American Antiquity (2023), **88**, 163–186
doi:10.1017/aaq.2023.4

ARTICLE

Replicability in Lithic Analysis

Justin Pargeter[1], Alison Brooks[2], Katja Douze[3], Metin Eren[4], Huw S. Groucutt[5], Jessica McNeil[6], Alex Mackay[7], Kathryn Ranhorn[8], Eleanor Scerri[9], Matthew Shaw[10], Christian Tryon[11], Manuel Will[12], and Alice Leplongeon[13]

[1]Department of Anthropology, New York University, NY, USA; Palaeo-Research Institute, University of Johannesburg, Johannesburg, South Africa, [2]Center for the Advanced Study of Human Paleobiology, Department of Anthropology, George Washington University, Washington, DC, USA; Human Origins Program, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA, [3]Laboratory Archaeology and Population in Africa, Section of Biology, Faculty of Science, University of Geneva, Geneva, Switzerland, [4]Department of Anthropology, Kent State University, Kent, OH, USA; Department of Archaeology, Cleveland Museum of Natural History, Cleveland, OH, USA, [5]Department of Classics and Archaeology, University of Malta, Msida, Malta; Extreme Events Research Group, Max Planck Institutes for the Science of Human History, Chemical Ecology, and Biogeochemistry, Jena, Germany, [6]Department of Anthropology, Harvard University, Cambridge, MA, USA, [7]Center for Archaeological Science, University of Wollongong, Wollongong, Australia; Department of Archaeology, University of Cape Town, Cape Town, South Africa, [8]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA; Institute of Human Origins, Arizona State University, Tempe, AZ, USA, [9]Pan-African Evolution Research Group, Max Planck Institute for the Science of Human History, Jena, Germany; Department of Prehistoric Archaeology, University of Cologne, Cologne, Germany, [10]Center for Archaeological Science, University of Wollongong, Wollongong, Australia, [11]Department of Anthropology, University of Connecticut, Storrs, CT, USA; Department of Anthropology, Harvard University, Cambridge, MA, USA; Human Origins Program, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA, [12]Department of Early Prehistory and Quaternary Ecology, University of Tübingen, Tübingen, Germany, and [13]Department of Archaeology, KU Leuven, Leuven, Belgium; UMR Histoire naturelle de l'Homme Préhistorique, Muséum national d'Histoire naturelle – Centre National de la Recherche Scientifique – Université de Perpignan Via Domitia, Paris, France
Corresponding author: Justin Pargeter, Email: justin.pargeter@nyu.edu

(Received 13 July 2022; revised 16 December 2022; accepted 9 January 2023)

Abstract

The ubiquity and durability of lithic artifacts inform archaeologists about important dimensions of human behavioral variability. Despite their importance, lithic artifacts can be problematic to study because lithic analysts differ widely in their theoretical approaches and the data they collect. The extent to which differences in lithic data relate to prehistoric behavioral variability or differences between archaeologists today remains incompletely known. We address this issue with the most extensive lithic replicability study yet, involving 11 analysts, 100 unmodified flakes, and 38 ratio, discrete, and nominal attributes. We use mixture models to show strong inter-analyst replicability scores on several attributes, making them well suited to comparative lithic analyses. Based on our results, we highlight 17 attributes that we consider reliable for compiling datasets collected by different individuals for comparative studies. Demonstrating this replicability is a crucial first step in tackling more general problems of data comparability in lithic analysis and lithic analyst's ability to conduct large-scale meta-analyses.

Resumen

La ubicuidad y la durabilidad de los artefactos líticos le da a los arqueólogos datos importantes sobre las dimensiones de la variabilidad del comportamiento humano. A pesar de su importancia, los artefactos líticos pueden ser problemáticos de estudiar ya que los especialistas en lítica difieren ampliamente en sus enfoques teóricos y en los datos que recogen. Si las diferencias en los datos líticos reflejan la variabilidad en el comportamiento prehistórico, o por el contrario están ligadas a las diferencias entre los arqueólogos que los estudian hoy es una cuestión aún parcialmente desconocida. Abordamos esta problemática con el estudio de replicabilidad lítica más amplio realizado hasta la fecha, incluyendo 11 especialistas, 100 lascas y 38 atributos continuos, discretos y nominales. Usando modelos de mezcla presentamos altos resultados

de replicabilidad entre los especialistas participantes sobre algunos atributos, lo que los hace adecuados para análisis líticos comparativos. Basados en nuestros resultados evidenciamos 17 atributos que consideramos fiables cuando recopilamos conjuntos de datos recogidos por diferentes individuos para análisis comparativos. Demostrar esta replicabilidad es un primer paso crucial para abordar problemas más generales de comparabilidad de datos en análisis líticos y la posibilidad de conducir meta-análisis a gran escala combinando múltiples conjuntos de datos.

Keywords: stone tools; attribute analysis; inter-analyst replicability

Palabras clave: herramientas líticas; análisis de atributos; replicabilidad entre especialistas en lítica

Comparative research greatly benefits Paleolithic archaeologists because of the field's vast temporal and spatial remit. However, comparative research requires datasets in which researchers know that the individual data points are robust, repeatable, and comparable; that is, data are collected similarly and compared against similar standards. Notably, such studies must recognize the error and uncertainty involved in specific types of data and the way they are collected. While archaeologists working with large datasets of published radiometric dates acknowledge this issue (Carleton and Groucutt 2021; Mauz et al. 2021; Scott et al. 2018; Stewart et al. 2021), lithic analysts have yet to deal with these issues systematically.

Stone tools are durable and ubiquitous, and they tend to pattern in space and time. Comparative lithic analysis therefore remains, and must remain, a cornerstone for understanding human behavioral evolution and the evolution of technology. Yet, to achieve this potential, stone tool analysts need to know the comparability of their units of analysis. Data incongruence is a significant problem given that stone tools make up most or all the archaeological record for a span of approximately 3–2 million years, particularly in Africa (Harmand et al. 2015; Shea 2016).

In lithic studies, researchers from different backgrounds often take different conceptual approaches to their analyses. They may record different kinds of data entirely, or even similar types of essential information in different ways (e.g., Andrefsky 2005; Holdaway and Stern 2004; Shea 2013; Van Peer 1992). For example, proponents of the *chaîne opératoire* approach generally focus on qualitative classification of tool production systems, whereas analysts in the Americanist attribute-based system often pursue quantitative measures of reduction intensity. This difference is true even for seemingly simple attributes such as "length," for which multiple definitions exist. Andrefsky (2005:100), for example, shows how analysts can measure flake length in at least two different ways: as a line perpendicular to the striking platform width or as the maximum distance from the proximal to the distal end along a line perpendicular to the striking platform width. Dogandžić et alia (2015) show that calculations of flake edge length and surface areas based on datasets where analysts recorded length using different methods are prone to large variance and errors. If variance exists in measuring even these basic lithic attributes, it is obvious that problems will arise when constructing and comparing lithics with large datasets generated by multiple analysts and analytical approaches.

Comparative lithic analysis aims to achieve high consistency and low error rates when recording and measuring attributes on lithic artifacts between observers. Increasing such inter-analyst replicability is a goal common to all empirical sciences. A lack of clear standards for assessing data quality and replicability has led to the recent "reproducibility crisis" (Baker 2016). Even though the importance of analytical standardization is undisputed, there are surprisingly few studies explicitly tackling analyst-induced variance in lithic technological analyses (Conard et al. 2004; Tixier 1963). Exceptions exist in the field of lithic use-wear (Crowther and Haslam 2007; Newcomer et al. 1986; Rots et al. 2006; Wadley et al. 2004), but these studies focus on tool use rather than lithic production strategies.

Researchers followed classic inter-analyst replicability studies by Fish (1978), Wilmsen and Robert (1978), and Dibble and Bernard (1980), with a limited number of more recent assessments for *quantifying* the effect of different observers on lithic data quality. Table 1 summarizes the most relevant studies for lithic analyses focused on assessing the extent, source, and relevance of inter-analyst

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

Table 1. Summary of Previous Lithic Inter-Analyst Replicability Studies. All numerical values are counts.

Study	Observers	Lithics	Attributes	Attributes	Attribute Classes
Fish 1978	3	25	6	Maximum length, technological length, width, thickness, striking platform angle, dorsal cortex	Ratio
Wilmsen and Robert 1978	4	4	8	Length, width, thickness, platform thickness, flake angle, cutting edge angle, distal edge angle, left lateral edge angle, right lateral edge angle	Ratio
Dibble and Bernard 1980	6	29	1	Edge angle (comparison of 4 different methods to measure edge angle)	Ratio
Perpére 1986	3	198	1	Levallois (Y/N)	Nominal
Boyd 1987	3	246	3	Working edge type, technological class, raw material type	Nominal
Calogero 1992	5	17	1	Raw material type	Nominal
Gnaden and Holdaway 2000	15	211	9	Length, width, thickness, platform thickness, platform width, termination type, cortex, platform type, artifact form	Ratio and nominal
Lycett et al. 2006	2	3	3	Length, width, thickness (all at 10% increments)	Ratio
Mackay 2008	7	58	1	Blade (Y/N)	Nominal
Driscoll 2011	47	20	5	Flake type, core type, debitage type, fragmentation, retouch type	Nominal
Proffitt and de la Torre 2014	4	765	9	Technological category, platform cortex, platform facets, knapping accidents, step scars, dorsal surface cortex, number of dorsal negatives, direction of dorsal negatives, Toth flake category	Ratio and nominal
Timbrell et al. 2022	7	6	5	Maximum dimension, width, thickness, shape (22 GMM)	Ratio
Pargeter et al. (this study)	11	100	38	This article Supplementary Table 2	Ratio, discrete, and nominal

Note: All numerical values are counts.

replicability. Principally, all studies in Table 1 found that replicability between analysts is an issue—larger than generally anticipated and with important ramifications for subsequent interpretations. Problems included measuring basic dimensions such as flake length, seemingly straightforward assessments such as counting the number of flake dorsal negative scars, and more complex inferences such as identifying whether a flake belongs to a specific technological system. Previous studies have a generally low number of examined attributes (median = 4; range = 1–9), a low number of observers (median = 5; range = 2–47, mostly from a close group of coworkers and students), and small lithic samples (Table 1). For the sake of this article, we exclude inter-analyst use-wear studies. As a result, we still lack the following:

- Quantitative measures for inter-analyst variability in lithic studies
- Tests to better understand the causes of inter-analyst differences
- Recommendations for fixing issues in inter-analyst replicability

We assembled the "Comparative Analyses of Middle Stone Age Artifacts in Africa" (CoMSAfrica; Will et al. 2019) group to address these issues with specific reference to the African Middle Stone Age (MSA), but the group immediately recognized that the problems of lithic inter-analyst replicability extend well beyond any time or place. In this article, we report on the group's first of three inter-analyst replicability studies, with this one focused on unretouched lithic flakes. We present data showing possible reasons for poor inter-analyst reliability on some of our attributes, and we suggest ways to improve the replicability of future lithic attribute analyses.

