Testing inter-observer error under a collaborative research framework for studying lithic shape variability

Lucy Timbrell (lucy.timbrell@liverpool.ac.uk)
University of Liverpool
Christopher Scott
University of Liverpool
Behailu Habte
National Museum of Ethiopia
Yosef Tefera
National Museum of Ethiopia
Hélène Monod
Musée de l'Homme
Mouna Qazzih
Institut National des Sciences de l'Archéologie et du Patrimoine
Benjamin Marais
Iziko Museums of South Africa
Wendy Black
Iziko Museums of South Africa
Christine Maroma
National Museums of Kenya
Emmanuel Ndiema
National Museums of Kenya
Struan Henderson
Mossel Bay Archaeological Project
Katherine Elmes
Mossel Bay Archaeological Project
Kimberly Plomp
University of the Philippines
Matt Grove
University of Liverpool

Research Article

Keywords: stone tools, metric measurements, geometric morphometrics, 3D printing, inter-observer reliability

Posted Date: June 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1752934/v1

Abstract

Evaluating error that arises through the aggregation of data recorded by multiple observers is a key consideration in many metric and geometric morphometric analyses of stone tool shape. One of the most common approaches involves the convergence of observers for repeat trails on the same set of artefacts, however this is logistically and financially challenging when collaborating internationally and/or at a large-scale. We present and evaluate a unique alternative for testing inter-observer error, involving the development of 3D printed copies of a lithic reference collection for distribution among observers. With the aim of reducing error, clear protocols were developed for photographing and measuring the replicas, and inter-observer variability was assessed on the replicas in comparison with a corresponding data set recorded by a single observer. Our results demonstrate that, when the photography procedure is standardized and dimensions are clearly defined, the resulting metric and geometric morphometric data are minimally affected by inter-observer error, supporting this method as an effective solution for assessing error under collaborative research frameworks. Collaboration is becoming increasingly important within archaeological and anthropological sciences in order to increase the accessibility of samples, encourage dual-project development between foreign and local researchers, and reduce the carbon-footprint of collection-based research. This study offers a promising validation of a collaborative research design whereby researchers remotely work together to produce comparable data capturing lithic shape variability.

1. Introduction

Shape analyses are becoming an increasingly popular methodology for examining lithic variability in the archaeological record. As such, traditional linear metrics and geometric morphometrics (GMM) are often employed to capture morphological information on stone tools (Cardillo 2010; Lycett and von Cramon-Taubadel 2015; Matzig et al. 2021). Combining morphological data from multiple observers is frequently necessary in studies of lithic assemblages, to increase sample size and/or to perform inter-site / inter-assemblage analyses, yet this can be problematic due to the possibility of introducing inter-observer error into the data (Lyman and VanPool 2009). Such error has multiple potential sources, can be introduced at various stages in the workflow, and can skew results by obscuring any ‘real’ signals in the data (Fruciano 2016); examining the magnitude of inter-observer error is therefore imperative to validate whether meta-analyses are robust. International researchers are increasingly being encouraged to work collaboratively in order to remotely produce archaeological and anthropological datasets (Chang and Alfaro 2015; O’Leary and Kaufman 2011; Scerri et al. 2020; Timbrell 2020, 2022) – in some case even crowdsourcing morphometric data (Chang and Alfaro 2015). However, it is frequently impossible for observers to converge on the same material to record repeat trials for an inter-observer repeatability assessment. Such tests therefore need to be appropriate for the specific research design, and customized solutions for evaluating error under collaborative research frameworks should be developed (Fruciano 2016). Here we present an innovative analysis of inter-observer error involving the compilation of standardized photographs and measurements of lithics from multiple observers for metric and GMM analysis (Timbrell 2022).

Traditionally, lithic shape variation has been examined through qualitative descriptions (Inizan et al. 1999), typological classification (Bordes 1961), and/or linear measurements (Roe 1964; McNabb 2017). Advancements in biological morphometrics and computing have meant that geometric morphometrics are now also routinely applied in the analysis of lithic morphologies (Bookstein 1991; Buchanan et al. 2018; Cardillo 2010; Lycett 2009; Serwatka and Riede 2016). GMM approaches are split into methods that use landmarks and outlines, the former representing shape through homologous points (landmarks) superimposed on a two-dimensional (2D) or three-dimensional (3D) object and the latter applying geometric descriptions of homologous outlines or surfaces (Mitteroecker 2021). Landmark based methods allow for specific aspects of morphology to be captured without the inclusion of random noise (i.e., shape dimensions that are not pertinent to the research question), however, their application to certain non-biological structures, such as lithics and other archaeological artefacts, is often more difficult as the identification of homologous landmarks can be subjective (Okumura and Araujo 2018). Outline based GMM, on the other hand, avoids certain issues of homology through quantifying the gross

shape of each specimen (Klingenberg 2008), making them ideal for describing shape variation of lithics in archaeological studies (e.g., Iovita 2009, 2011; Ivaonovaité et al. 2020; Matzig et al. 2021; Mesfin et al. 2020; Wang and Marwick 2020).

Assessment of the levels of inter- and intra-observer error under different methodological approaches to studying lithic shape is vital, and several studies have examined error in metric and GMM analyses at different phases of the workflow (Evin et al. 2020; Fagerton et al. 2014; Lyman and VanPool 2009; Macdonald et al. 2020; Menedez 2017; Osis et al. 2015; Perini et al. 2005; Robinson and Terhune 2017; von Cramon-Taubadel et al. 2007; Yezerinac et al. 1992). Problematic landmarks, i.e., those that are difficult to consistently locate, can be a source of error in landmark based GMM analysis (Fagerton et al. 2014; Menedez 2017; Robinson and Terhune 2017; von Cramon-Taubadel et al. 2007), even for experienced observers (Chang and Alfaro et al. 2015). von Cramon-Taubadel et al. (2007) found that repeating the digitization procedure was the most suitable method for assessing the precision of landmarks, with adequate landmark definitions imperative for reducing error. Yezerinac et al. (1992) also found that ill-defined measurements were a factor increasing error in metric data, in addition to operator experience, the precision of the measuring device, and the conditions under which the measurements are made, such as lighting. Combining metric measurements from more than one observer therefore is likely to be suitable only when the dimensions are standardized and easily measured, and the conditions, the precision and quality of the equipment, and the technique of recording the data are comparable (Lyman and VanPool 2009).

Comparatively, fewer studies have examined the levels of inter-observer error in outline based GMM methods. Evin et al. (2020), in an investigation of error between morphometric approaches, found that although methods that employ landmarks were the most sensitive to error, outline data saw relatively lower levels of intra-observer error compared to inter-observer error, with photography being an influential source of variance between observers. Digital photography is widely used in 2D GMM as it is inexpensive, easy to perform, and does not require extremely specialist knowledge or equipment, with the digitization of landmarks and/or outlines on the resulting images providing a 2D representation of the 3D object. The focal length and specifications of the lens used can, however, cause parallax error, optical distortion that occurs when the specimen is too close or not directly centered beneath the lens (fisheye). Nonetheless, several studies employing both landmark and outline methods suggest that 2D GMM data are minimally affected by parallax error, especially when the camera set-up is standardized and calibrated, with deviations small and constant enough for accurate analyses (Caple et al. 2018; MacDonald et al. 2020; Mullin and Taylor 2002; Riano et al. 2009). Overall, outline-based methods are likely more suitable for collaborative research designs in studies of lithic shape, due to the objectivity of data capture, the fact that landmark methods have high rates of inter-observer error, though this is more pertinent during landmark digitization than object photography (Evin et al., 2020), and the potential to reduce inter-observer error through the standardization of the photography procedure.

