 3D model, physical specimen, image, or sketch). While publications and manuals can draw conclusions from a large number of specimens, they tend to simplify the variation to single points. Physical reference collections are similarly limited as they often only include a handful of specimens for each species and typically contain samples from a limited subset of species. The usefulness of reference materials is therefore likely limited by their incapability to contain variance. Even descriptive criteria may not encapsulate the entire variation as they tend to be gross generalizations that are based on specimens that can derive from many geographic regions or worse, from very specific populations. Indeed, it has been argued that geographic and temporal variation affect the morphology of at least sheep astragali (Haruda et al. 2019; Pöllath et al. 2019). Thus, when an analyst is exceedingly reliant on reference materials, their accuracy may be lowered if they are unable to match the underlying population variances of the reference sample and the test sample, particularly when only parts of the bone morphology are used such as when the study specimen is fragmented. In a forthcoming article, we argue that the mismatch between the population variances and the variance encapsulated in reference materials can be reduced by using deep learning convolutional neural networks as they can take advantage of the entire bone morphology in a classification task, whereas comparative identification is reliant on defined, discrete features.

Impact of confidence

As expected, we found that confidence correlates with one's ability to classify sheep and goat astragali and that human analyst confidence correlates with the difficulty of the specimen, which is again as expected. It was also found that by setting a threshold to confidence scores it becomes possible to limit the evaluation of an assemblage to only those specimens that the analysts are the most confident about, which increased the average accuracy. This observation is especially true for analysts in Groups 3 and 4, whose accuracies improved from 82.82% and 87.27% for all answers to 95.24% and 94.74% for answers in the 'Very high' confidence category (using a threshold value of 75), respectively. When the confidence threshold is set to 85 for all groups, the mean accuracy is 93.33%, whereas the mean accuracy for all answers is 81.15%.

As analysts are more confident about classifying sheep than goats, limiting the answers to the most confident responses has the downside that the sheep to goat ratio shifts in favour of sheep, regardless of whether one analyses the true ratio, the ratio for correct answers, or the ratio for all answers. We have demonstrated this effect in detail in Online resource 10. Importantly, as low confidence answers would be likely classified as sheep/goat in zooarchaeological analyses, it is probable that this same effect is present in many published studies and sheep are therefore more than likely overrepresented in sheep:goat ratios that do not take ambiguous sheep/goat classifications into account.

Experience and confidence

In Driver's (1992) view, more experienced analysts are less willing to differentiate between morphologically similar species and therefore are presumably more likely to place bones in the sheep/goat category. However, our study found that analysts in Groups 1 and 4 (the more experienced analysts) have higher confidence scores than analysts in Groups 2 and 3. Because there is no prior information about what confidence level analysts used when they place samples in the 'sheep/goat' class, analysts could find the adoption of self-reported confidence scores beneficial in that it allows the communication of the rejection criteria. For example, the analyst can simply state that they included specimens that they were 75% confident to be correct and excluded those specimens with lower confidence from further analysis. Using self-reported confidence scores therefore informs peers about the analyst's certainty for each specimen, which in turn allows the reviewer to identify potential blind spots in the analysis. The analysts

themselves could use this data for self-reflection to find out areas of zooarchaeological identification that they need to improve upon should they keep track of their confidence scores over a long period. Furthermore, one additional avenue for taking advantage of confidence values could be to use them as prior probabilities following Wolfhagen and Price's (2017) methodology. However, unlike Wolfhagen and Price (2017, p.626), who mention that prior probabilities "express beliefs about the proportion of goats or sheep in an assemblage", confidence scores could be used to express beliefs about the specimen being a goat or a sheep. The difference in these definitions is larger than it seems: instead of using one prior for the entire study sample, confidence scores allow a unique prior for all bones.

In the future, similar studies that incorporate fragmented bones should be conducted because fragmentation has an impact on analyst accuracy (Pickering et al. 2006), which is likely to also have an effect on confidence. Furthermore, as it was found that analysts are more confident and better at classifying sheep astragali, it may be that there is a wider issue of reference materials being biased towards one species and it would be beneficial to look further into this issue. Overreliance on reference materials that are biased or depict limited variance would lead to systematic errors, which may erroneously lead to inflated or deflated confidence, depending on the population prevalence of the features used for classifying the bones. These errors would then be compounded in regional studies or meta analyses.