The CoMSAfrica Project

The CoMSAfrica project started as a three-day workshop at Harvard University (USA) in 2018 (Will et al. 2019). The workshop brought together 12 international lithic analysts (see author list) from seven countries working in different periods and regions of Africa, with varied methodological backgrounds (e.g., *chaîne opératoire* and attribute analysis) and levels of seniority (full professor to PhD student). The group aimed to compare African MSA lithic assemblages at the initial workshop. The project's long-term objective is to use African MSA lithic assemblages in comparative continental-scale studies to unpack spatial and temporal variation among Pleistocene *H. sapiens* populations.

Intense discussions at the initial meeting in 2018 made it clear that our initial goals were too ambitious and that any continental-scale comparisons were impossible until we understood differences in how group members recorded their lithic data. In 2018, we established a minimum number of attributes that each group member currently used or considered useful for reliable comparative analysis. We initially focused on unretouched flakes because they form the dominant category of all lithic assemblages and they carry important information about lithic production methods, techniques, and reduction intensity. This study forms the basis for working toward other future studies involving cores and retouched tools.

To maximize replicability, the group derived a set of definitions for each attribute from existing literature and lithic recording systems used by the project's members and others (e.g., Scerri et al. 2014; Shea 2013; Tostevin 2012; Wilkins et al. 2017). Again, although the group focuses on African MSA lithic assemblages, our current protocols relate to issues faced by lithic analysts working at almost any time or location.

In this study, we address the following seven research questions that arose from the process of data exploration:

(1) Which lithic attributes are analysts able to code more reliably, and which are they able to code less reliably?
(2) Does limiting the number of possible attribute states impact inter-analyst replicability?
(3) Do specific flake characteristics (i.e., differences in flake shape, etc.) impact inter-analyst replicability?

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

(4) Does the inclusion of images in definitions impact inter-analyst replicability?
(5) What degree of measurement precision is realistic in lithic analysis?
(6) Do differences in lithic flaking systems impact inter-analyst replicability?
(7) Do the analyst's experience and training impact inter-analyst replicability?

Methods and Materials

This study included lithic analysts from diverse research traditions and with varying levels of experience. We intentionally included multiple backgrounds to provide a wide range of inputs into the recording system and to avoid the false positives that might result when related analysts confront diverse sets of attributes. The study used alpha-numeric analyst IDs to ensure each analyst's anonymity. We asked analysts to list their years of experience conducting lithic analysis and to rate their training using *chaîne opératoire* and quantitative methods on a scale from 1 to 5 (1 = lowest, 5 = highest). The group averaged 18.5 years of experience (median = 13, range = 9–58), 3.2 on the self-reported quantitative training scale (median = 3, range = 1–5), and 3.5 on the self-reported *chaîne opératoire* training scale (median = 3.5, range = 2–5; see Supplemental Table 1).

After presenting their existing recording systems, the participants selected a common subset of attributes for this study. In subsequent discussions, all participants jointly agreed on a definition for each of the study's attributes. This approach ensured that the recording system represented the group's training backgrounds and experience levels. All attributes had to satisfy one common criterion: participants accepted them as potentially useful for studying African MSA human behavioral variability. The group selected 38 attributes for this study (see Supplemental Table 1).

The Flake Attributes

We divide our attributes into three broad classes:

(1) Ratio-scale attributes ($n = 17$): analysts recorded measurements on the flakes (e.g., flake length, width, thickness, and mass). Ratio-scales refer to data with a true zero and equal intervals between neighboring points.
(2) Discrete-scale attributes ($n = 5$): analysts counted attribute expressions as whole numbers (e.g., dorsal scar counts).
(3) Nominal attributes ($n = 16$): analysts selected options from a predefined list of descriptive characteristics (e.g., platform types). Analysts use nominal scales to label attributes with no quantitative value.

During data cleaning, we edited certain text inputs to lump answers with slight variations that otherwise referred to the same technological system (e.g., "Levallois variant" and "Levallois"). Following conventions in lithic analysis, we refer to subdivisions within each attribute as "states" (also called "expressions" or "levels"; Andrefsky 2005:65; Holdaway and Stern 2004:98). Each nominal attribute had between three and 10 attribute states. However, we left two ("Flake type" and "Reduction system") as free text and open to the analysts' unconstrained input, although in both cases, we included a range of suggested attribute states—16 for "Flake type" and 9 for "Reduction system" (see Supplemental Table 2). Besides these two examples, we designed the attribute states to be exhaustive and typically provided a range of prescribed states and one "other" state. For example, the attribute "Distal plan form" includes the states "Flat," "Pointed," "Rounded," and "Irregular," the last of which captures all nonconforming shapes (see Supplemental Table 2).

Before the analysis, the group agreed on a textual definition for all attributes, with instructions for their measurement. We added pictures to the definitions in some cases, particularly for size measurements. Wherever possible, we sourced attribute definitions and pictures from publications (see Supplemental Table 2). We typically did not define attribute states. The lack of a textual or "logical" description for attribute states was not a conscious part of our research design but reflected common practice in lithic artifact research—one that we do not recommend for future studies (see below).

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

We compiled the study's attributes into two main data recording systems. The first uses the open-source E4 data entry program by Shannon McPherron and Harold Dibble (www.oldstoneage.com). We used E4 to speed up data entry and reduce data entry errors. The E4 program creates conditional statements that allow certain variables to be skipped based on values entered for previous variables. The benefits of E4 over other data logging methods, such as Microsoft Excel, are that E4 prevents users from directly accessing a project database when entering data, and it prevents users from manually entering text inputs, which can lead to transcription errors. The program helps reduce data entry errors and increases data entry consistency. Several analysts were, however, more comfortable using Microsoft Excel (with a series of predefined columns and drop-down menus). This variance in the data recording method was a poor design choice that resulted in substantial time dedicated to data cleaning (see R code for details: https://osf.io/seh2t/?view_only=9097 ef58225b49e48f66afb220022fbf).

The Flake Assemblage

The flake assemblage comprised one raw material—chert—because its physical properties are analogous to many finer-grained raw materials found in African MSA and other lithic assemblages. Chert is also relatively fine grained, is homogenous, and fractures relatively reliably. This choice of raw material meant that the group worked with a raw material likely to produce a high proportion of flakes with "readable" technological characteristics.

One person (M. I. Eren) knapped all the flakes with a hard stone hammer and a direct freehand percussion technique. He used two continuous individual reduction strategies: recurrent unidirectional Levallois and a migrating multiplatform strategy in which he gave no platform surface preference. Admittedly, this is a limited framework, but with the study's otherwise complex recording methods, we decided to simplify the technological comparisons. These two reduction strategies cover a large amount of variance in African MSA lithic assemblages (Shea 2020) and occur in other periods and geographical areas. They also allowed us to test the attribute system on two different, but representative, flaking variants. During the study, analysts were unaware of these assemblage differences.

Eren reduced two cores until he had produced 100 flakes from each reduction strategy, and from these 100, we used a random number generator to select 50 flakes. The flakes, bagged separately, were boxed for shipping to each of the study's 11 participants. The team shared and shipped a set of digital calipers for flake measurements and used their own scales for mass measurements. Analysts examined the assemblages independently, without fixed protocols for lighting or the use of magnifying lenses, among other things (cf. best practices listed in Blumenschine et al. 1996). The participants did not know which flake assemblage corresponded to which knapping strategy. Analysts did not discuss observations until everyone had studied all the flakes, which took about two years.

Statistical Methods

Our primary research question is this: Which attributes are analysts able to code more reliably, and which are they able to code less reliably? To answer this question, we used replicability coefficients. Replicability describes the relative partitioning of variance in a measurement or other assessment into within-group and between-group sources of variance. Researchers generally refer to this measure as inter-rater repeatability (IRR; Hallgren 2012; Stoffel et al. 2017). We use inter-analyst replicability in this article.

We used a mixed effects model framework to estimate IRR and its uncertainty on the study's attributes using the rpt function in R version 4.0.3's rptR package (R Core Team 2021; Stoffel et al. 2017). Where analysts take repeated measures (e.g., quantifications of flake attributes) on the same objects (i.e., stone flakes), replicability estimation is calculated as the variance among group means (in our case, each flake measured 11 times) relative to the sum of group-level and data-level (individual measurements) variance. We included each analyst's anonymous ID and the two technological assemblage codes as random effect components to estimate the replicability at the level of each flake and across the two flaking systems. Higher replicability values show greater agreement between different analysts (1 = perfect agreement, 0 = no agreement). We modeled ratio data (i.e., flake maximum length) as approximating a normal distribution using rpt's Gaussian parameter. We modeled discrete attributes

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

(i.e., flake scars) using rpt's Poisson parameter. Dorsal cortex is a proportion (scored from 0 to 1) for which the rpt function does not yet have an inbuilt error function. Therefore, we omitted the dorsal cortex attribute from the IRR analysis, but we discuss qualitative observations on this attribute in this article. We also used standard deviation as a percentage of the mean for each ratio-scale attribute on each flake to track absolute error in our measurements. This measurement allowed us to determine if there are differences in relative (IRR) versus absolute (standard deviation as a percentage of the mean or the coefficient of variation [CV]) inter-analyst errors on the measurements.

For our nominal data, we used the first-order agreement coefficient (AC_1), where analysts classified flake attributes into one category among a limited number of possible categories (Gwet 2008). The AC_1 coefficient accounts for chance agreement between analysts in the presence of high agreement and can handle inputs from multiple raters. We implemented the analysis using the gwet.ac1.raw function in R's irrCAC package. We omitted instances where analysts either did not rate a specific flake or, for whatever reason, fewer than four analysts classified a flake. For some attributes, such as the "Reduction system," we ended up with 31 flakes for this attribute.

What constitutes a robust inter-analyst replicability estimate will depend on the nature of the study. Cohen (1960) provides a general guide that we use to interpret this study's IRR values: values ≤0 indicate no agreement, 0.01–0.20 indicate none to slight, 0.21–0.40 indicate fair, 0.41–0.60 indicate moderate, 0.61–0.80 indicate substantial, and 0.81–1.00 indicate strong agreement.

Our study also involved building several linear models to determine, for example, the impact of an analyst's prior experience on flake measurement performance. We built these models using R's base lm package, fitting different error functions (i.e., Gaussian and Poisson) to account for different response variable data scales or with two-way Analysis of Variance (ANOVA) using R's base aov package.

To evaluate potential causes for inter-analyst variance and whether some flakes led to a higher inter-analyst variance, we identified flake outliers using the Interquartile Range (IQR) for each ratio and discrete attribute. Here, a value is considered an outlier when it falls above the seventy-fifth or below the twenty-fifth percentile by a factor of 1.5 times the IQR. Because we were only interested in flakes with higher inter-analyst variance, we only considered outliers falling above the seventy-fifth percentile.

Results

Supplemental Tables 3–21 provide detailed summary data for each of the 38 attributes tested in this study. Supplemental Tables 3 and 4 document average results by analyst and flake for the ratio, discrete, and nominal attributes, whereas Supplemental Tables 5–21 provide summaries of the discrete attributes by flake. Here, we limit the results to our primary research questions.

Are Some Attributes More Replicable Than Others?