Although inter-observer error is a concern in any collaborative research design, collating data from multiple observers is often necessary in archaeological research, be it to increase sample sizes, facilitate interdisciplinary research and/or enable access to disparate data (Timbrell 2020). The latter is especially important when considering issues of income-disparity, childcare, and disability that can disproportionately disadvantage researchers who are unable to travel extensively to collect data. Global catastrophes, such as pandemics, climate change, and conflict, can also temporarily delay international research through the constraints imposed on travel and safety, requiring researchers to develop scientifically-sound remote models of data generation (Scerri et al. 2020). Timbrell (2022) presents such a framework, which involves the documentation of lithic shape by multiple collaborators. These types of approaches have additional benefits for decreasing the carbon footprint associated with accessing multiple international samples and fostering knowledge-sharing through dual project development and the division of responsibilities so that both foreign and local researchers take on principal roles within a given project, which is particularly crucial across the Global North-South divide (Chirikure 2015; Douglass et al. 2020; Else, 2022). Indeed, collaborative approaches accord with the open science initiative in archaeology, which advocates that data stewardship should be centered around researchers collecting and sharing data on behalf of the scientific community, as opposed for the betterment of a single individual's career (Marwick et al. 2017). While collaborative data collection offers a promising new framework for generating and sharing data internationally, the analysis of inter-observer error is imperative

to validate such an approach. Here, we present a unique test that involves the production of 3D printed replicas of a lithic reference collection, which can be distributed among observers and recorded following the same protocols used to collect the actual data. Using this approach, we evaluate whether the compilation of data from multiple observers is conducive to error, and thus could negatively bias the results of a collaborative study.

2. Materials

Six lithic points were knapped using fine-grained flint from Caistor Quarry, Caister St Edumunds, UK, and scanned for 3D printing at the University of Liverpool (Fig. 1). The reference tools varied in both size and shape, encapsulating a range of morphologies characteristic of the empirical sample to be studied in the main project (African Middle Stone Age assemblages). While flint is not a feature of African lithic assemblages, it could be considered representative of the finer-grained materials, such as obsidian, chert, and heat-treated silcrete, exploited during the Middle Stone Age (Key et al. 2021; Sahle et al. 2013). The tools were produced on flakes and retouched using: (1) direct freehand hard hammer percussion (quartzite hammerstones); (2) direct soft free hand hammer percussion using an antler hammer; (3) handheld pressure flaking using an antler tine supported in a tanned leather pad. Each tool was colored blue using craft enamel spray paint to aid scanning.

Next, each lithic was scanned with a freshly calibrated Einscan Pro 2X structured light scanner with colour camera, using combined feature and texture mapping in the high-resolution setting. Initial scans were performed with the lithics placed vertically in a foam holder using fixed scan mode aligned with an automated turntable and coded targets (scans taken every 11.25 degrees, i.e., 32 scans). The models were then completed by switching to the ‘align by’ feature using the turntable (32 scans) and the lithic was rescanned (2–3 times) until a complete model was achieved. All alignment was automatic to produce a watertight mesh; no holes were filled. Each model was sharpened using the Einscan high-setting and saved as .obj files without decimation (see Supplementary Online Table 1 for further data on each model).

The 3D models were processed for printing using Chitubox v1.8.1. Medium-sized automated supports were applied using this software at 90% total coverage to provide a strong foundation for the 3D prints. We used an Elegoo Mars 2 Resin printer, with a new printer film, using standard grey Elegoo LCD UV curing 405 Nm photopolymer resin with recommended Elegoo settings (Fig. 2). The prints were extracted from the print bed and the supports were removed by hand prior to being rinsed in ethanol and cured in direct sunlight. Each tool was printed six times to create six copies of the assemblage, resulting in 36 prints in total.

3. Methods

Each tool was assigned a number (Tool 1–6; Fig. 1) and a replica copy of the assemblage was sent to researchers at six independent institutions (Table 1). Data collection protocols, outlined in detail by Timbrell (2022) and described in Supplementary Online Resource S1, were developed to standardize the documentation of lithic shapes through photography and measurements. These procedures were followed by all observers across the study to produce equivalent data. Instructions for object position, camera position and settings, and lighting were specified and tightly controlled (Supplementary Online Resource S1). In addition, a scale (sourced in situ by the observers) was placed in each photograph to ensure a measure of size was recorded. Table 1 reports the camera and lenses used to capture each replica assemblage in the study; high quality equipment was accessed by all observers either through their institution directly or through funding provided by the project. Three basic measurements on each tool were also taken to record morphological length, width, and thickness (see Supplementary Online Fig. 1 for a schematic) at a resolution of 0.1mm. We defined length as the maximum dimension of the point, width as the maximum measurement in the perpendicular dimension to length, and thickness as the maximum measurement in the third dimension, following Shea (2020).

Table 1
Summary of the observers and the photography equipment used. This equipment was sourced locally; in most cases, the institutions already had access to the necessary apparatus, however in some cases it was rented and/or purchased and donated to the institution after the project, following guidelines provided by The Wenner Gren Foundation.

Assemblage number	Institution	Abbreviation	Country	Camera body	Camera lens
1	Institut National des Sciences de l'Archéologie et du Patrimoine	INSAP	Morocco	Nikon D7100	Nikon AF-S Micro Nikkor 105mm
2	Iziko Museums of South Africa	IM	South Africa	Canon 6D II	Canon 100mm 2.8 Macro
3	Mossel Bay Archaeological Project	MBAP	South Africa	Nikon D300s	Nikon AF Micro Nikkor 60mm 1:2.8D
4	National Museum of Ethiopia	NME	Ethiopia	Canon EOS DSLR 200D	Canon Tamron 60mm Macro Di II
5	National Museums of Kenya	NMK	Kenya	Nikon D5300	Nikon AF-S Micro Nikkor 40mm
6	Musée de l'Homme	MH	France	Nikon D5200	Nikon AF-S Nikkor 24-70mm

Prior to distribution among institutions, all 36 replicas were also recorded by a single observer (LT) to produce a comparative dataset. Photography was performed using a Canon M50 camera with an EF-S 60mm f/2.8 Macro USM lens and the three measurements were taken using digital calipers. This enabled us to determine the magnitude of intra-observer measurement error, for comparison with the magnitude of inter-observer error, had the project been carried out by a single individual under a traditional research framework.

Data were uploaded onto a communal data sharing platform (Google Drive) by each observer for processing and analysis by a single observer (LT). Analyses were performed in the R software environment (R Core Team 2020). Data and code can be found on the GitHub repository for the project: https://github.com/lucytimbrell/error_analysis_lithics/

3.1. Metric analyses

We first computed the intra-class correlation coefficient (ICC) using the 'psych' R package (Revelle, 2022) to assess the agreement between the six observers in measuring the six tools for length, width, and thickness. The ICC compares the variability within repeat measurements whilst contrasting variability between groups of measurements (Barlett and Frost 2008; Fruciano 2016; Koo and Li 2016; Shrout and Fleiss 1979). Specifically, we used a two-way mixed effects model to compute the ICC, with the set of observers considered a fixed effect. To assess the reliability of data collection, we next calculated and compared the mean, variance, technical error of measurement (TEM) and percentage technical error of measurement ($\%TEM$). The mean and variance (expressed as the standard deviation) were calculated for each measurement on each tool, with the TEM and $\%TEM$ calculated to compare pairs of observers across all measurements on all tools. The TEM reflects measurement precision between observers, and is calculated as:

$$TEM = \sqrt{\frac{\left(\sum_1^N\left(\sum_1^K M^2\right)\right) - \left(\left(\sum_1^K M\right)^2 / K\right)}{N(K-1)}}$$

where N is the number of subjects, K is the number of observers, and M is the measurement (modified from Ulijaszek and Kerr [1999]). The %TEM represents the magnitude of the error as a percentage of the mean of the measurement/variable studied. It is calculated as:

$$\%TEM = 100\left(\frac{TEM}{\bar{v}}\right)$$

where $\bar{v}$ is the average value of the raw measurements, taken across all measurements on all tools by multiple observers. The values obtained for these metrics must be subjectively assessed according to the research question, as there is no standard applied threshold of error deemed to be 'acceptable'. Following Lyman and VanPool (2009)'s analyses of projectile points, we propose that a $\%TEM$ of < 4 could be an acceptable level of error without negative consequences on the results. Lastly, we calculated the coefficient of reliability (R), which ranges from 0 to 1, with 1 indicating very high congruence between measures. We used the following formula outlined in Lyman and VanPool (2009):

$$R = \sigma_v^2 / \left(\sigma_v^2 + \sigma_d^2\right)$$

where σ_v^2 is the variance of all raw measurements on all tools taken by two observers and σ_d^2 is the variance of the difference between those two sets of measurements. Similarly to the ICC, the coefficient of reliability distinguishes between the variability between the specimens and that which results from random measurement error. However, whilst R can only be calculated between pairs of observers, the ICC represents an overall metric for measurement error across all observers.