Specimen difficulty

Perhaps unexpectedly, majority of the astragali were easy for the analysts, while a small number of them were very difficult. The qualitative analysis of the analysts' attention demonstrates how the analysts follow the descriptive criteria very faithfully. It is put forward here that the morphological variation of the difficult bones are at the fringes or beyond the morphological variation expressed by reference materials or indeed the analysts' own mental frame of reference. For example, the obliqueness of the medial articular ridge and the pointiness of the proximo-plantar projection are very similar for difficult goat and easy sheep specimens and vice versa, which obviously leads to a false classification as these features are primary features used in differentiating the two species (Boessneck 1969; Zeder and Lapham 2010). Thus, it is important to involve morphologically more varied references, but doing so is not easy nor always possible

due to the availability of specimens or space, let alone the fact that the analyst would have to make several comparisons for each possible species. Instead, computational models (such as deep learning convolutional neural networks) could be used in the future.

Conclusion

Our study has shown that human analysts can be very accurate at classifying sheep and goat astragali from photographs, but on average, their classification error rate is around one in five specimens. We did not find any evidence that using reference materials leads to a higher probability of a correct answer, and we argue that the reason is that morphological variation is not captured well by any single reference material type. Furthermore, analysts were found to be more confident and accurate when classifying sheep astragali, suggesting that either the analysts or the reference materials may be biased towards the identification of sheep. It is likely that this bias is also present in published sheep to goat ratios. This research has further shown that self-reported confidence scores can be used in place of the ambiguous sheep/goat class, which has several benefits to the researchers. For example, confidence scores can inform peers about potential blind spots in the analysis and analysts can use self-reported confidence scores to identify their own strengths and weaknesses if they keep track of the confidence scores.

Finally, analysts' focus maps demonstrate that human experts are somewhat inflexible about which features they use to classify the specimens. This behaviour leads to the misclassification of specimens that do not have very specific features that fall within the morphological variance insinuated by reference materials. As written and drawn comparative materials often have discrete boundaries and zooarchaeologists are trained to follow these guides, any bone that varies from the expected rigid shape has a higher probability of being misidentified. In some cases it may be enough that one feature of a bone is deemed to have a morphology of another species that all other features are disregarded. Therefore, we conclude that more flexible classification methods are needed for increased classification accuracy. Such a method should take a holistic approach to the species morphology and be able to encapsulate the full variance of the bones. We suggest that deep learning convolutional neural networks are able to achieve this and their usefulness will be shown in a forthcoming paper. Deep learning

convolutional neural network models also have the benefit of being portable as they only occupy digital space and can be utilised over the internet.

Author Declarations

Funding

Kone Foundation (Koneen Säätiö, funding identifiers: 201710513, 202006857 & 202102207) funded I.S.'s PhD studies, including the writing of this article.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics declaration

Each participant was asked to consent to the test as required to fulfil the ethics requirements set by UCL Institute of Archaeology Ethics Committee (ethics approval granted on 15 June 2020, Reference number 2020.020). No personal information was collected at any point, none of the participants were identifiable by the authors at any point during the study, nor are the participants identifiable from the data made available. All participants had the opportunity to remove their answers at any point during the test and afterwards.

Consent to participate

Not applicable

Consent for publication

Not applicable

Data availability

The data is made available in the Supplementary Information.

Code availability

Not applicable

Authors' contributions

I.S. wrote the main manuscript text, created the study website, performed the analyses, and created the figures and tables; L.D. edited the text and created Online Resource 1; L.M. edited the text; All authors were involved in the study design and reviewed the manuscript. J.S., L.M., and L.D. supervised I.S.'s PhD thesis.

References

Abdi H, Williams LJ (2010) Principal Component Analysis. Wiley Interdiscip Rev Comput Stat 2:1–47

Agresti A (2002) Categorical Data Analysis, 2nd edn. John Wiley & Sons, Inc, Hoboken, New Jersey

Atici L, Whitcher Kansa S, Lev-Tov J, Kansa EC (2013) Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. J Archaeol Method Theory 20:663–681. https://doi.org/10.1007/s10816-012-9132-9

Austen GE, Bindemann M, Griffiths RA, Roberts DL (2018) Species identification by conservation practitioners using online images: accuracy and agreement between experts. PeerJ 6:1–19. https://doi.org/10.7717/peerj.4157

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang 68:255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67:. https://doi.org/10.18637/jss.v067.i01

Bewick V, Cheek L, Ball J (2004) Statistics review 8: Qualitative data – tests of association. Crit Care 8:46–53. https://doi.org/10.1186/cc2428

Blumenschine RJ, Marean CW, Capaldo SD (1996) Blind Tests of Inter-analyst Correspondence and Accuracy in the Identification of Cut Marks, Percussion Marks, and Carnivore Tooth Marks on Bone Surfaces. J Archaeol Sci 23:493–507. https://doi.org/10.1006/jasc.1996.0047