Figure 1 shows the IRR results for the study's 17 ratio-scale attributes with 95% confidence intervals (see Supplemental Table 3). Ten of them show strong inter-analyst measurement agreement between the analysts. Seven measurements show less, but still substantial, agreement (IRR >0.6 and <0.8) between the 11 analysts. These seven attributes include the four platform measurements (width and three platform thickness measurement variants) and three technologically oriented size measurements (thickness at the proximal end, thickness at the distal end, and width at the distal end). As expected, flake mass showed the highest IRR values, with maximum flake dimension and technological length showing very high IRR scores.

Figure 2 presents the CV values for each ratio-scale attribute on each flake. Our measurement CV values show very low effective variance in the measurements (mean = 0.09, range = 0.009–0.18). The ordering of attributes along this measure follows approximately the same pattern seen in the IRR data (Figure 1). The results show that (a) our relative measure of error (IRR) tracks our absolute measure of error (CV), and (b) that simple measures such as CVs can trace some of the variance present in our more complex IRR calculations. This result also reaffirms the overall strong performance of our ratio-scale attributes.

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

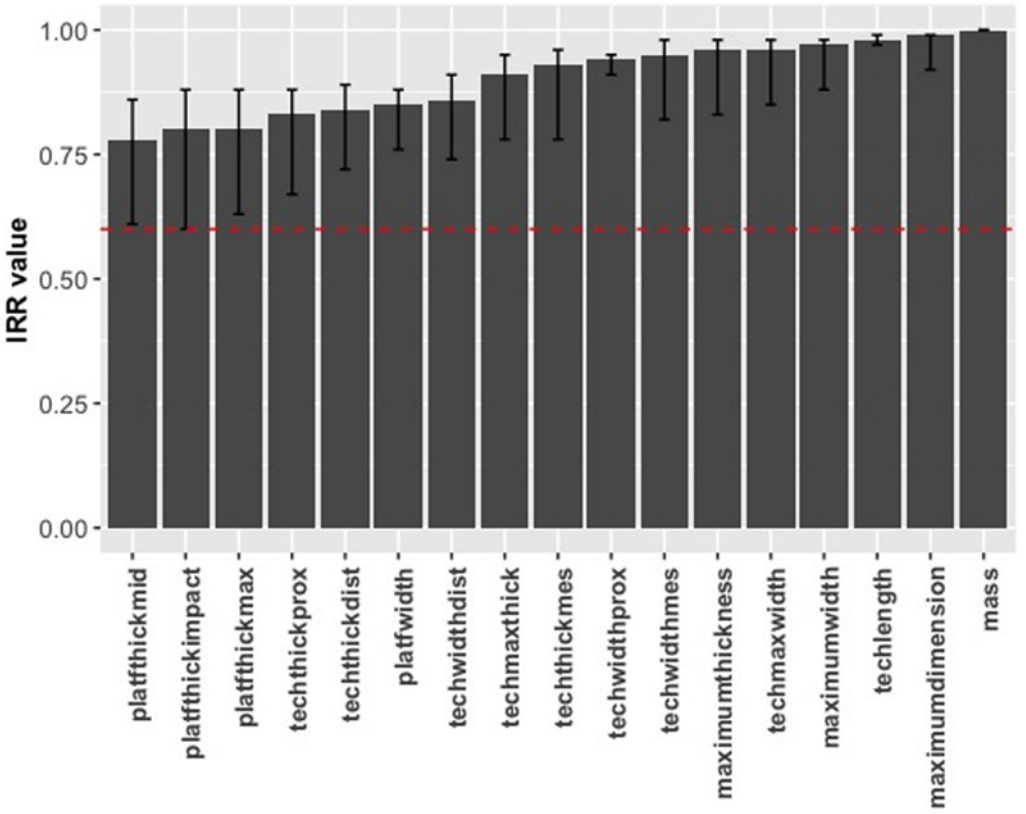

Figure 1. Summary IRR data for the study's 17 ratio-scale attributes. The dashed line indicates the cutoff for substantial agreement among raters. All measurements fall well above the substantial agreement threshold. Error bars show 95% confidence intervals.

Figure 3 shows the IRR results for our five discrete-scale attributes tracking flake dorsal scar characteristics. Three of these attributes (left sector scars, distal sector scars, and right sector scars) fall below the substantial agreement threshold. Proximal sector scar counts show an IRR value above the substantial agreement threshold. Total dorsal scar counts showed the highest IRR value within the substantial agreement threshold.

Figure 4 shows the overall IRR results for the study's 16 nominal attributes (see Supplemental Table 4). Five of these attributes show IRR values within the strong agreement range. The top-performing nominal attribute tracks analysts' ability to identify basic flake fracture mechanics features (i.e., bending, wedging, or Hertzian initiations), with the "free-text" input attributes ("Reduction system") also performing very well. Four nominal attributes show IRR values within the substantial agreement threshold (flake termination, form, completeness, and platform lipping). Seven of these attributes show IRR values below the substantial agreement threshold. Five lower-performing attributes relate to flake shape characteristics (ventral plan form, distal plan form, lateral edge shape, cross-section shape, and platform morphology).

Does Limiting the Number of Potential Attribute States Impact Inter-Analyst Replicability?

Having observed the study's wide-ranging (and generally lower) performance for inter-analyst replicability among our nominal data (Figure 4), we asked whether each attribute's number of states among which analysts could choose impacted some of this variability (see Supplemental Table 23).

A generalized linear model with a Poisson error parameter to account for the response variable's (attribute state counts) discrete scale shows a significant effect of attribute state counts on inter-analyst

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

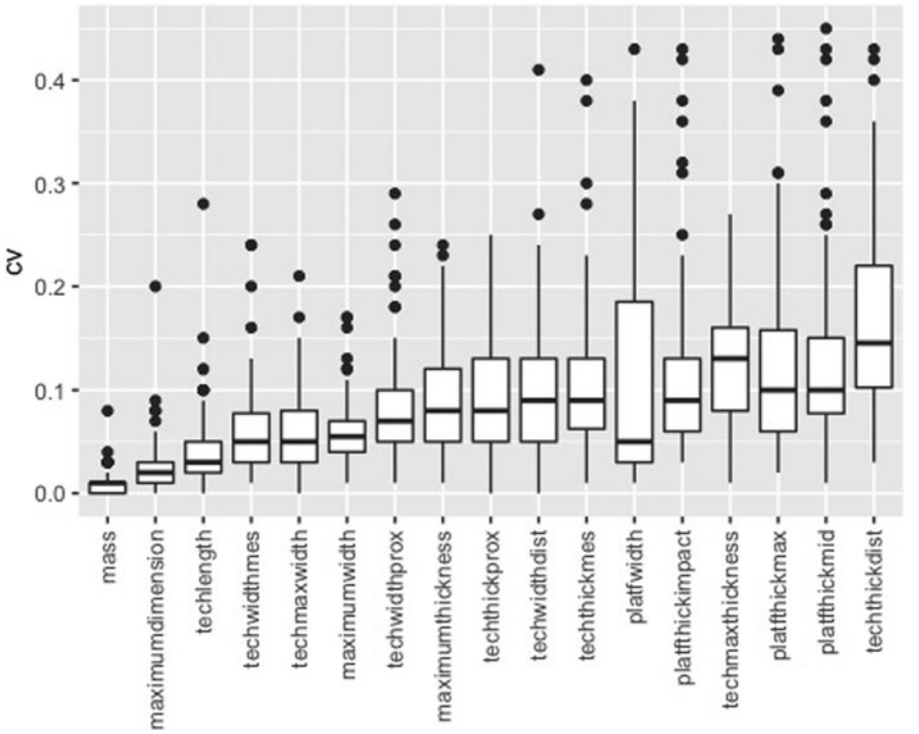

Figure 2. Summary showing the CV for each ratio-scale attribute on each flake. Outlier values with CV >0.5 are excluded from this plot.

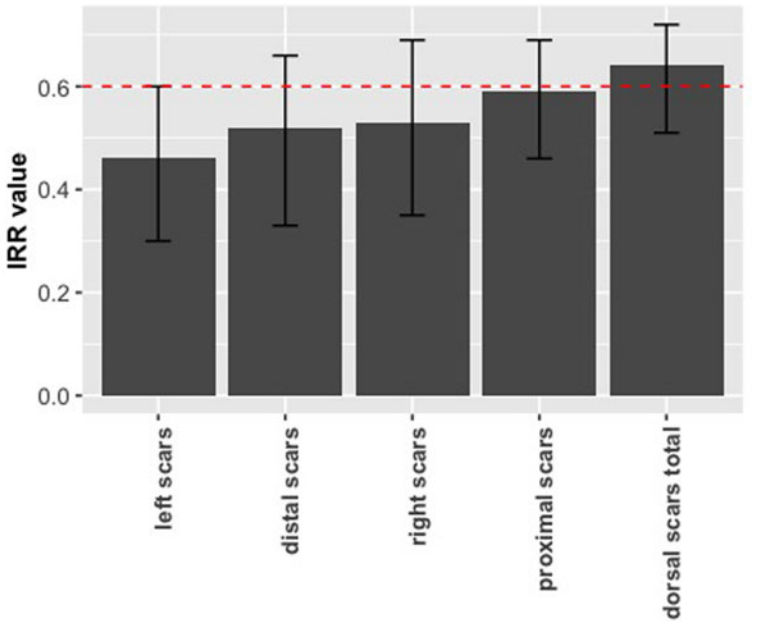

Figure 3. Summary IRR data for the study's discrete-scale attributes. The dashed line indicates the cutoff for substantial agreement among raters. Error bars show 95% confidence intervals.

replicability ($df = 15$, $p = 0.01$). Nominal attributes with more states tend to show lower inter-analyst replicability scores, whereas attributes with fewer states perform better. Three attributes (ventral plan form, distal plan form, and flake completeness) are notable outliers. Ventral and distal plan

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

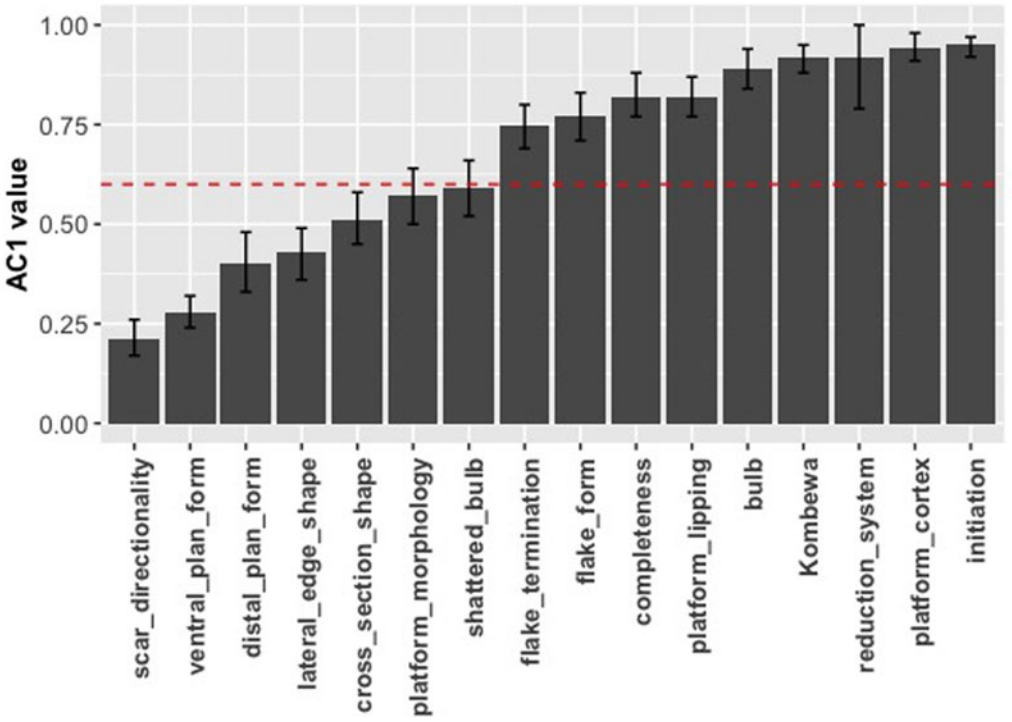

Figure 4. Summary AC1 data for attributes. The dashed line indicates the cutoff for substantial agreement among raters. Error bars show 95% confidence intervals.

forms have fewer states ($n = 5$ and $n = 4$), but analysts struggled to agree on their coding (see Supplemental Table 22). This result is likely because they require analysts to make complex decisions about flake shape and specific flake locations. Flake completeness has more states ($n = 8$), but analysts had fewer issues coding it. This is likely because the flakes in our assemblages had fewer breakages than the average archaeological assemblage.