Random error can inflate the amount of variance within a sample, resulting in a loss of statistical power as noise obscures true differences in means (Fruciano 2006; Yezerinac et al. 1992). To evaluate the levels of error in the multiple observer data in relation to the single observer data, we used two-sample t-tests to compare differences in mean and F-tests to compare differences in variance. If there is high inter- and/or intra-observer error, variation within replicas of the same tool will be increased and differences in the mean values for each tool will be significantly different.

3.2. Two-dimensional geometric morphometric analysis

In preparation for GMM analysis, each image was processed using the 'object select' tool in Adobe Photoshop, which automatically determines the contour of the object. Once the contour was highlighted, the object was filled with solid black to help facilitate the extraction of outline data. All processed images were then synthesized into a single thin-plate spline (.tps) file using tpsUtil, and the outline data were extracted using tpsDig2. The outline of each artefact was represented by an average of 2856 equidistant points, which were scaled through the specification of the pixel to centimeter ratio for each image (see Supplementary Online Fig. 2 for a visualization of the data). The outline data were saved as (x,y) coordinates within the .tps file and imported into R.

Using the 'Momocs' R package (Bonhomme et al. 2014), the outlines were standardized following Bonhomme et al. (2017) by normalizing to a common centroid, scaling to centroid size, and aligning along the long axis of the object. We then performed elliptic Fourier Analysis (EFA) to convert the geometric data to frequency data, with the outline decomposed into a series of repeating trigonometric functions, referred to as harmonics (Caple et al. 2017; Fig. 3). The appropriate number of harmonics were identified to capture sufficient information on shape; this was deemed to be 8 harmonics, achieving 99% harmonic power (Caple et al. 2017).

Next, we performed a principal components analysis (PCA) on the elliptic Fourier coefficients to reduce the dimensionality of the data. Principal components (PCs) are constructed to highlight the main axes of morphological variance (Zelditch et al. 2004). Like with the metric data, we calculated the ICC and R values to partition the variance from the inter-observer error for the PC scores of repeat captures (Daboul et al. 2018; Fruciano 2006). Due to the nature of PC scores, we were unable to obtain an informative relative measure of dispersion ($\%TEM$) and instead refer to the standard deviation (calculated as the square root of the variance) as absolute measures of dispersion. This is because, when the mean of a set of repeat captures falls close to the mean of a PC (~ 0) and has a low standard deviation (~ 0), the %TEM would be very high despite the tight

clustering of the repeated measures along that PC. In addition, we applied linear discriminant analysis (LDA) to the PC scores, with the equal sample sizes used as the prior group probabilities (1/6) of a repeat belonging to a certain group based on their outline shape alone (Mitteroecker and Bookstein 2011). In this analysis, we tested firstly whether the tools could be distinguished based on their shape alone, and then whether the observers could be identified. One would expect high classification results when discriminating between tools and low classification results when discriminating between observers if inter-observer error is low.

4. Results

4.1. Linear metric analysis

We first explored whether the measurements were recorded consistently on the replicas between observers. Figure 4 shows the distribution of the multiple observer data through boxplots; most of the measurements have very limited variance around the mean, and all tools were significantly different to each other across all measurements when tested using Tukey's Honestly Significant Difference (HSD; $p < 0.001$). Thickness is the most variable dimension recorded, probably because it is more difficult to orient the tool for this measurement than it is for length or width. Calculation of the coefficient of reliability between each pair of observers found that all values of R were > 0.999, suggesting that over 99% of the variance in each measurement is due to variability between the specimens as opposed to error. We calculated the TEM as 0.368 and the $\%TEM$ as 0.908, supporting that less than 1% of the variance in the dataset is related to measurement error. Finally, the ICC score confirmed that there is very high absolute agreement between the observers ($ICC = 1$, $p < 0.001$).

We then compared the measurements taken by multiple observers with those taken by a single observer as a means of comparing intra- and inter-observer error. We first calculated the coefficient of reliability for the single observer for each pair of replica assemblages – we found that all values were > 0.999, indicating very high congruence between repeat captures by the single observer. Table 2 reports the mean and standard deviation of length, width, and thickness for the single observer compared to multiple observers; two-sample t-tests found that there were almost no statistically significant differences in means between the data sets ($1/36 = p < 0.05$; Table 3). However, F-tests found that half of the measurements show statistically significance differences in variance, particularly along length and width (Table 3). This demonstrates that the single observer is generally less prone to error, which is likely due to a combination of the familiarity of this observer to both the metric definitions and the assemblage and the fact that the same equipment was used to measure all of the replicas. Nonetheless, that fact that these differences in variance only resulted in a single instance of significant difference in mean, plus the standard deviation does not exceeds 0.7 mm, suggests that the effects of inter-observer error are minimal on the results.

Table 2
Summary statistics reporting the mean (m) and standard deviation (sd) obtained for length, width and thickness, recorded by multiple observers versus a single observer for each tool (1–6). Standard deviation values have been rounded to 3 decimal places.

Tool	Length (mm)				Width (mm)				Thickness (mm)			
	Multiple		Single		Multiple		Single		Multiple		Single	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
1	86.2	0.471	86.3	0.175	40.9	0.308	40.9	0.103	13.4	0.281	13.3	0.248
2	67.6	0.266	67.6	0.089	37.3	0.258	37.5	0.228	10.3	0.141	10.4	0.075
3	66.0	0.613	66.3	0.137	23.4	0.266	23.3	0.225	6.87	0.472	6.72	0.075
4	74.6	0.279	74.4	0.299	48.4	0.374	48.5	0.103	11.9	0.151	11.8	0.105
5	59.7	0.133	59.7	0.075	27.4	0.335	27.6	0.082	9.45	0.281	9.48	0.147
6	87.3	0.405	87.4	0.063	44.7	0.659	44.6	0.126	14.3	0.415	14.2	0.117

Table 3
P-values from t-tests (difference in mean) and F-tests (difference in variance) comparing the metrics (length, width and thickness) for each tool (1–6) measured by multiple observers versus a single observer. Statistical significance ($p < 0.05$) is marked by an asterisk (*). All values have been rounded to 3 decimal places.

Tool	Length (mm)		Width (mm)		Thickness (mm)	
	T	F	T	F	T	F
1	0.815	0.049*	0.632	0.032*	0.673	0.792
2	0.678	0.032*	0.264	0.792	0.240	0.193
3	0.342	0.005*	0.498	0.724	0.475	0.001*
4	0.342	0.879	0.689	0.013*	0.037*	0.446
5	0.608	0.238	0.152	0.008*	0.804	0.182
6	0.575	0.001*	0.646	0.002*	0.653	0.015*

4.2. Geometric morphometric analysis

PCA was used to highlight variance in the multiple observer data. The first 3 PCs represented > 90% of the variation between the replicas, and thus were explored in this study. Figure 5 demonstrates the shape differences highlighted by PC1-3. PC1 represents 59.7% of the total variance, whilst PC2 and PC3 account for 33.4% and 3% respectively (see Supplementary Online Fig. 3 for scree plot of PC loadings and cumulative variance).

When the first 3 PCs are plotted against each other, clear clustering occurs, demonstrating that replicas of the same tool tend to share more similarities than that of different tools (Fig. 6). However, there is notable variation within tools along PC3, suggesting that inter-observer error deriving from photography equipment and set-up is prevalent in this dimension. PC3 is an axis of variation represented by slight asymmetries in convexity at the proximal end (Fig. 5), thus likely reflecting parallax error between observers. Additionally, we note some overlap between certain tool groups, although this is primarily because these tools share similar shapes once size is removed (Supplementary Online Fig. 2). For example, Tool 5 sometimes plots within the range of variation for Tool 1 and only shows statistically significant differences in mean from this tool along PC2 ($p < 0.008$; see Supplementary Online Table 2 for Tukey's HSD results comparing differences in mean between tools). To tease apart the variation between the tools and that associated with error, we calculated the coefficient of reliability between

each pair of observers, which ranged between 0.960 and 0.999 (Table 4), suggesting that < 4% of the variance is due to inter-observer error, which lies within our acceptable threshold. The ICC was computed using the first 3 PC scores to determine levels of similarity between the six observers, whilst taking into account the variability between the tools, and found an almost perfect agreement (ICC = 0.99, p < 0.001). Finally, we found that an LDA could discriminate accurately between the replica groups (94% classification accuracy) and could not differentiate between observers (0% classification accuracy).

