Journal of Paleolithic Archaeology
https://doi.org/10.1007/s41982-021-00087-4

Testing the Integrity of the Middle and Later Stone Age Cultural Taxonomic Division in Eastern Africa

Matt Grove[1] · James Blinkhorn[2,3]

Accepted: 1 February 2021/Published online: 10 May 2021

Abstract

The long-standing debate concerning the integrity of the cultural taxonomies employed by archaeologists has recently been revived by renewed theoretical attention and the application of new methodological tools. The analyses presented here test the integrity of the cultural taxonomic division between Middle and Later Stone Age assemblages in eastern Africa using an extensive dataset of archaeological assemblages. Application of a penalized logistic regression procedure embedded within a permutation test allows for evaluation of the existing Middle and Later Stone Age division against numerous alternative divisions of the data. Results suggest that the existing division is valid based on any routinely employed statistical criterion, but that is not the single best division of the data. These results invite questions about what archaeologists seek to achieve via cultural taxonomy and about the analytical methods that should be employed when attempting revise existing nomenclature.

Keywords Middle Stone Age · Later Stone Age · Cultural taxonomy · Lithic technology · Logistic regression · Permutation analysis

This article belongs to the Topical Collection: *Cultural taxonomies in the Palaeolithic - old questions, novel perspectives*
Guest Editors: Felix Riede and Shumon T. Hussain

✉ Matt Grove
matt.grove@liverpool.ac.uk

James Blinkhorn
blinkhorn@shh.mpg.de

[1] Department of Archaeology, Classics and Egyptology, University of Liverpool, 8-14 Abercromby Square, Liverpool L69 7WZ, UK

[2] Pan African Evolution Research Group, Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany

[3] Centre for Quaternary Research, Department of Geography, Royal Holloway, University of London, Egham TW20 0EX, UK

Introduction

A number of recent papers have revived long-standing debates concerning the validity of the cultural taxonomies adopted by archaeologists (e.g. Riede et al. 2020; Reynolds and Riede 2019a; Sauer and Riede 2019; Ivanovaite et al. 2020). Although the foregoing papers focus on the Late Upper Palaeolithic of Europe, they form the latest instalment of a debate that is as old as archaeology itself and has at various points encompassed all periods and regions (e.g. Bishop and Clark 1967; Dunnell 1971; Clark and Lindly 1991; Bisson 2000; O'Brien and Lyman 2002; Shea 2014, 2020). Of particular relevance to the current paper is the fact that Shea (2019) notes a clear parallel between the problems identified by Reynolds and Riede (2019a) for the European Upper Palaeolithic and those encountered in the analysis of eastern African archaeological material (Will et al. 2019; Shea 2020).

Debates on the validity of cultural taxonomy have a long history in African archaeology (e.g. Goodwin and Van Riet Lowe 1929; Bishop and Clark 1967; Shea 2020), with the added complication that most early classificatory schemes involved a 'bastardisation of European terminology' (Goodwin and Van Riet Lowe 1929:97) that was poorly suited to the African evidence. Indeed, Goodwin (1958:33) reflected that prior to the establishment of a purpose-built African terminology 'we had been trying to describe giraffe in terms of camel, or eland in terms of elk'. The inadequacy of European terminology prompted the establishment of a bipartite division of (southern) African material into Earlier and Later Stone Ages, ratified at the 24th annual meeting of the South African Association for the Advancement of Science in Pretoria, 1926 (Goodwin 1926). Continuing research by Goodwin and Van Riet Lowe (Goodwin 1928; Goodwin and Van Riet Lowe 1929) soon led them to recognize that the inclusion of a third period—the Middle Stone Age—was 'essential to cover the facts observed' (Goodwin 1946:74). This tripartite division subsequently became the norm throughout sub-Saharan Africa.

From the outset, it was recognized that there were varied regional and chronological facies within each of the major 'Ages', that certain industries could be regarded as transitional between them, and that the differences between them were quantitative rather than qualitative. Goodwin (1946:74) was careful to note that 'the three periods overlap to some extent… we only reach each new "Age" as the new technique becomes dominant'; delegates at the Third Pan-African Congress on Prehistory in 1955 resolved that 'more elasticity' was required in the use of the three Ages (Cole 1955:204) and adopted two intermediate stages between them (Cole 1955; Clark 1957). As the three Ages came to be used over a greater extent of the continent, it became clear that they could not be used as chronological markers and that 'time connotations must be separated from cultural concepts' (Bishop and Clark 1967:866); transitions from one Age to another could be protracted and did not occur simultaneously—nor even necessarily follow the same trajectories—in different geographical areas (Scerri et al. 2021).