Do Specific Flake Characteristics Impact Inter-Analyst Replicability?

One of the more complex issues our study faced is how different attributes might interact. For example, flake form and technological characteristics have the potential to impact the way analysts record different measurements. Flakes with more complex platform shapes or lateral edge types could complicate where analysts take specific measurements. This variance could impact inter-analyst replicability and increase "systematic errors" (i.e., errors that affect the central tendency of a size measurement [Gnaden and Holdaway 2000]). To examine this question more closely, we conducted ANOVA analyses with Bonferroni corrected post hoc comparisons for all our attributes against respective IRR values for attributes measured on those flake portions. For example, we compared the IRR values for platform measurements taken on different platform types.

Table 2 presents a subset of these results focused on attribute states that show statistically significant IRR results for each attribute. The data show that platform types, the presence/absence of platform cortex, flake lateral edge types, ventral plan form, and flake termination differences significantly impact measurements taken on these flake components. This result is particularly, but not exclusively, applicable to measurements of "technological" versus "maximum" dimensions. For example, lateral edge type shows significant differences in inter-analyst replicability in technological width measurements at the proximal and medial flake portions. Platform thickness measurements are complicated by different platform morphologies—especially those classed as "indeterminate"—and by the presence/absence of platform cortex, likely caused by the gradation of the cortex into regions less clearly part

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

Table 2. Summary ANOVA Results Comparing Instances Where Categorical Attribute's IRR Values Differed Significantly Based on Comparisons with Specific Attribute States.

Attribute	Comparison	Measurement	Est.	Lower	Upper	Adj. *P*-Value
Platform morphology	Indeterminate-Dihedral	Platform thickness midpoint	2.36	0.00	4.72	0.04
	Plain-Indeterminate	Platform thickness midpoint	−2.09	−4.08	−0.10	0.03
	Punctiform-Indeterminate	Platform thickness midpoint	−2.84	−5.68	−0.01	0.04
Platform cortex	Complete-Absent	Platform width	7.03	2.44	11.61	<0.01
	Partial-Absent	Platform thickness midpoint	2.58	1.02	4.15	<0.01
	Partial-Absent	Platform thickness maximum	2.10	0.51	3.68	<0.01
Lateral edge type	Parallel-Amorphous	Technological width medial	−2.98	−5.25	−0.71	<0.01
	Parallel-Divergent	Technological width medial	−2.28	−4.17	−0.39	<0.01
	Parallel-Amorphous	Technological width proximal	−2.64	−5.21	−0.07	0.04
	Ovoid-Convergent	Technological width proximal	2.96	0.07	5.85	0.04
	Parallel-Ovoid	Technological width proximal	−3.47	−6.04	−0.90	<0.01
	Parallel-Amorphous	Technological width proximal	−2.64	−5.21	−0.07	0.04
	Ovoid-Convergent	Technological width proximal	2.96	0.07	5.85	0.04
	Parallel-Ovoid	Technological width proximal	−3.47	−6.04	−0.90	<0.01
Ventral plane form	Flat-Bulbar	Maximum thickness	−0.46	−0.85	−0.07	0.01
	Concave-Bulbar	Technological maximum thickness	−0.81	−1.49	−0.12	0.01
	Flat-Bulbar	Technological maximum thickness	−0.76	−1.51	−0.01	0.04
Flake termination	Overshot-Feather	Technological thickness distal	2.24	0.23	4.24	0.02
	Overshot-Hinge	Technological thickness distal	2.76	0.69	4.84	<0.01

Note: All *p*-values are adjusted to account for multiple comparisons in the post hoc tests.

of the original striking platform. Flake thickness measurements show wider variability when flakes have larger bulbs, whereas distal thickness measurements are harder to record consistently on overshot flakes.

Another way to address this question is to examine all flake outliers for each ratio and discrete attribute to identify higher inter-analyst variance scores on specific flakes (Supplemental Table 23). Because we could not apply a systematic method to detect discrete attribute outliers, we do not consider these here (but see Supplemental Table 22 for a qualitative overview of [dis]agreement between analysts per nominal variable). Inter-analyst replicability scores for ratio and discrete-scale attributes were relatively high, but looking at outliers provides a means to explore potential ways of optimizing data recording in future lithic analyses. Supplemental Table 23 summarizes the main flake outlier characteristics for each attribute.

The qualitative assessment of flake outliers shows that some ($n = 14/110$ outliers) are due to high inter-analyst variance driven by one analyst's measurements, which may reflect human error when taking the measurement (e.g., typing error when entering the value). At least one flake consistently appears as an outlier (ID = 58) for several variables due to the breakage of its distal part during transport. We note the highest number of outliers for technological thickness and all four platform measurement attributes, which were also the variables that had lower—albeit substantial—agreement between analysts (IRR >0.6 and <0.8). Large inter-analyst variance in maximum dimension and technological measurements may be due to specific flake shapes (see above). For example, high inter-analyst variance in maximum size seems to occur when the flake maximum dimension is similar to the maximum width (see Figure 8). Inter-analyst variance occurs when the flakes have shapes that vary widely in width (e.g., flakes with expanding edges in the proximal part and convergent edges in the distal portion). Variance in thickness and width measurements may also occur on flakes with large and thick platforms and prominent bulbs, potentially inducing errors while measurements are taken despite the definitions provided (i.e., thickness and width should be measured independently from the platform).

Platform measurements seem more likely to vary between analysts when flakes have a cortical platform or no clear delimitation of the platform (e.g., Figure 9, ID9). In the case of *débordant* (core edge) flakes, issues occur with a blurred line between the platform and the lateral side of the flake (e.g., a relict platform unrelated to the removal of the flake). The large difference in recording dorsal cortex and dorsal scar count appeared to be due to diverse definitions for cortex—in particular, whether there should be a difference between cortex and naturally fractured or weathered surfaces—and what scar types should be counted (see Figure 9). Flakes with a higher inter-analyst variance seem to have a specific set of characteristics (including irregular shape, offset of technological axis compared to maximum dimension, cortical platforms), and they are often *débordant* flakes. In assemblages in which these categories of flakes are few, as in this experimental assemblage, there will be a nonsignificant impact on comparative analyses. Still, for the ones that include high proportions of such flakes, comparative studies should consider the issues raised here.

Does the Inclusion of Images in Definitions Impact Inter-Analyst Replicability?

In seeking to understand potential sources of variation within the group, we considered how visual aids in defining our attributes reduced inter-analyst replicability. The group predicted that analysts would code attributes with images in their definitions more reliably.

Comparisons between nominal attributes with and without images in their definitions show that the presence/absence of images does not significantly impact inter-analyst replicability (F [1,14] = 0.4, $p = 0.53$). The same is true for our ratio-scale and discrete attributes (F [1,21] = 0.01, $p = 0.9$). The best-performing discrete attribute (total dorsal scar counts) showed the highest IRR values for this attribute class. Still, it lacked a visual aid, as did many of our high-performing attributes. Although the nominal attributes' group mean differences are not significant, the IRR score variability around the mean seems different. It appears that including images in definitions reduces nominal attribute IRR variability. However, the sample size of nominal attributes with images is too small to make statistical conclusions. We hesitate to generalize too much from our sample because our definitions

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

were either taken from existing texts or arrived at after considerable group discussion. We hypothesize that our within-group consensus is higher than one would encounter among naive users of our recording system.

Do Differences in Lithic Flaking Systems Impact Inter-Analyst Replicability?

For a recording system like ours to achieve maximum comparability across sites and time, it needs to be robust to differences in lithic reduction strategies. To test the hypothesis that our recording system is insensitive to reduction strategy, we ran our mixed effects models separately on the two flaking systems (recurrent unidirectional Levallois and migrating multiplatform). We then compared the IRR results for each of our three attribute classes (ratio, discrete, and nominal). This comparison allowed us to track differences in replicability between these two broad reduction strategies. If our recording system is insensitive to reduction strategy differences, we should see minor differences in IRR values between the two assemblages (IRR <0.2).

Figure 5 presents the IRR assemblage difference variance contributions for our ratio-scale attributes. Negative values show lower IRR values in the Levallois assemblage, whereas positive values show lower IRR values in the multiplatform multidirectional assemblage. The data show minor inter-analyst replicability differences for all 17 attributes between the two lithic technological systems. About half (8/17) of these differences come from the multiplatform multidirectional assemblage. Technological thickness measurements at the flake proximal and distal ends show similarly high inter-analyst replicability differences in the two assemblages. The fact that knappers distribute mass across the flake differently in these two flaking systems likely drives these thickness differences (Tostevin 2012). This difference is because technologically driven variables can impact the recording of flake thickness at specific points along a flake's margin.