Table 4
Coefficient of reliability (R) values for pair-wise combinations of observers using the first 3 PC scores. For observer abbreviations and associated assemblage numbers, see Table 1. All values have been rounded to 3 decimal places.

	INSAP	IM	MBAP	NME	NMK
IM	0.988				
MBAP	0.978	0.960			
NME	0.984	0.975	0.995		
NMK	0.969	0.969	0.985	0.992	
MH	0.989	0.978	0.993	0.999	0.990

Next, we compared the levels of error obtained when collating photographs from multiple observers and that which arises when all replicas are photographed by the same observer. We performed another PCA with data acquired from both sets of images (see Supplementary Online Fig. 4–5 for PC contributions and loadings) and produced scatterplots of PC1-3. Figure 7 demonstrates clear clustering between tools recorded in both sets of data along PC1 and PC2. However, along PC3 there is clear variability within repeats when grouped by observer (multiple vs single). F-tests found that the variance among certain tools was only significantly higher for the multiple observers in three cases, i.e., Tool 4 and 1 along PC3 and Tool 4 along PC1 (Table 5). Two sample-t-tests found statistically significant differences in means between the data sets, but these are limited (5/36 = p < 0.05; Table 5). Table 6 and Fig. 7 demonstrate that the data collected by a single observer returns lower variance, though this pattern is not strong, and, in a few cases, it is slightly higher under this strategy, though not significantly so. We finally calculated the coefficient of reliability for the single observer between each of capture of the replica assemblages – Supplementary Online Table 3 shows that the R values ranged from 0.994 to 0.999, suggesting that < 1% of the variance in the single observer data is due to intra-observer error.

Table 5
P-values from t-tests (difference in mean) and F-tests (difference in variance) comparing the principal component (PC) scores of the repeats of each tool (1–6) captured by multiple observers verses a single observer. Statistical significance (p < 0.05) is marked by an asterisk (*). All values have been rounded to 3 decimal places.

Tool	PC1		PC2		PC3	
	T	F	T	F	T	F
1	0.282	0.068	0.556	0.141	0.110	0.001*
2	0.091	0.463	0.114	0.162	0.188	0.671
3	0.006*	0.119	0.067	0.873	0.335	0.115
4	0.082	0.029*	0.009*	0.384	0.099	0.006*
5	0.004*	0.663	0.003*	0.257	0.000*	0.411
6	0.954	0.056	0.095	0.157	0.441	0.939

Table 6
Summary statistics reporting mean (m) and standard deviation (sd) of principal component (PC) scores of the repeats of each tool (1–6), captured by multiple observers versus a single observer. All values have been rounded to 3 decimal places.

Tool	PC1				PC2				PC3			
	Multiple		Single		Multiple		Single		Multiple		Single	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
1	-0.034	0.006	-0.031	0.003	0.039	0.007	0.037	0.003	-0.011	0.016	-0.023	0.003
2	0.034	0.003	0.031	0.002	0.046	0.002	0.042	0.004	0.042	0.004	0.039	0.004
3	-0.131	0.001	-0.135	0.002	-0.111	0.003	-0.107	0.004	0.009	0.004	0.007	0.002
4	0.174	0.005	0.169	0.002	-0.074	0.004	-0.081	0.003	-0.004	0.01	-0.012	0.002
5	-0.03	0.001	-0.033	0.002	0.03	0.003	0.025	0.001	-0.014	0.001	-0.023	0.002
6	-0.008	0.007	-0.008	0.003	0.078	0.002	0.074	0.004	-0.006	0.006	-0.004	0.006

5. Discussion

Our results demonstrate that the levels of inter-observer error permeating shape data collated under a collaborative research framework, where the research protocols are outlined in detail, fall within the acceptable threshold. We found that, inevitably, increases in error occur as a consequence of relying on multiple observers, who each have access to different equipment, yet we do not deem this to be significant enough to highly distort the results towards a different conclusion about the data. Therefore, our innovative 3D printing approach and the results reported here have important implications for error assessments of linear metric and GMM data when recording lithic shape as well as the aggregation of data collected by multiple observers.

Outline based GMM was found to be slightly more sensitive to inter-observer error than metric methods. As Caple et al. (2018) point out, EFA involves global descriptors capturing around 99% of the variance in the outline shape, and therefore discrepancies between images lead to error in the coefficients dispersed throughout the full outline. Therefore, even if the error is not equally distributed, it is measured as such and consequently outline methods are often more sensitive to error than linear methods that capture only certain dimensions of an object. 2D outline based GMM provides comprehensive morphological information on the gross outline shape of an object, whereas linear metrics are able to capture aspects of the 3D shape but in much less detail; the increase in the morphological information captured, plus the added potential for automated data capture (e.g. Bonhomme et al. 2014; Matzig 2021) and impressive shape visualization (e.g. Fig, 5), will be worth the potential increase in error with 2D GMM in many scenarios. Our use of PCA to highlight axes of variance within lithic shape assemblages also demonstrates that inter-observer error does not affect all PCs equally. As outlined by Page (1976), subtle errors in each variable are combined in multivariate analyses and can be extracted by a single or small set of PCs, although they may also describe real aspects of covariance and so require careful consideration as to their source. When undertaking metric analyses, it is possible to assess error in each individual measurement; if the metrics are combined via dimension reduction methods such as PCA, the contributions of each individual measurement to each PC are readily identifiable through the PCA coefficients. This is less feasible with GMM data, particularly when using outlines and semi-landmarks, and in such cases, it is preferable to assess error on each of the leading PCs, as demonstrated above, rather than on each set of coordinates, which can be very numerous. Overall, error is impossible to avoid completely, and indeed the imperfect fidelity of cultural transmission means that copying errors can naturally occur during the knapping process and inflate variance between and within assemblages (Eerkens and Lipo 2005; Schillinger et al. 2014). In this sense, error is certain to arise within a data set capturing lithic variability; however, steps can be taken to ensure it is minimized, such as standardization of data acquisition, processing, and analytical procedures, calibration, high quality equipment, and assessment of error through repeat measures (Evin et al. 2020; Lyman and VanPool 2009; Robinson and Terhune 2017;

Yezerinac et al. 1992). In the case of the current study, we determine that inter-observer error is low enough for accurate analyses under both methods, especially as the high ICC and R values demonstrate acceptable levels of congruence between the six observers.

Through the development of clear research protocols, our results demonstrate that multiple observers can successfully work together to produce sets of comparable data for aggregation. We believe that collaborative research designs, such as the one reported in Timbrell (2022), play an integral role in addressing the vulnerabilities of international research to disruption, revealed most recently in 2020 by the outbreak of coronavirus (COVID-19), which halted both domestic and international travel as well as social interaction. Our results suggest that, as well as single researchers visiting multiple collections to independently access lithic samples, international colleagues are also able to work together *in situ* to generate data, thereby building resilience in archaeological practice (Douglass et al., 2020; Scerri et al., 2020). We stress though that collaborative research designs should involve an equitable partnership in relation to the data, following the imminent Cape Town statement (see Else, 2022), with all researchers being involved in all stages of the research, from planning and protocol development to publication and dissemination (Chirikure 2015; Douglass et al. 2020). In this way, dual project development can enable local researchers to benefit from international archaeological research, thereby avoiding some (but not all) of the neo-colonial 'helicopter' practices that have been hugely criticized in archaeological and anthropological sciences, particularly in Africa (Ackermann 2019; Athreya and Ackermann 2019; Sahle 2021). We have provided here an initial pilot test of collaborative data collection using a 3D printing approach. This approach is unique and, to our knowledge, has not yet been applied in the context of lithic variability nor inter-observer error assessments. We propose that future studies should aim to reproduce our approach with more expanded samples of replica artefacts, and discuss three important aspects of potential future study design below.