Bochenski ZM (2008) Identification of skeletal remains of closely related species: the pitfalls and solutions. J Archaeol Sci 35:1247–1250. https://doi.org/10.1016/j.jas.2007.08.013

Boessneck J (1969) Osteological Differences between Sheep (Ovis aries Linné) and Goat (Capra hircus Linné). In: Brothwell D, Higgs E (eds) Science in Archaeology. A Comprehensive Survey of Progress and Research, 2nd edn. Thames and Hudson, London, pp 331–358

Boessneck J, Miller H-H, Teichert M (1964) Osteologische Unterscheidungsmerkmale zwischen Schaf (*Ovis aries* Linné) une Ziege (*Capra hircus* Linné). Kühn-Archiv 78:1–129

Bolker BM, Brooks ME, Clark CJ, et al (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24:127–135. https://doi.org/10.1016/j.tree.2008.10.008

Buckley M, Whitcher Kansa S, Howard S, et al (2010) Distinguishing between archaeological sheep and goat bones using a single collagen peptide. J Archaeol Sci 37:13–20. https://doi.org/10.1016/j.jas.2009.08.020

Burgman MA, McBride M, Ashton R, et al (2011) Expert Status and Performance. PLoS One 6:7. https://doi.org/10.1371/journal.pone.0022998

Culley C, Janzen A, Brown S, et al (2021) Iron Age hunting and herding in coastal eastern Africa: ZooMS identification of domesticates and wild bovids at Panga ya Saidi, Kenya. J Archaeol Sci 130:13. https://doi.org/10.1016/j.jas.2021.105368

Davis SJM (2017) Towards a metrical distinction between sheep and goat astragali. In: Rowley-Conwy P, Serjeantson D, Halstead P (eds) Economic Zooarchaeology. Studies in Hunting, Herding and Early Agriculture. Oxbow Books, Oxford, pp 50–82

Davis SJM (2016) Hacia una distinción métrica entre los astrágalos de oveja y cabra. In: Lloveras L, Rissech C, Nadal J, Fullola JM (eds) What bones tell us. El que ens expliquen els ossos, Monografies del SERP 12. Universitat de Barcelona, Barcelona, pp 35–58

Dinno A (2015) Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. Stata J 15:292–300. https://doi.org/10.1177/1536867x1501500117

Domínguez-Rodrigo M (2012) Critical review of the MNI (minimum number of individuals) as a zooarchaeological unit of quantification. Archaeol Anthropol Sci 4:47–59. https://doi.org/10.1007/s12520-011-0082-z

Driver JC (1992) Identification, classification and zooarchaeology. Circaea 9:35–47

Dunn OJ (1964) Multiple Comparisons Using Rank Sums. Technometrics 6:241–252

Endsley MR, Kiris EO (1995) The Out-of-the-Loop Performance Problem and Level of Control in Automation. Hum Factors 37:381–394. https://doi.org/10.1518/001872095779064555

Ericsson KA, Lehmann AC (1996) Expert and Exceptional Performance: Evidence of Maximal Adaptation to Task Constraints. Annu Rev Psychol 47:273–305. https://doi.org/10.1146/annurev.psych.47.1.273

Fernandez H (2001) Ostéologie comparée des petits ruminants eurasiatiques sauvages et domestiques (genres *Rupicapra*, *Ovis*, *Capra* et *Capreolus*): diagnose différentielle du squelette appendiculaire. Université de Genève

Gilmour GH (1997) The nature and function of astragalus bones from archaeological contexts in the Levant and eastern Mediterranean. Oxford J Archaeol 16:167–175. https://doi.org/10.1111/1468-0092.00032

Giovas CM, Lambrides ABJ, Fitzpatrick SM, Kataoka O (2017) Reconstructing prehistoric fishing zones in Palau, Micronesia using fish remains: A blind test of inter-analyst correspondence. Archaeol Ocean 52:45–61. https://doi.org/10.1002/arco.5119

Gobalet KW (2001) A Critique of Faunal Analysis; Inconsistency among Experts in Blind Tests. J Archaeol Sci 28:377–386. https://doi.org/10.1006/jasc.2000.0564

Greenlee DM, Dunnell RC (2010) Identification of fragmentary bone from the Pacific. J Archaeol Sci 37:957–970. https://doi.org/10.1016/j.jas.2009.11.029

Halstead P, Collins P, Isaakidou V (2002) Sorting the Sheep from the Goats: Morphological Distinctions between the Mandibles and Mandibular Teeth of Adult *Ovis* and *Capra*. J Archaeol Sci 29:545–553. https://doi.org/10.1006/jasc.2001.0777