Figure 6 presents the IRR assemblage difference variance contributions for our discrete-scale attributes. These attributes concern flake scar patterns counted in different flake sectors. Again, we see minimal inter-analyst replicability differences for these five attributes between the two lithic reduction

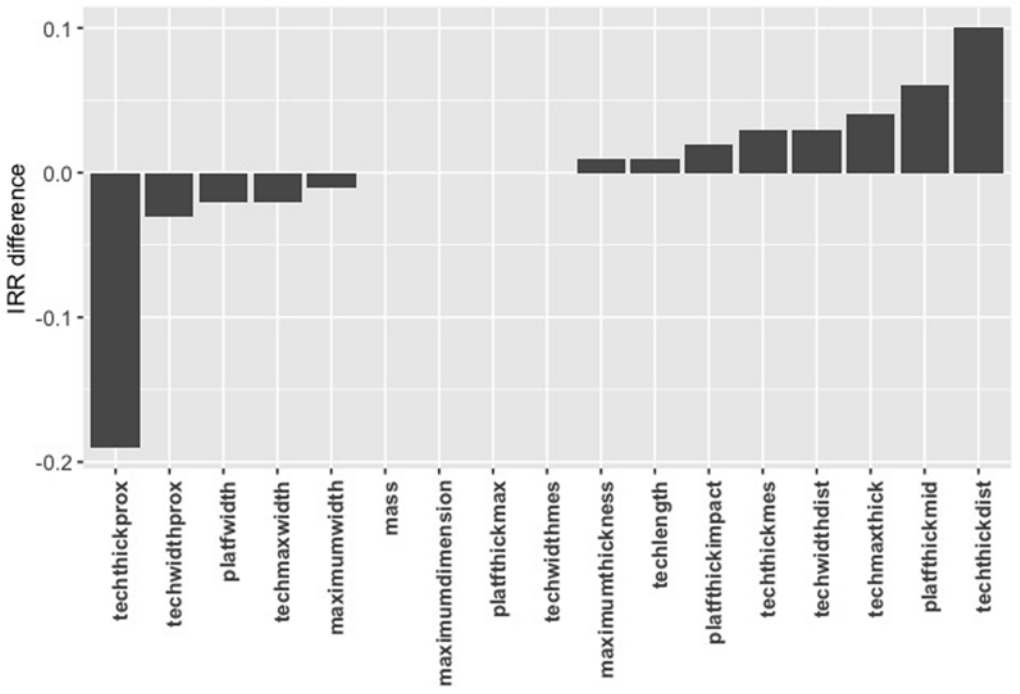

Figure 5. Comparisons of the ratio-scale inter-analyst replicability differences on our two assemblages. Levallois values are arbitrarily converted to negative numbers for graphical reasons.

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

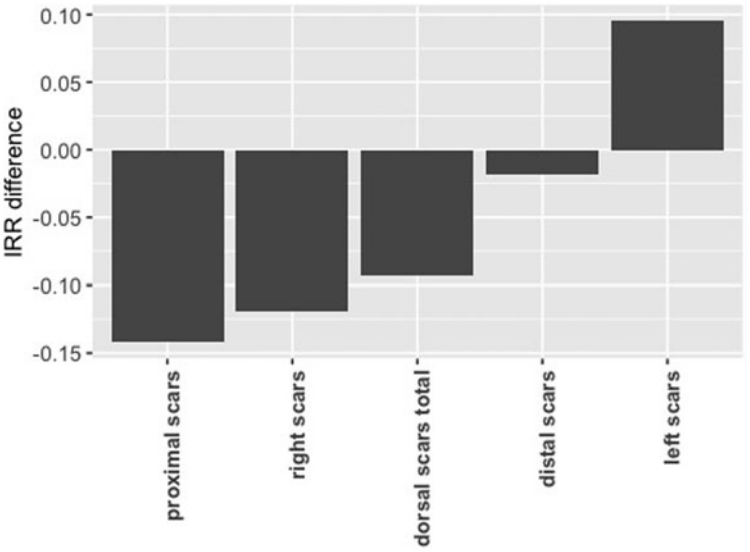

Figure 6. Comparisons of the discrete-scale inter-analyst replicability differences on our two assemblages. Levallois values are arbitrarily converted to negative numbers for graphical reasons.

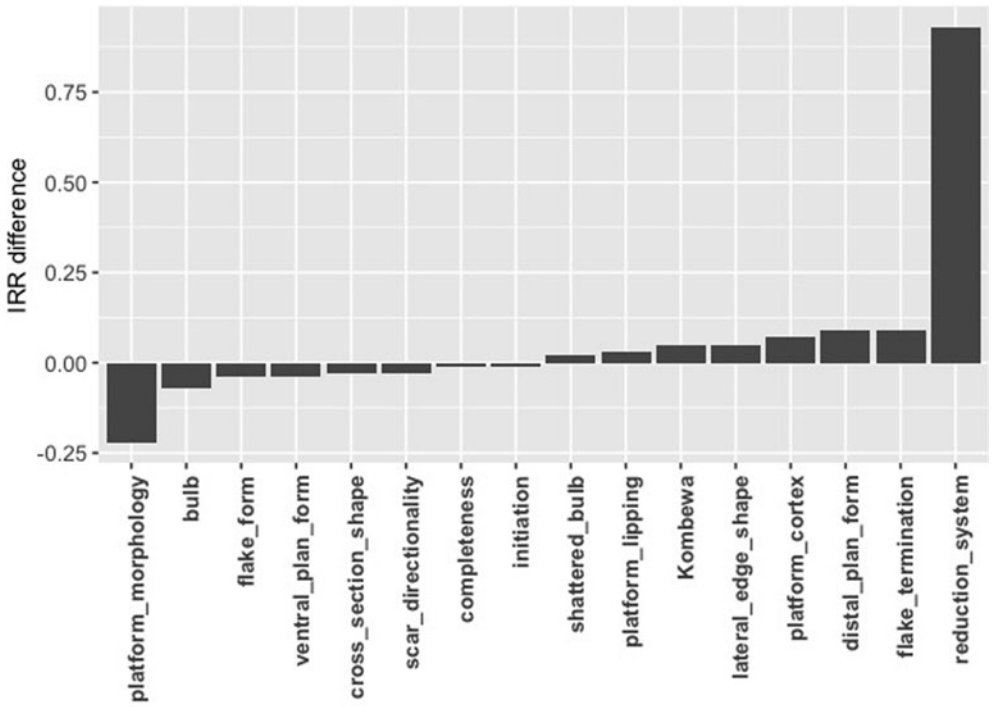

Figure 7. Comparisons of the nominal attribute inter-analyst replicability differences on our two assemblages. Levallois values are arbitrarily converted to negative numbers for graphical reasons.

strategies. Surprisingly, most (4/5) of these differences come from the Levallois assemblage. It is important to note that in at least one commonly used lithic recording system (Van Peer 1992), counting flake scar patterns according to flake sectors is an important component of diagnosing variability within the Levallois approach.

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

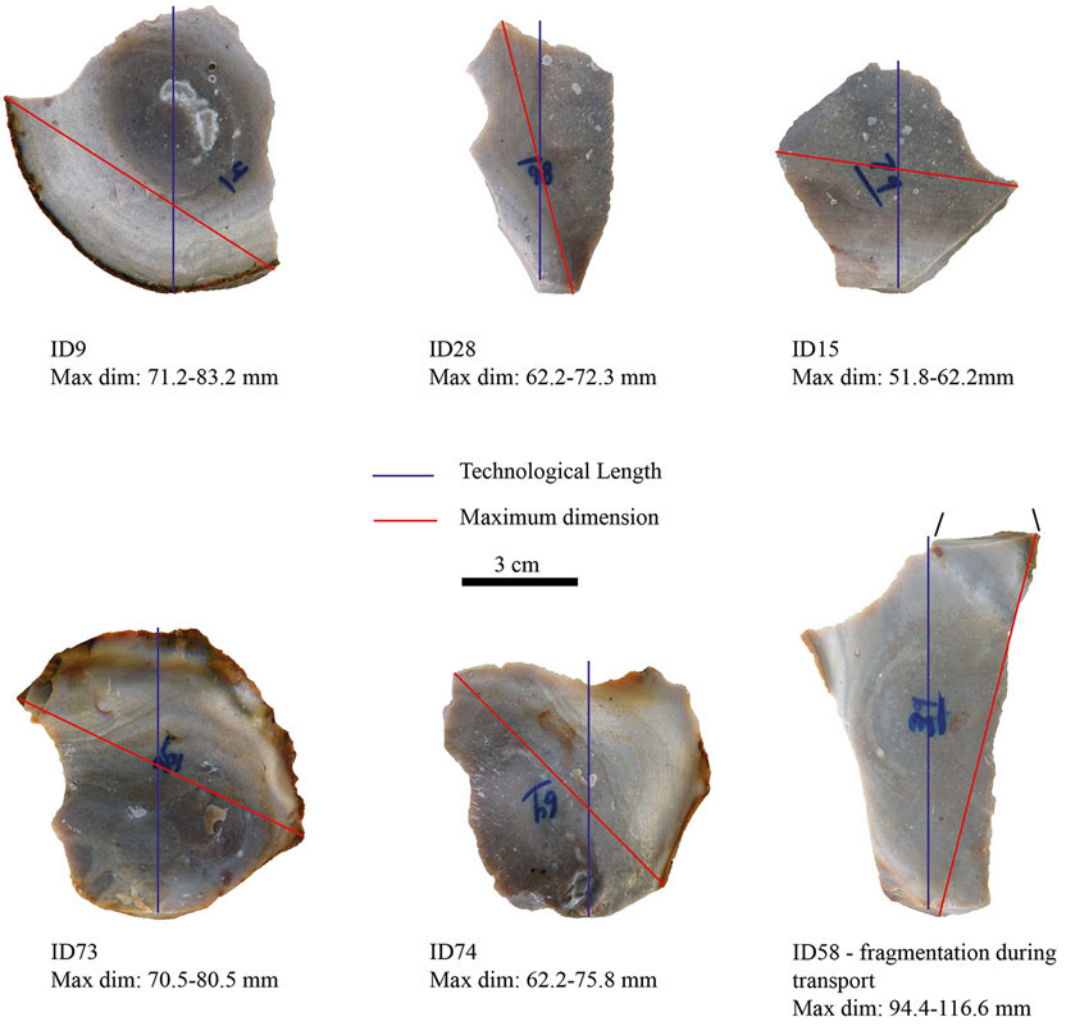

Figure 8. Examples of flake outliers for the maximum dimension attribute. All flakes are oriented ventral face up according to their technological axis, with their proximal part at the bottom. Values show different analysts' range of values on each flake. (Color online)

Figure 7 presents the IRR assemblage difference variance contributions for our nominal attributes. Most of these attributes (14/16) show minor inter-analyst replicability differences between the two lithic technological systems. The data show an even split of the differences ($n = 8$) between the two reduction strategies. Two notably high differences are in the platform morphology (difference = −0.22) and reduction system attributes (difference = 0.93). These results show that analysts agreed less on platform morphologies in the Levallois assemblage than in the multiplatform multidirectional assemblage. They also show that analysts tended to agree when identifying a flake as belonging to the Levallois reduction system, but they struggled to identify flakes from the multiplatform multidirectional system.

Does an Analyst's Experience and Quantitative Training Impact Inter-Analyst Replicability?