The Burg Wartenstein symposium of 1965 was highly critical of the typology developed by Goodwin and Van Riet Lowe, and even more critical of its rather lax subsequent use; indeed, Isaac's proposal that 'the terms "Earlier", "Middle", and "Later" Stone Age in Africa should be abolished for all formal usage' was agreed unanimously (Bishop and Clark 1967:867). Kleindienst (1967) noted that publications

of indicative assemblages with adequate descriptions were scarce and that as such prehistorians had tended to use 'the same terms but with different definitions, different connotations, and at different levels of abstraction' (Kleindienst 1967:828); her extensive lexicon makes such problems abundantly clear. Although the Burg Wartenstein delegates advocated the abandonment of the three Ages, the major issue for archaeologists over the subsequent decades was that no concrete suggestions had been made as to what would replace them (Sampson 1974; Parkington 1993; Underhill 2011). As such, the three Ages have retained their dominance over the African record; Clark's (1969) mode system provides a useful accompaniment, and Shea's (2020) EAST typology may yet have significant impact, but of the assemblages analysed below, all are designated by their excavators as either Middle or Later Stone Age.

Goodwin and Van Riet Lowe's (1929) distinction between Middle and Later Stone Age implements rests upon differences in the preparation of the striking platform and in the nature of the resultant flakes. MSA flakes are marked by faceted striking platforms and convergent edges, unlike the flat striking platforms and parallel edges of the LSA. The essentially triangular MSA flake is therefore 'eminently suitable for use as a point, and, indeed, the typical implement throughout the Middle Stone Age is the worked point in a variety of forms' (Goodwin and Van Riet Lowe 1929:98). These basic distinctions persist; contemporary researchers stress decreases in prepared core technologies and retouched points together with increases in the production of backed pieces, prismatic blades and bladelets, and bipolar reduction as signalling the transition from the MSA to the LSA (Gossa et al. 2012; Pleurdeau et al. 2014; Masao 2015; Lahr and Foley 2016; Leplongeon et al. 2017; Shipton et al. 2018; Tryon 2019).

Although the primary distinction between MSA and LSA assemblages has been established on the basis of changes in lithic technology, increases in frequency of a number of other elements of material culture have also been aligned with this transition (e.g. Tryon 2019). Ground stone tools appear during transitional sequences at Mumba, Nasera, and Kisese II (Mehlman 1989; Tryon et al. 2018; Tryon 2019), while bone tools demonstrate erratic early appearances (e.g. Pante et al. 2020) before increasing in frequency during the LSA (e.g. Langley et al. 2016; Shipton et al. 2018). The use of ochre is associated with the earliest MSA at Olorgesailie (Brooks et al. 2018) but becomes widespread only in the late MSA and LSA (Tryon 2019; D'Errico et al. 2020). Finally, the appearance of disk beads made from ostrich eggshell may be a true marker of the transition, with the earliest examples found in eastern Africa around 50 ka at Mumba and Magubike (Gliganic et al. 2012; Miller & Willoughby 2014). The earliest examples of engraved ostrich eggshell in eastern Africa date to ~43 ka at Goda Buticha and are associated with an MSA industry (Assefa et al. 2018). The transition between MSA and LSA has therefore been identified across a range of material classes, but the ubiquity of stone tools, and their durability in the archaeological record, provides a robust means to examine change through time that is less impacted by patterns of selective preservation.

In eastern Africa, the MSA first appears ~300ka at Olorgesailie (Brooks et al. 2018) and persists until the end of MIS 3, ~30ka (e.g. Ossendorf et al. 2019); the eastern African LSA first appears ~67ka at Panga ya Saidi and persists into the Holocene (Shipton et al. 2018). The chronological overlap between the two industrial complexes is therefore substantial, and it should be noted that individual 'LSA' technologies are by no means absent from MSA assemblages (Blinkhorn and Grove 2018), whilst some important MSA technologies persist within the LSA (e.g. Ranhorn and Tryon 2018;

Shipton et al. 2018). Mirroring Goodwin's (1946) cautions concerning the lack of clear-cut divisions, Ranhorn and Tryon et al. (2018) suggest that proportional rather than categorical differences may be critical, while Grove and Blinkhorn (2020) find that the use of co-occurring constellations of technologies rather than individual *fossiles directeurs* allows for robust discrimination between industrial complexes. Using machine learning algorithms, Grove and Blinkhorn (2020) demonstrate that the co-occurrence of Levallois flakes, retouched points, core tools, and scrapers is indicative of the MSA, whilst an alternative constellation of blades, backed pieces, and bipolar reduction signals the LSA.

The three Ages thus remain dominant but disputed, and as Robertshaw (1990:8) wryly notes, discussions of typology and nomenclature in African archaeology have 'often generated a great deal of heat but very little light'. The analyses reported below statistically test the validity of the division between assemblages labelled MSA and LSA using a combination of weighted binary logistic regression and permutation analysis. Whilst the primary aim is to assess the integrity of this particular division within the Stone Age of eastern Africa, the subsidiary aim is to provide a blueprint for the kind of analysis that might be used to test cultural taxonomic integrity in other periods and regions.