Haruda AF (2017) Separating Sheep (*Ovis aries* L.) and Goats (*Capra hircus* L.) Using Geometric Morphometric Methods: An Investigation of Astragalus Morphology

from Late and Final Bronze Age Central Asian Contexts. Int J Osteoarchaeol 27:551–562. https://doi.org/10.1002/oa.2576

Haruda AF, Varfolomeev V, Goriachev A, et al (2019) A new zooarchaeological application for geometric morphometric methods: Distinguishing *Ovis aries* morphotypes to address connectivity and mobility of prehistoric Central Asian pastoralists. J Archaeol Sci 107:50–57. https://doi.org/10.1016/j.jas.2019.05.002

Heilen M, Altschul JH (2013) The Accuracy and Adequacy of In-Field Artifact Analysis. An Experimental Test at Two Archaeological Sites in the Western United States. Adv Archaeol Pract 1:121–138. https://doi.org/10.7183/2326-3768.1.2.121

Holmgren R (2004) “Money on the hoof” The astragalus bone – religion, gaming and primitive money. In: Frizell BS (ed) Pecus. Man and Animal in Antiquity: Proceedings of the conference at the Swedish Institute in Rome, September 9-12, 2002. The Swedish Institute in Rome, Rome, pp 212–220

Kain MP, Bolker BM, McCoy MW (2015) A practical guide and power analysis for GLMMs: detecting among treatment variation in random effects. PeerJ 3:24

Koerper HC, Whitney-Desautels NA (1999) Astragalus Bones: Artifacts Or Ecofacts? Pacific Coast Archaeol Soc Q 35:69–80

Kruskal WH, Wallis WA (1952) Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc 47:583–621. https://doi.org/10.1080/01621459.1952.10483441

Lau H, Whitcher Kansa S (2018) Zooarchaeology in the era of big data: Contending with interanalyst variation and best practices for contextualizing data for informed reuse. J Archaeol Sci 95:33–39. https://doi.org/10.1016/j.jas.2018.03.011

Lee Y, Nelder JA (1996) Hierarchical Generalized Linear Models. J R Stat Soc Ser B 58:619–656. https://doi.org/10.1111/j.2517-6161.1996.tb02105.x

Lee Y, Nelder JA (2006) Double hierarchical generalized linear models. J R Stat Soc Ser C Appl Stat 55:139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x

Lloveras L, Moreno-García M, Nadal J, Thomas R (2014) Blind test evaluation of accuracy in the identification and quantification of digestion corrosion damage on leporid bones. Quat Int 330:150–155. https://doi.org/10.1016/j.quaint.2013.07.033

Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. Eur J Soc Psychol 51:485–504. https://doi.org/10.1002/ejsp.2752

Lyman RL, VanPool TL (2009) Metric Data in Archaeology: A Study of Intra-Analyst and Inter-Analyst Variation. Am Antiq 74:485–504

Morin E, Ready E, Boileau A, et al (2017) Problems of Identification and Quantification in Archaeozoological Analysis, Part I: Insights from a Blind Test. J Archaeol Method Theory 24:886–937. https://doi.org/10.1007/s10816-016-9300-4

Moscatelli A, Mezzetti M, Lacquaniti F (2012) Modeling psychophysical data at the population-level: The generalized linear mixed model. J Vis 12:1–17. https://doi.org/10.1167/12.11.26

Nims R, Butler VL (2017) Assessing reproducibility in faunal analysis using blind tests: A case study from northwestern North America. J Archaeol Sci Reports 11:750–761. https://doi.org/10.1016/j.jasrep.2017.01.012

Noddle B (1974) Ages of Epiphyseal Closure in Feral and Domestic Goats and Ages of Dental Eruption. J Archaeol Sci 1:195–204. https://doi.org/10.1016/0305-4403(74)90042-9

Pedregosa F, Varoquax G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

Pickering TR, Egeland CP, Schnell AG, et al (2006) Success in Identification of Experimentally Fragmented Limb Bone Shafts: Implications for Estimates of Skeletal Element Abundance in Archaeofaunas. J Taphon 4:97–108

Pilaar Birch SE, Scheu A, Buckley M, Çakırlar C (2019) Combined osteomorphological, isotopic, aDNA, and ZooMS analyses of sheep and goat remains from Neolithic Ulucak, Turkey. Archaeol Anthropol Sci 11:1669–1681. https://doi.org/10.1007/s12520-018-0624-8