The first aspect relates to the use of statistics and simple metrics for reporting inter-observer error. Statistics such as the ICC and $\%TEM$ express the error variance relative to the overall variance of the sample; variance is decomposed into that due to genuine variation among the artefacts and that due to variation among the observers (including that due to different individuals, their different cameras, lenses, etc.). Whilst this approach has many advantages, one immediate drawback is that these statistics are directly affected by the magnitude of genuine variation in both the sample of artefacts and in the dimensions measured. A given, constant level of measurement error will appear large when the artefacts measured are highly standardized, but small when the artefacts measured are highly variable. Even if one were to measure the widths and lengths of a set of highly standardized artefacts, a given level of measurement error would appear smaller the further the ratio of width to length is from unity, as this would increase the magnitude of genuine variation in the measurements taken. For this reason, it is always valuable to present simple indices of *absolute* error (such as standard deviation or variance) for *single* measurements alongside the indices of relative error variance across all measurements provided by the ICC and $\%TEM$. Such simple indices are valuable in assessing inter-observer error even when the ultimate study involves more sophisticated morphological analyses, such as those based on GMM. In the current study, Table 2 presents such indices, and demonstrates that levels of error are minimal (the largest standard deviation among multiple observers for a single measurement = 0.613mm).

The second aspect relates to the exploration of the effects of the raw material used for production of the reference collection on the results of comparative studies. In this study, we used flint because it was available and accessible at the University of Liverpool, where the materials were prepared. This fine-grained raw material tends to produce well-defined features and edges, and so it would be interesting to replicate the approach with a more coarse-grained material, such as quartzite, chert, calcrete or sandstone. This is especially pertinent in our case as the shapes obtained from these materials are likely to be more representative of the actual African stone tools that have been recorded in the main project. However, we note that heat-treated silcrete may achieve a grain as fine as flint (Key et al. 2021), and that obsidian can be even finer-grained than flint; since both silcrete and obsidian are raw materials commonly found in African Middle Stone Age assemblages, we suggest that the flint used here acts as a suitable middle ground in terms of granularity and can therefore be considered as broadly comparable to those raw materials studied in the main project.

Finally, an aspect of variation between individual replicas that we did not explicitly measure is that which can arise through 3D printing. Zeng and Zou (2019) outline some of the factors that can affect the precision of 3D printing, which include slicing and support errors. However, we propose that, even if there are printing errors present in our replicas, these are likely minimal due to the highly comparable data obtained across the project. Additionally, printing errors should not contribute to differences between the two data collection strategies as both the multiple observers and the single observer recorded measurements from the same set of replicas.

6. Conclusion

Aggregating lithic shape data requires careful consideration in to order reduce potential sources of inter-observer error that can result in detrimental consequences on the results and their interpretation. Our analysis of metric and outline-based 2D GMM data from multiple observers found that the former performed slightly better than the latter in our tests of inter- and intra-observer error, primarily due to differences in the nature and detail of the morphological information obtained, though both approaches returned levels of error deemed acceptable for accurate analyses. Standardization of the data collection procedure is vital for ensuring that congruence between observers is maintained, though we note that this alone cannot completely eradicate error as we find that variability between observers can still be detected within our data to a (sometimes) significant extent. Nonetheless, we believe that producing replica samples through 3D printing could have many useful applications within archaeological and anthropological sciences beyond the study of error in the analysis of lithic assemblages and should be adopted more widely in assessments of inter-observer error as an integral component of international collaborations between institutions.

Author declarations:

Declarations

Author declarations:

Funding:

This project was supported by funding awarded to LT by the Leakey Foundation (*Movement, interaction, and structure: modelling population networks and cultural diversity in the African Middle Stone Age*), the Wenner Gren Foundation (Gr. 10157) and the Lithic Studies Society (Jacobi Bursary, 2020).

Conflicts of interest:

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval/declarations:

Not applicable

Consent to participate:

Not applicable

Consent for publication:

Not applicable

Data/code availability statement:

All data and R code can be found on the project's repository, and was made available for the peer-review of this article: https://github.com/lucytimbrell/error_analysis_lithics/

Author contributions:

All authors contributed to the study conception and design. Funding was acquired by LT. CS and LT performed the material preparation. Data collection was carried out by HM, MQ, BM, CM, BH, YT, SH, KE and LT, under the supervision of EM, WB, KP and MG. LT and MG performed data processing and analysis. The first draft of the manuscript was written by LT and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Ackermann RR (2019) Reflections on the history and legacy of scientific racism in South African paleoanthropology and beyond, J Hum Evol 126: 106–111. https://doi.org.10.1016/j.jhevol.2018.11.007
2. Athreya S, Ackermann RR (2019) Colonialism and narratives of human origins in Asia and Africa. In: Porr M, Matthews JM (eds) Interrogating Human Origins: Decolonisation and the Deep Past. Routledge, London. https://doi.org/10.4324/9780203731659-4
3. Bartlett JW, Frost C (2008) Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics and Gynecology 31:466–75. https://10.1002/uog.5256
4. Bonhomme V, Picq S, Gaucherel C, Claude J (2014) Momocs: Outline Analysis Using R, J Stat Softw 56: 1–24. https://doi.org/10.18637/jss.v056.i13
5. Bonhomme V, Forster E, Wallace M, Stillman E, Charles M, Jones G (2017) Identification of Inter- and Intra-species Variation in Cereal Grains Through Geometric Morphometric Analysis, and its Resilience under Experimental Charring. J Archaeol Sci 86: 60–67. https://doi.org/10.1016/j.jas. 2017.09.010
6. Bookstein FL (1991) Morphometric tools for landmark data: geometry and biology. Cambridge University Press, Cambridge
7. Bordes F (1961) Typologie du Paléolithique Ancien et Moyen. Bordeaux
8. Buchanan B, Andrews B, O'Brien M, Eren MI (2018) An assessment of stone weapon tip standardization during the Clovis–Folsom transition in the Western United States, Am Antiq 83:721–734 https://doi.org/10.1017/aaq.2018.53
9. Caple J, Byrd J, Stephan CN (2017) Elliptical Fourier Analysis: Fundamentals, Applications, and Value for Forensic Anthropology. Int J Leg Med 131: 1675–1690. https://doi.org/10.1007/s00414-017-1555-0
10. Caple J, Byrd J, Stephan CN (2018) The utility of elliptical Fourier analysis for estimating ancestry and sex from lateral skull photographs. Forensic Sci Int 289: 352–362. https://doi.org/10.1016/j.forsciint.2018.06.009
11. Cardillo M (2010). Some Applications of Geometric Morphometrics to Archaeology. In: Elewa AMT (ed.) Morphometrics for Nonmorphometricians. Springer, Berlin, Heidelberg. 325–341. https://doi.org/10.1007/978-3-540-95853-6_15
12. Chang J, Alfaro ME (2015) Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data, Methods Eco Evol 7(4): 472–482 https://doi.org/10.1111/2041-210X.12508
13. Chirikure S (2015) "Do as I Say and Not as I Do". On the Gap Between Good Ethics and Reality in African Archaeology. In: Haber A, Shepherd N (eds.) After Ethics. Ethical Archaeologies: The Politics of Social Justice, vol 3. Springer, New York. https://doi.org/10.1007/978-1-4939-1689-4_3.
14. Daboul A, Ivanovska T, Bülow R, Biffar R, Cardini A (2018) Procrustes-based geometric morphometrics on MRI images: An example of inter-operator bias in 3D landmarks and its impact on big datasets. PLoS One 13(5):e0197675. https://doi.org10.1371/journal.pone.0197675
15. Douglass K (2020). Amy ty lilin-draza'ay: Building archaeological practice on principles of community, Afr Archaeol Rev 37: 481:485. https://doi.org/10.1007/s10437-020-09404-8