A final question concerns the impact of individual differences on inter-analyst replicability. Our analyst survey data (Supplemental Table 1) allowed us to determine three individual difference metrics on the inter-analyst replicability outcomes: years of experience, training in quantitative methods, and training in the *chaîne opératoire* approach. It seems reasonable to hypothesize that analysts with

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

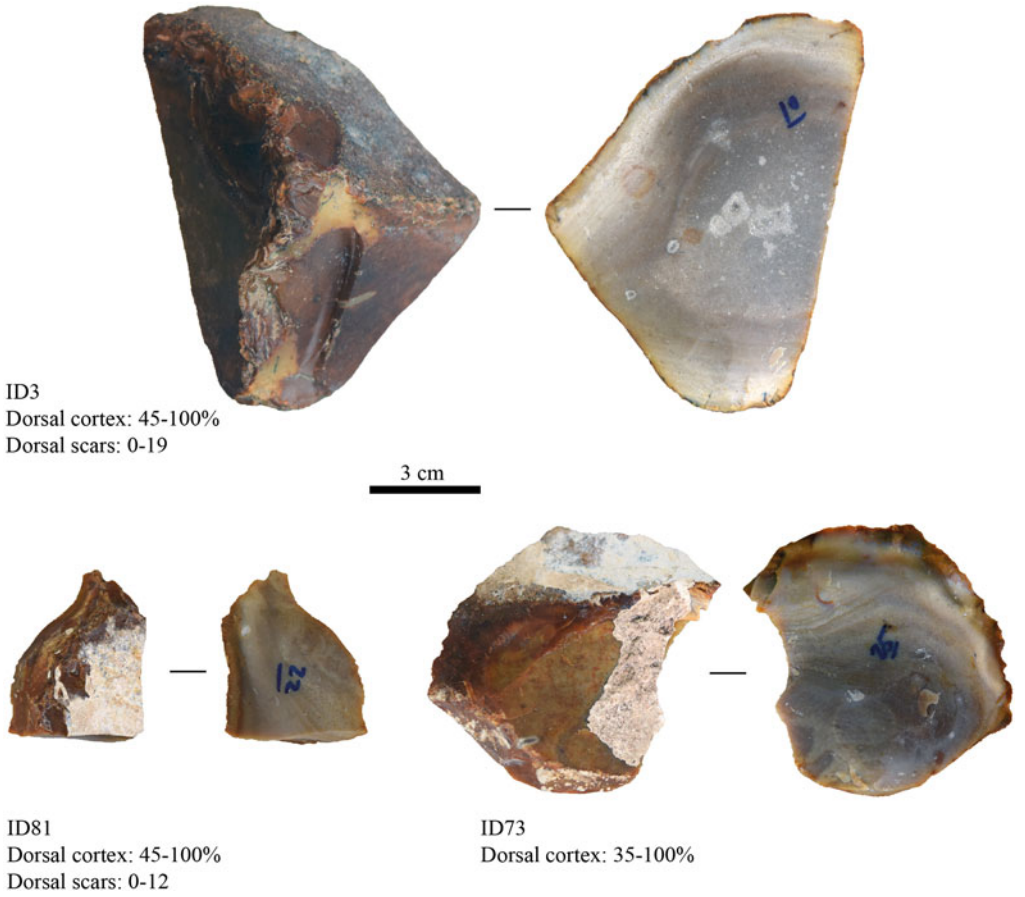

Figure 9. Examples of flake outliers for the dorsal cortex and dorsal scar count attribute. All flakes are oriented according to their technological axis, with their proximal part at the bottom. Values show different analysts' range of values on each flake. (Color online)

greater quantitative training and expertise will more consistently record ratio and discrete-scale attributes.

Here, we focus on how these measures impact the recording of attributes as they provide the most straightforward means of assessing individual measurement performance relative to the group. To do this, we first compared each analyst's distance from the group's average measurements. We then averaged these values across an analyst's suite of measures to derive a single performance metric. A surprising result is that overall years of experience showed a nonsignificant relationship with measurement performance (F [1, 7] = 0.2, R^2 = 0.1, p = 0.65). Our average measurement performance metric is significantly and positively correlated (F [1, 7] = 5.6, R^2 = 0.39, p = 0.04) with an analyst's self-reported ranking of quantitative training levels (1 = lowest, 5 = highest).

Chaîne opératoire training levels show a negative but nonsignificant (F [1, 7] = 5, R^2 = 0.33, p = 0.06) relationship with the average measurement performance metric. There are at least two possible explanations for this. First, most of the data we collected in this study align with more quantitative approaches to lithic analysis. Although it is an oversimplification to state that *chaîne opératoire* approaches are opposed to quantitative research (Soressi and Geneste 2011), they tend to emphasize qualitative readings of artifacts. At least within our group, the data collected for this study were more often unfamiliar to those analysts who employ a *chaîne opératoire* approach extensively. Less experience in collecting some of the data described here likely drives some of the lower inter-rater replicability. The second possibility is that individual responses to our basic survey overestimate or underestimate expertise in particular approaches among the group. Visual inspection of both these results

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

shows that the correlations are driven predominantly by individuals who report very low or high levels of training in each field.

Discussion

We observe that ratio-scale attributes (e.g., maximum flake dimension and technological length) show strong inter-analyst replicability scores, making them simple and immediately suitable for comparative lithic analyses. The discrete-scale attributes, mostly concerned with dorsal negatives on flakes, showed a comparatively low inter-analyst replicability score, likely due to the complexity of identifying flake scar patterns. This result implies that inter-analyst variation stemmed from dividing flakes into "sectors" (left, right, proximal, and distal) rather than the actual counting of dorsal scars and with differences in counting scars occurring in specific sectors versus originating from specific sectors. However, proximal scars performed better than the other sectors. This difference is likely because flakes originating from the proximal end can—often do—retain initiations, making their orientation easier to work out, whereas those from the laterals are more difficult to position. It is important for future work to address these results in more detail because these arbitrary sectors are a central part of several systems used to describe unretouched flakes (e.g., Crew 1975; Tostevin 2012; Van Peer 1992). Finally, our nominal attributes showed more variation in inter-analyst replicability scores, suggesting that additional work is required to ensure that they are reliable for future comparative research.

Why Do Certain Attributes Show Lower Inter-Analyst Replicability Scores and How Can Comparability Be Increased?

Our study showed that increased attribute state counts significantly decrease inter-analyst replicability. More choices increase the chance that analysts code features in different ways. A simple fix would be to collapse certain attribute states into more manageable and reliable classes. However, for platform morphology and directionality, the analysts selected four or more different attribute states for 30% to 60% of the flakes, respectively. This result suggests that it is unlikely that collapsing certain attribute states would provide a satisfying solution for the lower agreement values on these variables (see Supplemental Table 22). Future work could test whether reducing attribute state counts in more complex nominal attributes will allow analysts to track meaningful variability.

Flake and platform shapes and certain technological features (e.g., thick bulbs of percussion) created several issues for the attributes' IRR scores. We cannot deal with flake shape and technological variability simply because lithic artifacts have widely variable shapes and technical features. However, our results show that attribute states such as "amorphous" or "indeterminate" can create uncertainty for analysts taking specific measurements. Still, they do not impact the overall agreement between analysts (see Supplemental Figure 1). The "indeterminate" attribute state serves as a placeholder for times when analysts cannot code a specific attribute but intend to note that there is clear uncertainty with that decision. Future work should look to understand better the flake qualities "indeterminate" states describe and to explain better how the word "indeterminate" is used (e.g., "it cannot be determined" vs. "I cannot determine it").

Our study did not observe any effect of years of experience on analyst performance. Instead, we found significant training background effects. This result suggests that increasing replicability in lithic analysis is more about changing training than increasing experience per se. We suggest that training programs including mixed basic quantitative methods and *chaîne opératoire*–like instructions might help standardize similar lithic attribute analyses. Our group agreed on a unified set of definitions for all the study's attributes and yet, we still found significant inter-analyst differences in some attributes. This result suggests that programs need to provide training to analysts wishing to engage in comparative lithic research, including beta-testing attribute definitions before engaging in primary data collection.

We included illustrations in our attribute definitions, where possible, to increase inter-analyst replicability. Our results show that these images did not significantly impact inter-analyst replicability.

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

One reason is that static images provide a snapshot of the more dynamic measurement process. They also represent one's interpretation of an idealized and specific attribute that does not necessarily capture how analysts measure attributes on variable artifacts. Regardless, images provide a useful set of information that can help strandardize recording systems and that might be more helpful for naive analysts learning to identify specific attributes.

Do Differences in Lithic Flaking Systems Impact Inter-Analyst Replicability?

Our two assemblages show minor differences in inter-analyst replicability. We can identify specific areas where greater IRR differences arise between the two assemblages. These include analysts' difficulties in classifying Levallois platforms, measuring technological thickness, and identifying flakes from the multiplatform multidirectional cores (in the absence of those cores). This result shows biases and weaknesses in existing datasets that researchers could address in future analyses. Notably, the study's recording system is robust to differences in technological strategies, and researchers can use it to compare these technological variants. Given how representative these two flaking systems are of MSA (and Middle Paleolithic) technological variability, our findings are therefore likely to be generalizable to flakes made from other core reduction strategies.

How Do Our Results Compare to Prior Inter-Analyst Replicability Lithic Studies?

The previous studies listed in Table 1 agree that inter-analyst replicability is an issue that researchers should address more thoroughly in lithic analysis. Unfortunately, most of these studies suffer in having few analysts, a small number of tested attributes, small lithic samples, and they generally serve as a starting point for further comparative lithic analyses. With some notable exceptions (e.g., Gnaden and Holdaway 2000; Proffitt and de la Torre 2014), these studies also lack quantitative assessment of inter-analyst replicability. As a further caveat, we did not examine accuracy as some experiments have done (e.g., Gnaden and Holdaway 2000; Proffitt and de la Torre 2014)—testing measurements against a true standard or "correct answer."

In common with this study, Fish (1978) and Gnaden and Holdaway (2000) found metric attributes such as length or thickness highly reliable. Maximum width and platform thickness performed well in our study, as they did for Fish (1978) and Wilmsen and Robert (1978), but they showed high inter-analyst variability in another experiment (Gnaden and Holdaway 2000). These latter authors attributed some of the systematic errors to differences in definitions that we ruled out in our study, underlining their general importance. The high variance in recording dorsal cortex aligns our study with others (Fish 1978; Gnaden and Holdaway 2000), demonstrating the need for more precise definitions of what researchers should consider "true" cortical faces. Interestingly, cortex identification on dorsal surfaces was among the most reliable attributes in another experiment on an Oldowan assemblage (Proffitt and de la Torre 2014), which may speak more to raw material differences between these studies. As our study did, Proffitt and de la Torre (2014) found that the direction of dorsal negatives compared poorly among researchers. Although our study found substantial to strong IRR values for many attributes (>0.6), Proffitt and de la Torre (2014) report mostly moderate levels of agreement (0.4–0.6), which likely stemmed from their use of three different raw materials, including quartzite, which performed the worst. Chert, which we used, had the highest agreement values among their analysts. Driscoll's (2011) quartz-based study also found low replicability between observers for several discrete variables. Timbrell and colleagues' (2022) recent study examined shape variable replicability via outline 2D GMM and linear measurements. They found inter-analyst error to be low enough for accurate analyses with both methods. Unfortunately, no previous study had ratio, discrete, and nominal variables in their design, precluding comparisons to this experiment on the level of different measurement scales.

Limitations and Recommendations

We hope to overcome several limitations in future studies, including testing the impact of different raw materials and including a greater range of flake production strategies. Unretouched flakes comprise only one (if the most abundant) component of the Paleolithic record that lithic analysts study. The

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

Table 3. Recommended Attributes Showing Strong Inter-Analyst Replicability Scores (>0.8) in the Current Study.