Pöllath N, Alibert P, Schafberg R, Peters J (2018) Striking new paths – Distinguishing ancient *Ovis orientalis* from its modern domestic descendant (Karakul breed) applying Geometric and traditional Morphometric approaches to the astragalus. In: Çakırlar C, Chahoud J, Berthon R, Pilaar Birch S (eds) Archaeozoology of the Near East XII. Barkhuis Publishing & University of Groningen, Groningen, pp 207–

Pöllath N, Schafberg R, Peters J (2019) Astragalar morphology: Approaching the cultural trajectories of wild and domestic sheep applying Geometric Morphometrics. J Archaeol Sci Reports 23:810–821. https://doi.org/10.1016/j.jasrep.2018.12.004

Prendergast ME, Janzen A, Buckley M, Grillo KM (2019) Sorting the sheep from the goats in the Pastoral Neolithic: morphological and biomolecular approaches at Luxmanda, Tanzania. Archaeol Anthropol Sci 11:3047–3062. https://doi.org/10.1007/s12520-018-0737-0

Prummel W, Frisch H-J (1986) A Guide for the Distinction of Species, Sex and Body Side in Bones of Sheep and Goat. J Archaeol Sci 13:567–577. https://doi.org/10.1016/0305-4403(86)90041-5

Python Software Foundation (2016) Python Language Reference, version 3.6.0

R Core Team (2021) R: A Language and Environment for Statistical Computing

Salvagno L (2020) The Neglected Goat. A new method to assess the role of the goat in the English Middle Ages. Archaeopress Publishing Ltd, Oxford

Salvagno L, Albarella U (2017) A morphometric system to distinguish sheep and goat postcranial bones. PLoS One 12:e0178543

Stroup WW (2013) Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press, Boca Raton, Florida

Thompson CG, Kim RS, Aloe AM, Becker BJ (2017) Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. Basic Appl Soc Psych 39:81–90. https://doi.org/10.1080/01973533.2016.1277529

Tillett BJ, Field IC, Bradshaw CJA, et al (2012) Accuracy of species identification by fisheries observers in a north Australian shark fishery. Fish Res 127–128:109–115. https://doi.org/10.1016/j.fishres.2012.04.007

Twiss KC, Wolfhagen J, Madgwick R, et al (2017) Horses, Hemiones, Hydruntines? Assessing the Reliability of Dental Criteria for Assigning Species to Southwest Asian Equid Remains. Int J Osteoarchaeol 27:298–304. https://doi.org/10.1002/oa.2524

Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Welker F, Soressi M, Rendu W, et al (2015) Using ZooMS to identify fragmentary bone from the Late Middle/Early Upper Palaeolithic sequence of Les Cottés, France. J Archaeol Sci 54:279–286. https://doi.org/10.1016/j.jas.2014.12.010

Wolfhagen J, Price MD (2017) A probabilistic model for distinguishing between sheep and goat postcranial remains. J Archaeol Sci Reports 12:625–631. https://doi.org/10.1016/j.jasrep.2017.02.022

Wolverton S (2013) Data Quality in Zooarchaeological Faunal Identification. J Archaeol Method Theory 20:381–396

Zeder MA (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci U S A 105:11597–11604. https://doi.org/10.1073/pnas.0801317105

Zeder MA, Lapham HA (2010) Assessing the reliability of criteria used to identify postcranial bones in sheep, *Ovis*, and goats, *Capra*. J Archaeol Sci 37:2887–2905. https://doi.org/10.1016/j.jas.2010.06.032

Zeder MA, Pilaar SE (2010) Assessing the reliability of criteria used to identify mandibles and mandibular teeth in sheep, *Ovis*, and goats, *Capra*. J Archaeol Sci 37:225–242. https://doi.org/10.1016/j.jas.2009.10.002

Zhang Q, Couloigner I (2005) A New and Efficient K-Medoid Algorithm for Spatial Clustering. In: Gervasi O, Gavrilova ML, Kumar V, et al. (eds) Computational Science and Its Applications – ICCSA 2005. ICCSA 2005. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 181–189

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14. https://doi.org/10.1111/j.2041-210x.2009.00001.x

Abdi H, Williams LJ (2010) Principal Component Analysis. Wiley Interdiscip Rev Comput Stat 2:1–47

Agresti A (2002) Categorical Data Analysis, 2nd edn. John Wiley & Sons, Inc, Hoboken, New Jersey

Atici L, Whitcher Kansa S, Lev-Tov J, Kansa EC (2013) Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. J Archaeol Method Theory 20:663–681. https://doi.org/10.1007/s10816-012-9132-9

Austen GE, Bindemann M, Griffiths RA, Roberts DL (2018) Species identification by conservation practitioners using online images: accuracy and agreement between experts. PeerJ 6:1–19. https://doi.org/10.7717/peerj.4157