16. Else, H. (2022) African researchers lead campaign for equity in global collaborations. Nature News. https://www.nature.com/articles/d41586-022-01604-3
17. Evin A, Bonhomme V, Claude J (2020) Optimizing digitalization effort in morphometrics, Biol Methods Protoc 5(1): bpaa023. https://doi.org/10.1093/biomethods/bpaa023
18. Fagertun J, Harder S, Rosengren A, Moeller C, Werge T, Paulsen RR, Hansen TF (2014) 3D facial landmarks: inter-operator variability of manual annotation. BMC Med Imaging 14:35 https://doi.org/10.1186/1471-2342-14-35
19. Fruciano C (2016) Measurement error in geometric morphometrics. Dev Genes Evol 226, 139–158 https://doi.org/10.1007/s00427-016-0537-4
20. Inizan ML, Reduron-Ballinger M, Roche H, Tixier J (1999) Technology and Terminology of Knapped Stone. Cercle de recherches et d'études préhistoriques.
21. Iovita R (2009) Ontogenetic scaling and lithic systematics: Method and application, J Archaeol Sci 36(7): 1447–1457. https://doi.org/10.1016/j.jas.2009.02.008
22. Iovita R (2011) Shape variation in Aterian Tanged Tools and the Origins of Projectile Technology: A Morphometric Perspective on Stone Tool Function. PLoS One 6(12): e29029. https://doi.org/10.1371/journal.pone.0029029
23. Ivanovaitė L, Serwatka K, Hoggard CS, Sauer F, Riede F (2020) All these fantastic cultures? Research history and regionalization in the Late Palaeolithic tanged point cultures of Eastern Europe. Eur J Archaeol 23(2): 162–185. https://doi.org/10.1017/eaa.2019.59
24. Key A, Pargeter J, Schmidt P (2021) Heat treatment significantly increases the sharpness of silcrete stone tools, Archaeometry 63(3): 447–466 https://doi.org/10.1111/arcm.12619
25. Klingenberg C (2008) Novelty and 'Homology-free' Morphometrics: What's in a Name? Evol Biol 35: 186–190. https://doi.org/10.1007/s11692-008-9029-4.
26. Koo T, Li M (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15. https://doi.org/10.1016/j.jcm.2016.02.012
27. Lycett SJ (2009) Quantifying transitions: Morphometric approaches to Palaeolithic variability and technological change. In: Camps M, Chauhan P (eds) Sourcebook of Paleolithic Transitions: Methods, Theories, and Interpretations, Springer New York, pp 9–92. https://doi.org/10.1007/978-0-387-76487-0_5.
28. Lycett SJ, von Cramon-Taubadel N (2015) Toward a "Quantitative Genetic" Approach to Lithic Variation, J Archaeol Method Theory 22, 646–675 https://doi.org/10.1007/s10816-013-9200-9
29. Lyman LL, VanPool TL (2009) Metric Data in Archaeology: A Study of Intra-Analyst and Inter-Analyst Variation, Am Antiq 74(3), 485–504. http://dx.doi.org/10.1017/S0002731600048721
30. MacDonald DA, Royal K, Buchanan B (2020) Evaluating the effects of parallax in archaeological geometric morphometric analyses. Archaeol Anthropol Sci 12, 149. https://doi.org/10.1007/s12520-020-01111-4
31. Marwick B, Guedes JA, Barton CM et al. (2017) Open Science in Archaeology. The SAA Archaeol Rec, http://dx.doi.org/10.17605/OSF.IO/3D6XX
32. Matzig, D.N. 2021. outlineR: An R package to derive outline shapes from (multiple) artefacts on JPEG images. Zenodo. https://doi.org/10.5281/ZENODO.4527469
33. Matzig DN, Hussain ST, Riede F (2021) Design Space Constraints and the Cultural Taxonomy of European Final Palaeolithic Large Tanged Points: A Comparison of Typological, Landmark-Based and Whole-Outline Geometric Morphometric Approaches. J Palaeolithic Archaeol 4(27). https://doi.org/10.1007/s41982-021-00097-2
34. McNabb J (2017) Journeys in space and time. Assessing the link between Acheulean Handaxes and genetic explanations, J Archaeol Sci Rep (13): 403.
35. Menéndez LP (2017) Comparing Methods to Assess Intraobserver Measurement Error of 3D Craniofacial Landmarks Using Geometric Morphometrics Through a Digitizer Arm, J Forensic Sci 62(3) https://doi.org/10.1111/1556-4029.13301

36. Mesfin I, Leplongeon A, Pleurdeau D, Borel A (2020) Using morphometrics to reappraise old collections: The study case of the Congo Basin Middle Stone Age bifacial industry. J Lithic Stud 7(1). https://doi.org/10.2218/jls.4329
37. Mitteroecker P (2021) Morphometrics in Evolutionary Developmental Biology. In: de la Rosa LN, Muller GB (eds) Evolutionary Development Biology. Springer, Cham, pp 941–951. http://doi.org/10.1007/978-3-319-32979-6_119.
38. Mitteroecker P, Bookstein F (2011). Linear Discrimination, Ordination, and the Visualization of Selection Gradients in Modern Morphometrics, Evol Biol 38(1):100–144. https://doi.org/10.1007/s11692-011-9109-8
39. Mullin SK, Taylor PJ (2002) The effects of parallax on geometric morphometric data. Computers in Biology and Medicine 32(6): 455–464. https://doi.org/10.1016/S0010-4825(02)00037-9
40. O'Leary MA, Kaufman S (2011) MorphoBank: phylophenomics in the 'cloud', Cladistics (27), 5, 529–537 https://doi.org/10.1111/j.1096-0031.2011.00355.x
41. Okumura M, Araujo AGM (2018) Archaeology, biology, and borrowing: A critical examination of geometric morphometrics in archaeology, J Archaeol Sci 101: 149–158. https://doi.org/10.1016/j.jas.2017.09.015
42. Osis S, Hettinga B, Macdonald S, Ferber R (2015) A novel method to evaluate error in anatomical marker placement using a modified generalized Procrustes analysis. Comput Methods Biomech Biomed Eng 18:1108–1116, https://doi.org/10.1080/10255842.2013.873034
43. Page, JW (1976) A note on interobserver error in multivariate analyses of populations. Am J Phys Anthropol 44(3), 521–525. https://doi.org/10.1002/ajpa.1330440315
44. Perini TA, de Oliveira GL, Ornellas JDS, de Oliveira FP (2005) Technical error of measurement in anthropmetry. Rev Bras Med Esporte 11(1), 86–90. https://doi.org/10.1590/S1517-86922005000100009
45. R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
46. Riaño HC, Jaramillo N, Dujardin J-P (2009) Growth changes in Rhodnius pallescens under simulated domestic and sylvatic conditions, Infect Genet Evol 9:162–168. https://doi.org/10.1016/j.meegid.2008.10.009
47. Robinson C, Terhune CE (2017) Error in geometric morphometric data collection: Combining data from multiple sources, Am J Phys Anthropol 164(1): 62–75. https://doi.org/10.1002/ajpa.23257
48. Roe DA (1964) The British Lower and Middle Paleolithic: Some Problems, Methods of Study and Preliminary Results. Proceedings of the Prehistoric Society.
49. Sahle Y (2021) Fossil men: The quest for the oldest skeleton and the origins of humankind Am J Phys Anthropol 176(2): 340–341. https://doi.org/10.1002/ajpa.24359
50. Sahle Y, Hutchings WK, Braun et al. (2013) Earliest stone-tipped projectiles from the Ethiopian Rift date to > 279,000 years ago, PloS One 8(11), e78092 https://doi.org/10.1371/journal.pone.0078092
51. Serwatka K, Riede F (2016). 2D geometric morphometric analysis casts doubt on the validity of large tanged points as cultural markers in the European Final Palaeolithic, J Archaeol Sci Rep 9: 150–159. https://doi.org/10.1016/j.jasrep.2016.07.018
52. Scerri EML, Kühnert D, Blinkhorn J et al. (2020) Field-based sciences must transform in response to COVID-19, Nat Ecol Evol 4: 1571–1574. https://doi.org/10.1038/s41559-020-01317-8
53. Schillinger K, Mesoudi A, Lycett S (2014) Copying Error and the Cultural Evolution of "Additive" vs. "Reductive" Material Traditions: An Experimental Assessment, Am Antiq 79(1), 128–143. https://doi.org/10.7183/0002-7316.79.1.128
54. Shea JJ (2020) Prehistoric Stone Tools of Eastern Africa: A Guide. Cambridge University Press.
55. Shott MJ & Trail BW (2010) Exploring new approaches to lithic analysis: Laser scanning and geometric morphometrics, Lithic Technol 35(2): 195–220. https://doi.org/10.1080/01977261.2010.11721090
56. Shrout PE, Fleiss JL (1979) Intraclass Correlation: Uses in Assessing Rater Reliability, Psychol Bull 86: 420–28. https://doi.org/10.1037//0033-2909.86.2.420