Attribute	Units	Data type
General metric descriptors		
Mass	g	Ratio
Maximum dimension	mm	Ratio
Maximum width	mm	Ratio
Maximum thickness	mm	Ratio
Technological measures		
Technological length	mm	Ratio
Technological maximum width	mm	Ratio
Technological maximum thickness	mm	Ratio
Reduction intensity measures		
Total dorsal scar count		Nominal
Platform cortex presence/absence		Categorical
Reduction strategy indicators		
Initiation type		Categorical
Kombewa		Categorical
Reduction system		Categorical
Platform lipping		Categorical
Bulb (?)		Categorical
Flake termination		Categorical
Flake form		Categorical
Completeness measures		
Flake completeness		Categorical

extent to which we can extend our broad measures of inter-analyst replicability to other artifact classes (e.g., cores and retouched tools) is unknown.

Another area for future research is the need to understand different sources of variation in lithic data. Random variation about the mean can arise due to minor differences in data collection. More worrisome are differences that occur because of systematic errors, which stem from unclear definitions that affect a measurement's central tendency (Gnaden and Holdaway 2000). Figure 2 shows the range of CV values for our ratio-scale variables, which despite ranging from roughly 0.01 to 0.15, are still very low when compared to any non-machine-based method of data collection (cf. Eerkens and Lipo 2005). This result suggests that random variation, although present in our study, is relatively modest. Systematic errors are likely more difficult to detect but were undoubtedly present in our dataset, especially in our efforts to measure exterior platform angle using the modified caliper method initially described by Dibble and Bernard (1980).

One major limitation of our study is that we do not track variations in exterior platform angle and flake curvature measurements. We currently have little basis other than two studies (Andrefsky 2005; Dibble and Bernard 1980) from which to assess these attributes' effectiveness. Our group could not reliably use the modified caliper method as published by Dibble and Bernard (1980), and future work should aim to retest and refine this method. There is also a general implicit assumption among many lithic analysts that platform angles and curvature values are difficult to measure accurately. Addressing this issue could require 3D scans on new software to record exterior platform

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

angle and flake curvature accurately (Valletta et al. 2020; Yezzi-Woodley et al. 2021). It remains uncertain how these patterns of inter-analyst variation might impact assemblage-level comparisons. Although costly and impractical on larger lithic assemblages with hundreds or thousands of specimens, such research would produce results of broad relevance to comparative lithic analysts.

Based on our current study, we identify several ways to advance replicability in stone tool analysis. First, a standard set of transparent, clear, and agreed-upon definitions of attributes are necessary for any comparative study. Ratio-scale attributes fared by far the best, and key attributes such as mass, maximum dimension, technological length, maximum width, and thickness are easy to record. They can form a good comparative baseline, given that many researchers already use them. Many nominal attributes showed higher replicability than expected by us and likely others—such as flake form—and increasing comparability can further be achieved by decreasing the number of attribute states in some cases.

In contrast, many attributes associated with flake shape showed lower replicability in our study. As a way forward, we recommend increasing sample sizes while using photogrammetry or morphometric methods designed to capture shape quantitatively (e.g., Bretzke and Conard 2012; Grosman et al. 2008; Iovita 2011; Magnani et al. 2020; Ranhorn et al. 2019). Ideally, we should explore these options using approaches that are increasingly accessible as costs decline and that researchers can capture on widely available devices (e.g., Cerasoni et al. 2022; Porter et al. 2016). The same goes for measuring angles (such as EPA) and curvature. For data recording, analysts should use relational databases built using programs such as E4, instead of spreadsheets. Instruction should ideally work with static images and dynamic visuals, such as short training videos showing how to measure in three dimensions. Moreover, researchers might more reliably record some variables (i.e., flake scar sectors) on images rather than actual implements.

Our results show that enhancing replicability in comparative studies in the MSA, or any other period, is not dependent on experience but rather on basic training in quantitative methods. To be clear, quantitative data are not "better" than more qualitative interpretations. They are simply more replicable, and illustrations and technological readings using *chaîne opératoire* and allied approaches remain an essential component of lithic analysis because they provide complementary information.

As stated at the outset, this project's initial and long-term goal was to assemble large datasets to explore patterning at the subcontinental and smaller scales across Africa, also making use of the enormous quantities of data already gathered by researchers over the last decades. Based on our experiences thus far and the results presented above, Table 3 lists those variables ($n = 17$) that we consider reliable when compiling datasets (published or otherwise) collected by different individuals for comparative ends, using definitions consistent with those we used here. Note that Table 3 provides general guidelines for interpreting inter-analyst replicability, with values >0.6 considered "substantially" reliable (Cohen 1960). However, our results stem from definitions and protocols extensively worked out through hundreds of hours of collaborative conversation and writing. We cannot extend these measures uncritically to other research teams. Consequently, we take a more conservative view, favoring those variables that have IRR scores of >0.8 and that showed minimal effects of interactions with other variables. We group the attributes according to potential uses for exploring a range of lithic artifact research questions, including basic metric parameters, flake propagation measures, measures of reduction intensity, core reduction strategies, and basic flake breakage indicators.

Conclusions

The issue of data comparability is particularly acute in the analysis of lithic (stone) artifacts that dominate the Paleolithic record. Addressing this issue, we presented the most extensive study yet on replicability in lithic analysis based on a total of 11 analysts, 100 lithics, 38 attributes, and hundreds of hours of collaborative conversation and writing. Although initially geared toward the African Middle Stone Age record, based on the diverse range of experience of the participating researchers, this study has broad applicability to analyzing stone tools across all regions and periods. The most salient finding of our study is that the 11 international expert lithic analysts performed well across many of the attributes tested in the study. Ratio-scale attributes fared the best, but several

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

nominal-scale attributes showed promise when used with the definitions provided in this study. We conclude that high replicability in lithic analysis is possible, providing the baseline for any comparative study, at least under specific methodological designs. Apart from its general relevance for the field of lithic analysis, this finding is important given this project's original goal of comparing lithic assemblages across the MSA of Africa, including datasets already collected by researchers and new ones.

Acknowledgments. This article is CoMSAfrica publication number 2.

Funding Statement. We thank the Radcliffe Institute for Advanced Study for funding the initial CoMSAfrica meeting held in 2018. We also thank the Swiss National Science Foundation (SNSF, grant #IZSEZ0_186545) for funding the second workshop in 2019 in Switzerland, and the Faculty of Sciences of the University of Geneva for supporting both financially and logistically this second CoMSAfrica meeting. Alice Leplongeon's research is funded by a grant from the Research Foundation in Flanders (FWO, grant Q36#12U9220N).

Data Availability Statement. Detailed results of all analyses and assessments of the data structure are available in this article's supplementary materials and through the Open Science Framework (https://osf.io/seh2t/?view_only=9097ef58225b49e48f66afb220022fbf).

Competing Interests. The authors declare none.

Supplemental Material. For supplemental material accompanying this article, visit https://doi.org/10.1017/aaq.2023.4.

Supplemental Figure 1. Visual summary of the gwet results with and without indeterminates.

Supplemental Table 1. Results of the prior experience survey for all study participants.

Supplemental Table 2. Overview and definitions for the study's attributes.

Supplemental Table 3. Summary IRR statistics for the study's ratio and discrete scale attributes.

Supplemental Table 4. Summary IRR statistics for the study's nominal scale attributes.

Supplemental Table 5. Detailed summary statistics broken down by flake and subject for the study's ratio-scale attributes.

Supplemental Table 6. Detailed summary statistics broken down by flake and subject for the study's platform lipping nominal scale attribute.

Supplemental Table 7. Detailed summary statistics broken down by flake and subject for the study's bulb type nominal scale attribute.

Supplemental Table 8. Detailed summary statistics broken down by flake and subject for the study's platform morphology nominal scale attribute.

Supplemental Table 9. Detailed summary statistics broken down by flake and subject for the study's flake initiation nominal scale attribute.

Supplemental Table 10. Detailed summary statistics broken down by flake and subject for the study's flake scar directionality nominal scale attribute.

Supplemental Table 11. Detailed summary statistics broken down by flake and subject for the study's flake form nominal scale attribute.

Supplemental Table 12. Detailed summary statistics broken down by flake and subject for the study's reduction system nominal scale attribute.

Supplemental Table 13. Detailed summary statistics broken down by flake and subject for the study's kombewa presence nominal scale attribute.

Supplemental Table 14. Detailed summary statistics broken down by flake and subject for the study's shattered bulb nominal scale attribute.

Supplemental Table 15. Detailed summary statistics broken down by flake and subject for the study's flake distal plan form nominal scale attribute.

Supplemental Table 16. Detailed summary statistics broken down by flake and subject for the study's flake initiation nominal scale attribute.

Supplemental Table 17. Detailed summary statistics broken down by flake and subject for the study's flake lateral edge type nominal scale attribute.

Supplemental Table 18. Detailed summary statistics broken down by flake and subject for the study's flake platform cortex scale attribute.

Supplemental Table 19. Detailed summary statistics broken down by flake and subject for the study's flake section nominal scale attribute.

Supplemental Table 20. Detailed summary statistics broken down by flake and subject for the study's flake completeness nominal scale attribute.

Supplemental Table 21. Detailed summary statistics broken down by flake and subject for the study's flake ventral plan form nominal scale attribute.

Supplemental Table 22. Detailed summary statistics for the study's outlier flakes according to the nominal scale attributes.

Supplemental Table 23. Detailed overview of the number of outlier flakes for each of study's ratio scale attributes and explanations for why they were classed as outliers.

Author Contributions. CT and MW started the CoMSAfrica project by organizing a workshop funded by the Radcliffe Institute for Advanced Study in November 2018. All authors except ME designed the study during this workshop and subsequent online meetings (2018–2019). Following the initial identification of the attributes to be included in the study during the workshop, AL, JM, and KR led a working group on the definitions of the variables, which were refined and agreed upon by the whole group. JM and KR implemented the database in E4 and Microsoft Excel format. JP, MS, and KD worked on developing a common communication platform and a digital place where researchers could collaborate and share data. AM, ES, AL, CT, and JP developed a plan to create a test assemblage. ME made the test assemblage.

All authors analyzed the test assemblage (2019–2021). Analyses were agreed upon and refined by all authors during a workshop organized at the University of Geneva and funded by the Swiss National Science Foundation (application for funding led by KD, HG, and MW). JP conducted the statistical analyses. AL performed the data cleaning and generated summary statistics.

All authors contributed to the writing of the article: AM, MW, CT, and JP wrote the initial draft.

JP and AL reported the results and discussion sections. All authors contributed to the revisions of the article.

References Cited

Andrefsky, William. 2005. *Lithics: Macroscopic Approaches to Analysis*. Cambridge University Press, Cambridge.

Baker, Monya. 2016. Reproducibility Crisis. *Nature* 533:353–366.

Blumenschine, Robert J., Curtis W. Marean, and Salvatore D. Capaldo. 1996. Blind Tests of Inter-Analyst Correspondence and Accuracy in the Identification of Cut Marks, Percussion Marks, and Carnivore Tooth Marks on Bone Surfaces. *Journal of Archaeological Science* 23:493–507.

Boyd, Clifford C. 1987. Interobserver Error in the Analysis of Nominal Attribute States: A Case Study. *Tennessee Anthropologist* 12:88–95.

Bretzke, Knut, and Nicholas J. Conard. 2012. Evaluating Morphological Variability in Lithic Assemblages Using 3D Models of Stone Artifacts. *Journal of Archaeological Science* 39:3741–3749.