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang 68:255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67:. https://doi.org/10.18637/jss.v067.i01

Bewick V, Cheek L, Ball J (2004) Statistics review 8: Qualitative data – tests of association. Crit Care 8:46–53. https://doi.org/10.1186/cc2428

Blumenschine RJ, Marean CW, Capaldo SD (1996) Blind Tests of Inter-analyst Correspondence and Accuracy in the Identification of Cut Marks, Percussion Marks, and Carnivore Tooth Marks on Bone Surfaces. J Archaeol Sci 23:493–507. https://doi.org/10.1006/jasc.1996.0047

Bochenski ZM (2008) Identification of skeletal remains of closely related species: the pitfalls and solutions. J Archaeol Sci 35:1247–1250. https://doi.org/10.1016/j.jas.2007.08.013

Boessneck J (1969) Osteological Differences between Sheep (Ovis aries Linné) and Goat (Capra hircus Linné). In: Brothwell D, Higgs E (eds) Science in Archaeology. A Comprehensive Survey of Progress and Research, 2nd edn. Thames and Hudson, London, pp 331–358

Boessneck J, Miller H-H, Teichert M (1964) Osteologische Unterscheidungsmerkmale zwischen Schaf (*Ovis aries* Linné) une Ziege (*Capra hircus* Linné). Kühn-Archiv 78:1–129

Bolker BM, Brooks ME, Clark CJ, et al (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24:127–135. https://doi.org/10.1016/j.tree.2008.10.008

Buckley M, Whitcher Kansa S, Howard S, et al (2010) Distinguishing between archaeological sheep and goat bones using a single collagen peptide. J Archaeol Sci 37:13–20. https://doi.org/10.1016/j.jas.2009.08.020

Burgman MA, McBride M, Ashton R, et al (2011) Expert Status and Performance. PLoS One 6:7. https://doi.org/10.1371/journal.pone.0022998

Culley C, Janzen A, Brown S, et al (2021) Iron Age hunting and herding in coastal eastern Africa: ZooMS identification of domesticates and wild bovids at Panga ya Saidi, Kenya. J Archaeol Sci 130:13. https://doi.org/10.1016/j.jas.2021.105368

Davis SJM (2017) Towards a metrical distinction between sheep and goat astragali. In: Rowley-Conwy P, Serjeantson D, Halstead P (eds) Economic Zooarchaeology. Studies in Hunting, Herding and Early Agriculture. Oxbow Books, Oxford, pp 50–82

Davis SJM (2016) Hacia una distinción métrica entre los astrágalos de oveja y cabra. In: Lloveras L, Rissech C, Nadal J, Fullola JM (eds) What bones tell us. El que ens expliquen els ossos, Monografies del SERP 12. Universitat de Barcelona, Barcelona, pp 35–58

Dinno A (2015) Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. Stata J 15:292–300. https://doi.org/10.1177/1536867x1501500117

Domínguez-Rodrigo M (2012) Critical review of the MNI (minimum number of individuals) as a zooarchaeological unit of quantification. Archaeol Anthropol Sci 4:47–59. https://doi.org/10.1007/s12520-011-0082-z

Driver JC (1992) Identification, classification and zooarchaeology. Circaea 9:35–47

Dunn OJ (1964) Multiple Comparisons Using Rank Sums. Technometrics 6:241–252

Endsley MR, Kiris EO (1995) The Out-of-the-Loop Performance Problem and Level of Control in Automation. Hum Factors 37:381–394. https://doi.org/10.1518/001872095779064555

Ericsson KA, Lehmann AC (1996) Expert and Exceptional Performance: Evidence of Maximal Adaptation to Task Constraints. Annu Rev Psychol 47:273–305. https://doi.org/10.1146/annurev.psych.47.1.273

Fernandez H (2001) Ostéologie comparée des petits ruminants eurasiatiques sauvages et domestiques (genres *Rupicapra*, *Ovis*, *Capra* et *Capreolus*): diagnose différentielle du squelette appendiculaire. Université de Genève

Gilmour GH (1997) The nature and function of astragalus bones from archaeological contexts in the Levant and eastern Mediterranean. Oxford J Archaeol 16:167–175. https://doi.org/10.1111/1468-0092.00032

Giovas CM, Lambrides ABJ, Fitzpatrick SM, Kataoka O (2017) Reconstructing prehistoric fishing zones in Palau, Micronesia using fish remains: A blind test of inter-analyst correspondence. Archaeol Ocean 52:45–61. https://doi.org/10.1002/arco.5119

Gobalet KW (2001) A Critique of Faunal Analysis; Inconsistency among Experts in Blind Tests. J Archaeol Sci 28:377–386. https://doi.org/10.1006/jasc.2000.0564