57. Timbrell L (2020) Strength in numbers: combining old datasets to answer new questions. In Kaercher K, Arntz M, Bomentre N, Hermoso Buxán XL, Day K, Ki S, Macleod R, Muñoz Mojado H, Timbrell L, Wisher I (eds) New Frontiers in Archaeology: Proceedings of the Cambridge Annual Student Archaeology Conference 2019, Archaeopress: Access Archaeology. ISBN 978-1-78969-794-0
58. Timbrell L (2022) A collaborative model for lithic shape digitization in museum settings AfricArXiv. osf.io/preprints/africarxiv/ba2f8.
59. Ulijaszek SJ, Kerr DA (1999). Anthropometric measurement error and the assessment of nutritional status. Br J Nutr 82, 165–177 https://doi.org/10.1017/s0007114599001348
60. von Cramon-Taubadel N, Frazier BC, Lahr MM (2007) The problem of assessing landmark error in geometric morphometrics: Theory, methods, and modifications, Am J Phys Anthropol 134, 24–35. https://doi.org/10.1002/ajpa.20616
61. Wang L-Y, Marwick B (2020) Standardization of ceramic shape: A case study of Iron Age pottery from northeastern Taiwan, J Archaeol Sci Rep 33: 102554. https://doi.org/10.1016/j.jasrep.2020.102554
62. Yezerinac SM, Lougheed SC, Handford P (1992) Measurement Error and Morphometric Studies: Statistical Power and Observer Experience, Systematic Biol 41(4): 471–482, https://doi.org/10.1093/sysbio/41.4.471
63. Zelditch ML, Swiderski DL, Sheets DH, Fink WL (2004) Geometric Morphometrics for Biologists: A Primer, Academic Press.
64. Zeng L, Zou X (2019) Error Analysis and Experimental Research on 3D Printing, IOP Conf Ser Mater Sci Eng 592, 012150.

Figures

Figure 1

The six 3D printed replica tools. Original lithics were knapped and scanned by CS in preparation for 3D printing. Example photos were taken by SH. Scale = 3cm.

Figure 2

Photographs from the 3D printing process. A) The 3D model of the tool is sent to the machine for printing. B) The resulting 3D prints once removed from the supports are cleaned using ethanol. 3D printing was carried out by LT and CS.

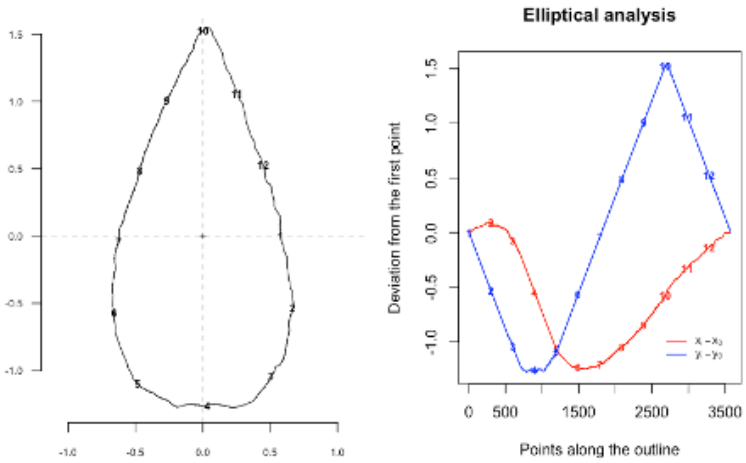

Figure 3

A schematic of the Elliptic Fourier fitting process that generates the raw shape data for geometric morphometrics. Coefficients of sine and cosine terms (harmonics) are computed to reconstruct the x (blue) and y (red) coordinates from an

arbitrary starting point moving along the outline.

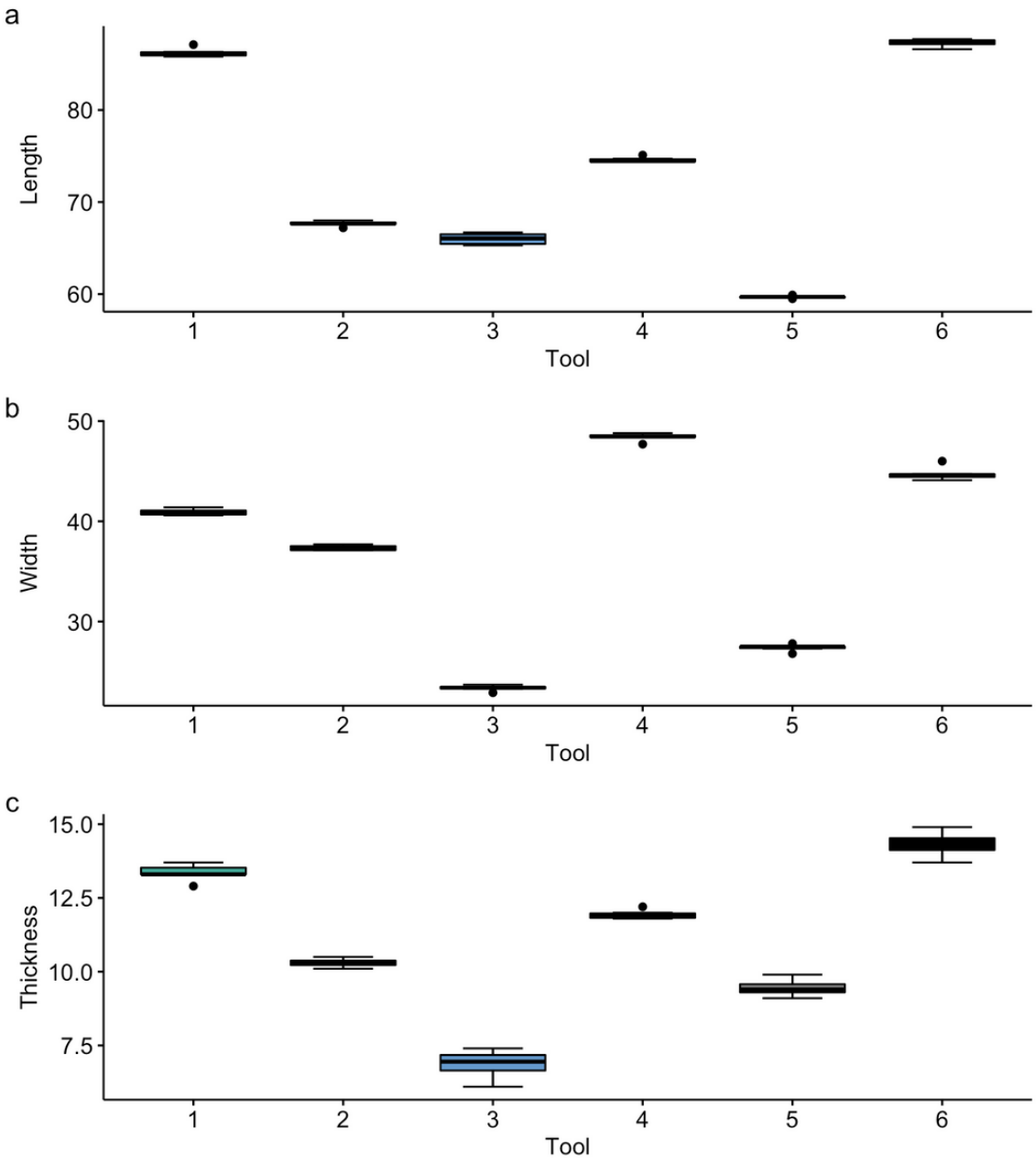

Figure 4

Boxplots demonstrating the distribution of length, width, and thickness (mm) collected by multiple observers for each tool (1-6).

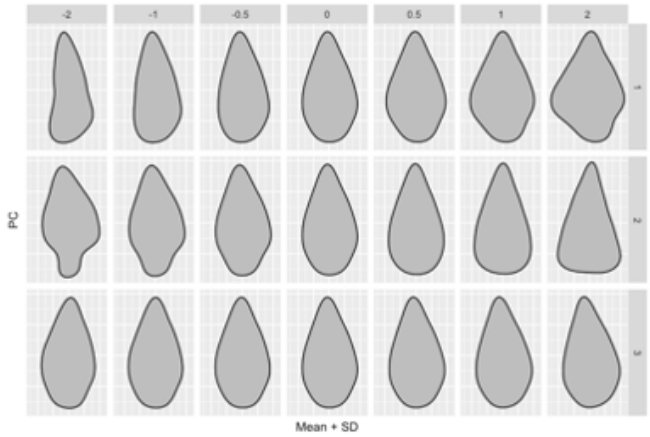

Figure 5

Principal component (PC) contributions along the first 3 axes of variance within the multiple observer outline data.