Calogero, Barbara. 1992. Lithic Misidentification. *Man in the Northeast* 43:87–90.

Carleton, Christopher W., and Huw S. Groucutt. 2021. Sum Things Are Not What They Seem: Problems with Point-Wise Interpretations and Quantitative Analyses of Proxies Based on Aggregated Radiocarbon Dates. *Holocene* 31:630–643.

Cerasoni, Jacopo Niccolò, Felipe do Nascimento Rodrigues, Yu Tang, and Emily Yuko Hallett. 2022. Do-It-Yourself Digital Archaeology: Introduction and Practical Applications of Photography and Photogrammetry for the 2D and 3D Representation of Small Objects and Artefacts. *PLoS ONE* 17(4):e0267168. https://doi.org.10.1371/journal.pone.0267168.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20:37–46.

Conard, Nicholas J., Marie Soressi, John E. Parkington, Sarah Wurz, and Royden Yates. 2004. A Unified Lithic Taxonomy Based on Patterns of Core Reduction. *South African Archaeological Bulletin* 59:13–17.

Crew, Harry L. 1975. An Examination of the Variability of the Levallois Method: Its Implications for the Internal and External Relationships of the Levantine Mousterian. Ph dissertation, UC Davis, Department of Anthropology, University of Michigan, Ann Arbor.

Crowther, Alison, and Michael Haslam. 2007. Blind Tests in Microscopic Residue Analysis: Comments on Wadley et al. (2004). *Journal of Archaeological Science* 34:997–1000.

Dibble, Harrold L., and M. C. Bernard. 1980. A Comparative Study of Edge Angle Measurement Techniques. *American Antiquity* 45:857–865.

Dogandžić, Tamara, David R. Braun, and Shannon P. McPherron. 2015. Edge Length and Surface Area of a Blank: Experimental Assessment of Measures, Size Predictions and Utility. *PLoS ONE* 10(9):e0133984. https://doi.org/10.1371/journal.pone.0133984.

Driscoll, Killian. 2011. Vein Quartz in Lithic Traditions: An Analysis Based on Experimental Archaeology. *Journal of Archaeological Science* 38:734–745.

Eerkens, Jelmer W., and Carl P. Lipo. 2005. Cultural Transmission Theory and the Archaeological Record: Providing Context to Understanding Variation and Temporal Changes in Material Culture. *Journal of Archaeological Research* 15:239–274.

Fish, Paul R. 1978. Consistency in Archaeological Measurement and Classification: A Pilot Study. *American Antiquity* 43(1):86–89.

Gnaden, Denis, and Simon Holdaway. 2000. Understanding Observer Variation When Recording Stone Artifacts. *American Antiquity* 65:739–748.

Grosman, Leore, Oded Smikt, and Uzy Smilansky. 2008. On the Application of 3-D Scanning Technology for the Documentation and Typology of Lithic Artifacts. *Journal of Archaeological Science* 35:3101–3110.

Gwet, Kilem L. 2008. Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology* 61:29–48.

Hallgren, Kevin A. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology* 8:23.

Harmand, Sonia, Jason E. Lewis, Craig S. Feibel, Christopher J. Lepre, Sandrine Prat, Arnaud Lenoble, Xavier Boes, et al. 2015. 3.3-Million-Year-Old Stone Tools from Lomekwi 3, West Turkana, Kenya. *Nature* 521:310–315.

Holdaway, Simon, and Nicola Stern. 2004. *A Record in Stone: The Study of Australia's Flaked Stone Artifacts*. Aboriginal Studies Press, Melbourne.

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

Iovita, Radu. 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.

Lycett, Stephen J., Noreen von Cramon-Taubadel, and Robert A. Foley. 2006. A Crossbeam Co-ordinate Caliper for the Morphometric Analysis of Lithic Nuclei: A Description, Test and Empirical Examples of Application. *Journal of Archaeological Science* 33:847–861.

Mackay, Alex. 2008. On the Production of Blades and Its Relationship to Backed Artefacts in the Howiesons Poort at Diepkloof, South Africa. *Lithic Technology* 33:87–99.

Magnani, Matthew, Matthew Douglass, Whittaker Schroder, Jonathan Reeves, and David R. Braun. 2020. The Digital Revolution to Come: Photogrammetry in Archaeological Practice. *American Antiquity* 85:737–760.

Mauz, Barbara, Loïc Martin, Michael Discher, Chantal Tribolo, Sebastian Kreutzer, Chiara Bahl, Andreas Lang, and Nobert Mercier. 2021. On the Reliability of Laboratory Beta-Source Calibration for Luminescence Dating. *Geochronology* 3:371–381.

Newcomer, Mark, Roger Grace, and Romana Unger-Hamilton. 1986. Investigating Microwear Polishes with Blind Tests. *Journal of Archaeological Science* 13:203–217.

Perpère, Marie. 1986. Apport de la typométrie à la définition des éclats Levallois: l'exemple d'Ault. *Bulletin de la Société Préhistorique Française* 83:115–118.

Porter, Samantha Thi, Morgan Roussel, and Marie Soressi. 2016. A Simple Photogrammetry Rig for the Reliable Creation of 3D Artifact Models in the Field: Lithic Examples from the Early Upper Paleolithic Sequence of Les Cottés (France). *Advances in Archaeological Practice* 4:71–86.

Proffitt, Thomas, and Ignacio de la Torre. 2014. The Effect of Raw Material on Inter-Analyst Variation and Analyst Accuracy for Lithic Analysis: A Case Study from Olduvai Gorge. *Journal of Archaeological Science* 45:270–283.

Ranhorn, Kathryn L., David R. Braun, Rebecca E. Biermann Gürbüz, Elliot Greiner, Daniel Wawrzyniak, and Alison S. Brooks. 2019. Evaluating Prepared Core Assemblages with Three-Dimensional Methods: A Case Study from the Middle Paleolithic at Skhūl (Israel). *Archaeological and Anthropological Sciences* 11:3225–3238.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rots, Veerle, Louis Pirnay, Ph Pirson, and Odette Baudoux. 2006. Blind Tests Shed Light on Possibilities and Limitations for Identifying Stone Tool Prehension and Hafting. *Journal of Archaeological Science* 33:935–952.

Scerri, Eleanor M. L., Nick A. Drake, Richard Jennings, and Huw S. Groucutt. 2014. Earliest Evidence for the Structure of *Homo Sapiens* Populations in Africa. *Quaternary Science Reviews* 101:207–216.

Scott, E. Marian, Philip Naysmith, and Gordon T. Cook. 2018. Why Do We Need ^{14}C Inter-Comparisons?: The Glasgow-^{14}C Inter-Comparison Series, a Reflection over 30 Years. *Quaternary Geochronology* 43:72–82.

Shea, John J. 2013. *Stone Tools in the Paleolithic and Neolithic Near East: A Guide*. Cambridge University Press, Cambridge.

Shea, John J. 2016. *Stone Tools in Human Evolution: Behavioral Differences among Technological Primates*. Cambridge University Press, Cambridge.

Shea, John J. 2020. *Prehistoric Stone Tools of Eastern Africa: A Guide*. Cambridge University Press, Cambridge.

Soressi, Marie, and Jean-Michel Geneste. 2011. The History and Efficacy of the *Chaîne Opératoire* Approach to Lithic Analysis: Studying Techniques to Reveal Past Societies in an Evolutionary Perspective. *PaleoAnthropology* 2011: 334–350.

Stewart, Mathew, W. Christopher Carleton, and Huw S. Groucutt. 2021. Climate Change, Not Human Population Growth, Correlates with Late Quaternary Megafauna Declines in North America. *Nature Communications* 12:1–15.

Stoffel, Martin A., Shinichi Nakagawa, and Holger Schielzeth. 2017. rptR: Repeatability Estimation and Variance Decomposition by Generalized Linear Mixed-Effects Models. *Methods in Ecology and Evolution* 8:1639–1644.

Timbrell, Lucy, Christopher Scott, Behailu Habte, Yosef Tefera, Hélène Monod, Mouna Qazzih, Benjamin Marais, et al. 2022. Testing Inter-Observer Error under a Collaborative Research Framework for Studying Lithic Shape Variability. *Archaeological and Anthropological Sciences* 14:209. https://doi.org/10.1007/s12520-022-01676-2.

Tixier, Jacques. 1963. *Typologie de l'épipaléolithique du Maghreb*. Mémoires du Centre de recherches anthropologiques, préhistoriques et ethnographiques 2. Arts et métiers graphiques, Paris.

Tostevin, Gilbert B. 2012. *Seeing Lithics: A Middle-Range Theory for Testing for Cultural Transmission in the Pleistocene*. Oxbow Books, Oakville, California.

Valletta, Francesco, Uzy Smilansky, A. Nigel Goring-Morris, and Leore Grosman. 2020. On Measuring the Mean Edge Angle of Lithic Tools Based on 3-D Models–A Case Study from the Southern Levantine Epipalaeolithic. *Archaeological and Anthropological Sciences* 12:Article 49.

Van Peer, Philip. 1992. *The Levallois Reduction Strategy*. Prehistory Press, Madison, Wisconsin.

Wadley, Lyn, Marlize Lombard, and Bonnie Williamson. 2004. The First Residue Analysis Blind Tests: Results and Lessons Learnt. *Journal of Archaeological Science* 31:1491–1501.

Wilkins, Jayne, Kyle S. Brown, Simen Oestmo, Telmo Pereira, Kathryn L. Ranhorn, Benjamin J. Schoville, and Curtis W. Marean. 2017. Lithic Technological Responses to Late Pleistocene Glacial Cycling at Pinnacle Point Site 5-6, South Africa. *PLoS ONE* 12(3):e0174051. https://doi.org/10.1371/journal.pone.0174051.

Will, Manuel, Christian Tryon, Matthew Shaw, Eleanor M. L. Scerri, Kathryn Ranhorn, Justin Pargeter, Jessica McNeil, Alex Mackay, Alice Leplongeon, and Huw S. Groucutt. 2019. Comparative Analysis of Middle Stone Age Artifacts in Africa (CoMSAfrica). *Evolutionary Anthropology* 28:57–59.

https://doi.org/10.1017/aaq.2023.4 Published online by Cambridge University Press

Wilmsen, Edwin N., and Frank H. H. Robert, Jr. 1978 *Lindenmeier, 1934–1974: Concluding Report on Investigations*. Smithsonian Contributions to Anthropology 24. Smithsonian Institution, Washington, DC.

Yezzi-Woodley, Katrina, Jeff Calder, Peter J. Olver, Paige Cody, Thomas Huffstutler, Alexander Terwilliger, J. Anne Melton, Martha Tappen, Reed Coil, and Gilbert Tostevin. 2021. The Virtual Goniometer: Demonstrating a New Method for Measuring Angles on Archaeological Materials Using Fragmentary Bone. *Archaeological and Anthropological Sciences* 13: Article 106.

Cite this article: Pargeter, Justin, Alison Brooks, Katja Douze, Metin Eren, Huw S. Groucutt, Jessica McNeil, Alex Mackay, et al. 2023. Replicability in Lithic Analysis. *American Antiquity* **88**(2):163–186. https://doi.org/10.1017/aaq.2023.4.