Greenlee DM, Dunnell RC (2010) Identification of fragmentary bone from the Pacific. J Archaeol Sci 37:957–970. https://doi.org/10.1016/j.jas.2009.11.029

Halstead P, Collins P, Isaakidou V (2002) Sorting the Sheep from the Goats: Morphological Distinctions between the Mandibles and Mandibular Teeth of Adult *Ovis* and *Capra*. J Archaeol Sci 29:545–553. https://doi.org/10.1006/jasc.2001.0777

Haruda AF (2017) Separating Sheep (*Ovis aries* L.) and Goats (*Capra hircus* L.) Using Geometric Morphometric Methods: An Investigation of Astragalus Morphology from Late and Final Bronze Age Central Asian Contexts. Int J Osteoarchaeol 27:551–562. https://doi.org/10.1002/oa.2576

Haruda AF, Varfolomeev V, Goriachev A, et al (2019) A new zooarchaeological application for geometric morphometric methods: Distinguishing *Ovis aries* morphotypes to address connectivity and mobility of prehistoric Central Asian pastoralists. J Archaeol Sci 107:50–57. https://doi.org/10.1016/j.jas.2019.05.002

Heilen M, Altschul JH (2013) The Accuracy and Adequacy of In-Field Artifact Analysis. An Experimental Test at Two Archaeological Sites in the Western United States. Adv Archaeol Pract 1:121–138. https://doi.org/10.7183/2326-3768.1.2.121

Holmgren R (2004) “Money on the hoof” The astragalus bone – religion, gaming and primitive money. In: Frizell BS (ed) Pecus. Man and Animal in Antiquity: Proceedings of the conference at the Swedish Institute in Rome, September 9-12, 2002. The Swedish Institute in Rome, Rome, pp 212–220

Kain MP, Bolker BM, McCoy MW (2015) A practical guide

and power analysis for GLMMs: detecting among treatment variation in random effects. PeerJ 3:24

Koerper HC, Whitney-Desautels NA (1999) Astragalus Bones: Artifacts Or Ecofacts? Pacific Coast Archaeol Soc Q 35:69–80

Kruskal WH, Wallis WA (1952) Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc 47:583–621. https://doi.org/10.1080/01621459.1952.10483441

Lau H, Whitcher Kansa S (2018) Zooarchaeology in the era of big data: Contending with interanalyst variation and best practices for contextualizing data for informed reuse. J Archaeol Sci 95:33–39. https://doi.org/10.1016/j.jas.2018.03.011

Lee Y, Nelder JA (1996) Hierarchical Generalized Linear Models. J R Stat Soc Ser B 58:619–656. https://doi.org/10.1111/j.2517-6161.1996.tb02105.x

Lee Y, Nelder JA (2006) Double hierarchical generalized linear models. J R Stat Soc Ser C Appl Stat 55:139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x

Lloveras L, Moreno-García M, Nadal J, Thomas R (2014) Blind test evaluation of accuracy in the identification and quantification of digestion corrosion damage on leporid bones. Quat Int 330:150–155. https://doi.org/10.1016/j.quaint.2013.07.033

Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. Eur J Soc Psychol 51:485–504. https://doi.org/10.1002/ejsp.2752

Lyman RL, VanPool TL (2009) Metric Data in Archaeology: A Study of Intra-Analyst and Inter-Analyst Variation. Am Antiq 74:485–504

Morin E, Ready E, Boileau A, et al (2017) Problems of Identification and Quantification in Archaeozoological

Analysis, Part I: Insights from a Blind Test. J Archaeol Method Theory 24:886–937. https://doi.org/10.1007/s10816-016-9300-4

Moscatelli A, Mezzetti M, Lacquaniti F (2012) Modeling psychophysical data at the population-level: The generalized linear mixed model. J Vis 12:1–17. https://doi.org/10.1167/12.11.26

Nims R, Butler VL (2017) Assessing reproducibility in faunal analysis using blind tests: A case study from northwestern North America. J Archaeol Sci Reports 11:750–761. https://doi.org/10.1016/j.jasrep.2017.01.012

Noddle B (1974) Ages of Epiphyseal Closure in Feral and Domestic Goats and Ages of Dental Eruption. J Archaeol Sci 1:195–204. https://doi.org/10.1016/0305-4403(74)90042-9

Pedregosa F, Varoquax G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

Pickering TR, Egeland CP, Schnell AG, et al (2006) Success in Identification of Experimentally Fragmented Limb Bone Shafts: Implications for Estimates of Skeletal Element Abundance in Archaeofaunas. J Taphon 4:97–108