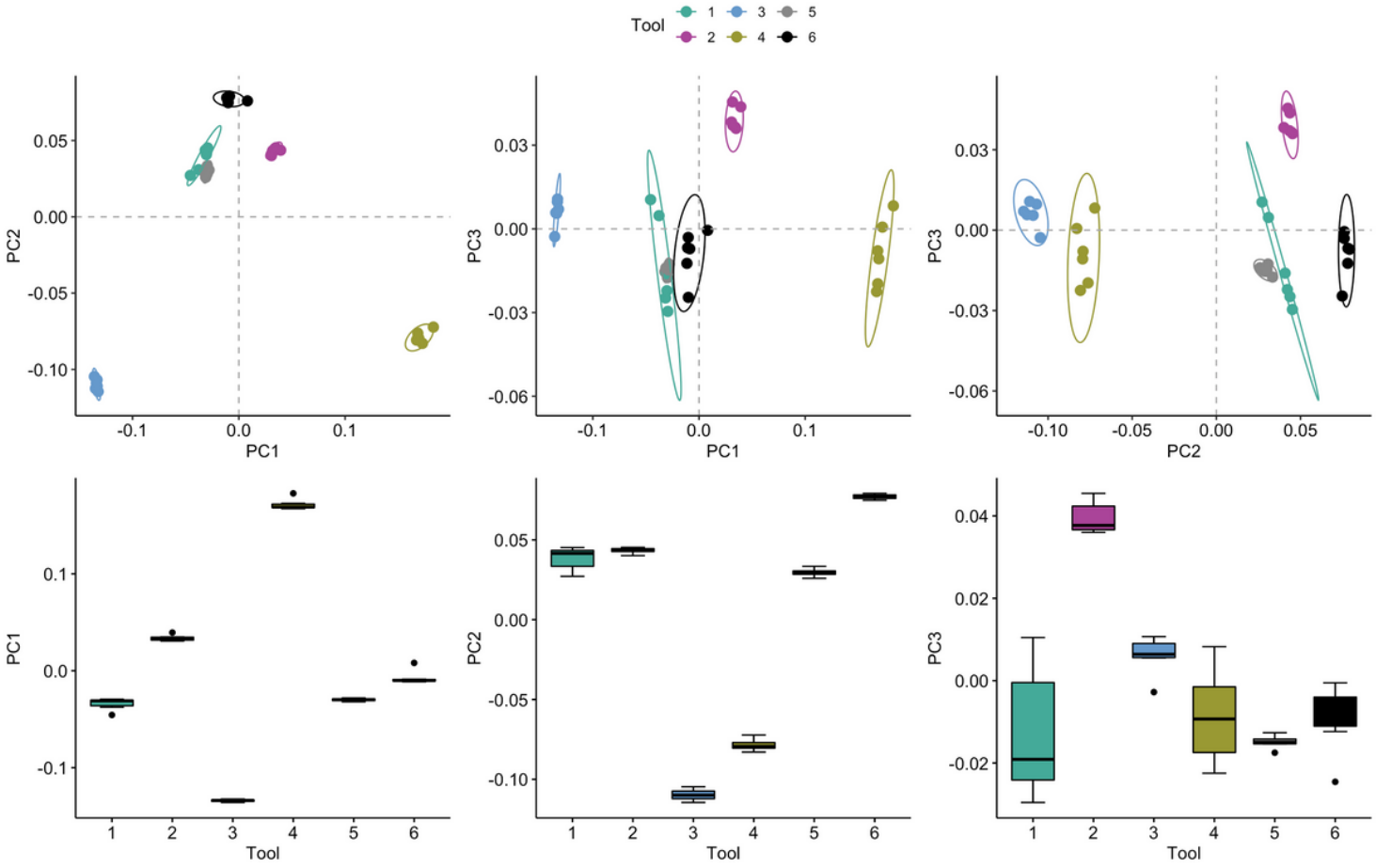

Figure 6

Scatterplots (top row) and boxplots (bottom row) of repeat capture scores along principal components (PC) 1-3, demonstrating the clustering within tools (1-6). PC1 represents 59.7% of the total variance, whilst PC2 and PC3 account for 33.4% and 3% respectively.

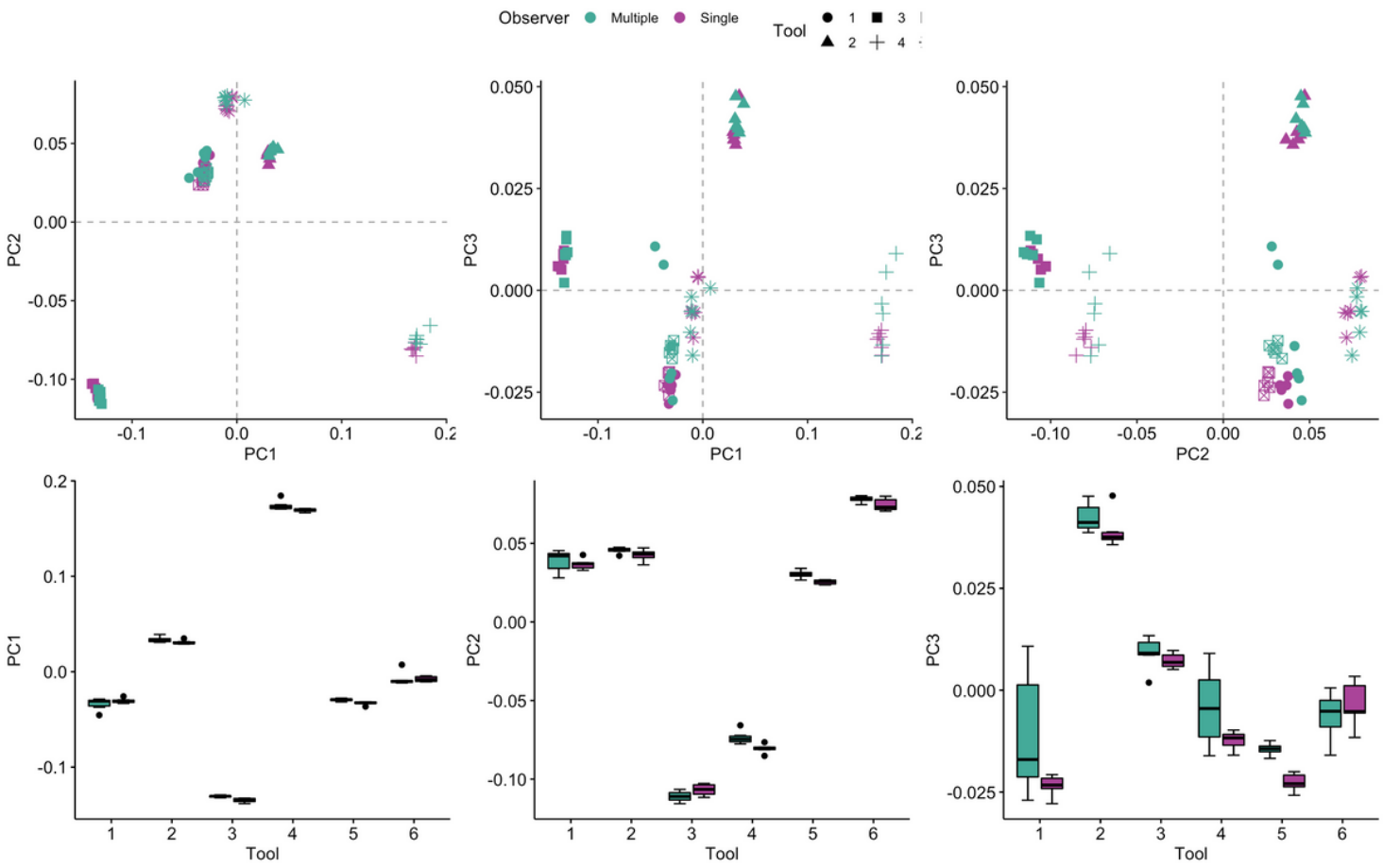

Figure 7

Scatterplots (top row) and boxplots (bottom row) of repeat capture scores along principal components (PC) 1-3, demonstrating the clustering within tools (symbols) and between data sets (colors). PC1 represents 60.4 % of the total variance, whilst PC2 and PC3 account for 33.5% and 3.3% respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- draft2somerroranalysis.docx