Pilaar Birch SE, Scheu A, Buckley M, Çakırlar C (2019) Combined osteomorphological, isotopic, aDNA, and ZooMS analyses of sheep and goat remains from Neolithic Ulucak, Turkey. Archaeol Anthropol Sci 11:1669–1681. https://doi.org/10.1007/s12520-018-0624-8

Pöllath N, Alibert P, Schafberg R, Peters J (2018) Striking new paths – Distinguishing ancient *Ovis orientalis* from its modern domestic descendant (Karakul breed) applying Geometric and traditional Morphometric approaches to the astragalus. In: Çakırlar C, Chahoud J,

Berthon R, Pilaar Birch S (eds) Archaeozoology of the Near East XII. Barkhuis Publishing & University of Groningen, Groningen, pp 207–226

Pöllath N, Schafberg R, Peters J (2019) Astragalar morphology: Approaching the cultural trajectories of wild and domestic sheep applying Geometric Morphometrics. J Archaeol Sci Reports 23:810–821. https://doi.org/10.1016/j.jasrep.2018.12.004

Prendergast ME, Janzen A, Buckley M, Grillo KM (2019) Sorting the sheep from the goats in the Pastoral Neolithic: morphological and biomolecular approaches at Luxmanda, Tanzania. Archaeol Anthropol Sci 11:3047–3062. https://doi.org/10.1007/s12520-018-0737-0

Prummel W, Frisch H-J (1986) A Guide for the Distinction of Species, Sex and Body Side in Bones of Sheep and Goat. J Archaeol Sci 13:567–577. https://doi.org/10.1016/0305-4403(86)90041-5

Python Software Foundation (2016) Python Language Reference, version 3.6.0

R Core Team (2021) R: A Language and Environment for Statistical Computing

Salvagno L (2020) The Neglected Goat. A new method to assess the role of the goat in the English Middle Ages. Archaeopress Publishing Ltd, Oxford

Salvagno L, Albarella U (2017) A morphometric system to distinguish sheep and goat postcranial bones. PLoS One 12:e0178543

Stroup WW (2013) Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press, Boca Raton, Florida

Thompson CG, Kim RS, Aloe AM, Becker BJ (2017) Extracting the Variance Inflation Factor and Other

Multicollinearity Diagnostics from Typical Regression Results. Basic Appl Soc Psych 39:81–90. https://doi.org/10.1080/01973533.2016.1277529

Tillett BJ, Field IC, Bradshaw CJA, et al (2012) Accuracy of species identification by fisheries observers in a north Australian shark fishery. Fish Res 127–128:109–115. https://doi.org/10.1016/j.fishres.2012.04.007

Twiss KC, Wolfhagen J, Madgwick R, et al (2017) Horses, Hemiones, Hydruntines? Assessing the Reliability of Dental Criteria for Assigning Species to Southwest Asian Equid Remains. Int J Osteoarchaeol 27:298–304. https://doi.org/10.1002/oa.2524

Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Welker F, Soressi M, Rendu W, et al (2015) Using ZooMS to identify fragmentary bone from the Late Middle/Early Upper Palaeolithic sequence of Les Cottés, France. J Archaeol Sci 54:279–286. https://doi.org/10.1016/j.jas.2014.12.010

Wolfhagen J, Price MD (2017) A probabilistic model for distinguishing between sheep and goat postcranial remains. J Archaeol Sci Reports 12:625–631. https://doi.org/10.1016/j.jasrep.2017.02.022

Wolverton S (2013) Data Quality in Zooarchaeological Faunal Identification. J Archaeol Method Theory 20:381–396

Zeder MA (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci U S A 105:11597–11604. https://doi.org/10.1073/pnas.0801317105

Zeder MA, Lapham HA (2010) Assessing the reliability of criteria used to identify postcranial bones in sheep, *Ovis*,

and goats, *Capra*. J Archaeol Sci 37:2887–2905. https://doi.org/10.1016/j.jas.2010.06.032

Zeder MA, Pilaar SE (2010) Assessing the reliability of criteria used to identify mandibles and mandibular teeth in sheep, *Ovis*, and goats, *Capra*. J Archaeol Sci 37:225–242. https://doi.org/10.1016/j.jas.2009.10.002

Zhang Q, Couloigner I (2005) A New and Efficient K-Medoid Algorithm for Spatial Clustering. In: Gervasi O, Gavrilova ML, Kumar V, et al. (eds) Computational Science and Its Applications – ICCSA 2005. ICCSA 2005. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 181–189

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14. https://doi.org/10.1111/j.2041-210x.2009.00001.x