Methods

Data

The archaeological database used is that documented in Grove and Blinkhorn (2020), with the exception that the putative LSA assemblage from Nasera levels 4 and 5 is omitted. In the neural network study of Grove and Blinkhorn (2020) that sought to distinguish between LSA and MSA assemblages in eastern Africa, Nasera 4/5 was the only assemblage misclassified. Recent radiocarbon dates on ostrich egg shell beads obtained from stratigraphic positions above and below this assemblage by Ranhorn and Tryon et al. (2018) suggest that it is somewhat older than originally suggested by Mehlman (1989), and whilst chronological age is certainly not a valid proxy for industrial affiliation, both Ranhorn and Tryon et al. (2018) and Grove and Blinkhorn (2020) argue that this assemblage's LSA status is questionable. Further to this, when employing the typology used by Grove and Blinkhorn (2020), this assemblage is identical to Mumba UV 38, which is unequivocally MSA. The database employed below thus consists of 91 assemblages (LSA $n = 30$; MSA $n = 61$) evaluated on the basis of the presence or absence of 16 technologies (see Grove and Blinkhorn 2020 and Supplementary Materials for further details).

The 16 technologies used in the database were Backed Pieces, Bipolar Technology, Blade Technology, Borer, Burin, Centripetal Technology, Core Tool, Denticulate, Levallois Blade Technology, Levallois Flake Technology, Levallois Point Technology, Notch, Platform Core, Point Technology, RT Bifacial, and Scraper. These technologies were chosen following a comprehensive search of the literature and were amalgamated from various synonymous terms used in the literature by previous authors. The terms employed encompass the full breadth of terminology used to describe stone tool assemblages for Late Pleistocene eastern Africa. Although previous researchers have

in some cases employed different designations to refer to indistinguishable artefact forms (e.g. radial core as opposed to discoidal core), the amalgamation of such terms into a reduced taxonomy of 16 technologies goes a long way towards obviating this problem. The database utilizes existing classifications employed by the researchers who excavated or analysed a given assemblage; this is the case both for the designation of technocomplexes (i.e. 'MSA' or 'LSA') and for distinctions between multiple assemblages from the same site (e.g. Panga ya Saidi 5 or Panga ya Saidi 6). Although there may be differences in excavation and analytical techniques that lead to different concepts of what constitutes a distinct assemblage, the designations provided in the published literature provide the logical starting point for any subsequent analysis. Further details, including a comprehensive breakdown of synonymous terms, can be found in the Supplementary Materials. The locations of the assemblages used in the analyses are shown in Fig. 1.

Whilst the analyses presented below focus on differences in the technologies comprising LSA and MSA assemblages, it should be noted that differences in artefact size and in raw material use have also been suggested to distinguish between these two industrial complexes. Decreases in artefact size from the MSA to the LSA have been previously noted (e.g. Leakey et al. 1972; Eren et al. 2013; Tryon and Faith 2016; Shipton et al. 2018), with Pargeter and Shea (2019) stressing the significance of miniaturization as a trend through time. Decreases in artefact size have also been identified in sequences at individual sites (e.g. Shipton et al. 2018). In terms of raw material use, an increasing focus on more fine-grained materials and those that appear in smaller clast sizes has been documented (e.g. Leakey et al. 1972; Shipton et al. 2018). Nonetheless, the analyses of Grove and Blinkhorn (2020) demonstrate that technological shifts alone afford considerable discriminatory power; as the analyses presented below rely only on technological differentiation, they can be regarded as conservative in terms of their assessment of the integrity of the LSA/MSA division.

The basic hypothesis to be tested is that the division of these 91 assemblages into two classes—labelled LSA and MSA—is statistically valid in the sense that a model that distinguishes between these two classes can be obtained with a deviance lower than that obtained via alternative divisions of the data. This hypothesis can in fact be formulated and tested in both weak and strong forms. The weak form employs a standard *p*-value, such that validity is claimed when, for example, the probability of obtaining a deviance as low as that obtained for the LSA/MSA division in a permutation sample is less than one in twenty (equating to $\alpha = 0.05$). The null hypothesis in this case is that the LSA/MSA division is invalid because it leads to a model deviance that could have occurred at random with a relatively high probability. The strong form of the hypothesis is that the LSA/MSA division is valid in that it leads to a model deviance that is lower than that obtained via *any other possible division of the data*. In this case, the null hypothesis is simply that the LSA/MSA division is invalid because it is not the single best division of the data.

General statistical practice regards the strong hypothesis as overly conservative, and there are computational obstacles to testing this hypothesis precisely as stated. An exact permutation test would be required to assess the deviance of every other possible division in the data, and this is computationally intractable (including the split into LSA and MSA, there are $2^{91} \approx 2.48 \times 10^{27}$ possible divisions of the data). This is a common problem for permutation analyses, however, and a truly random sample of several

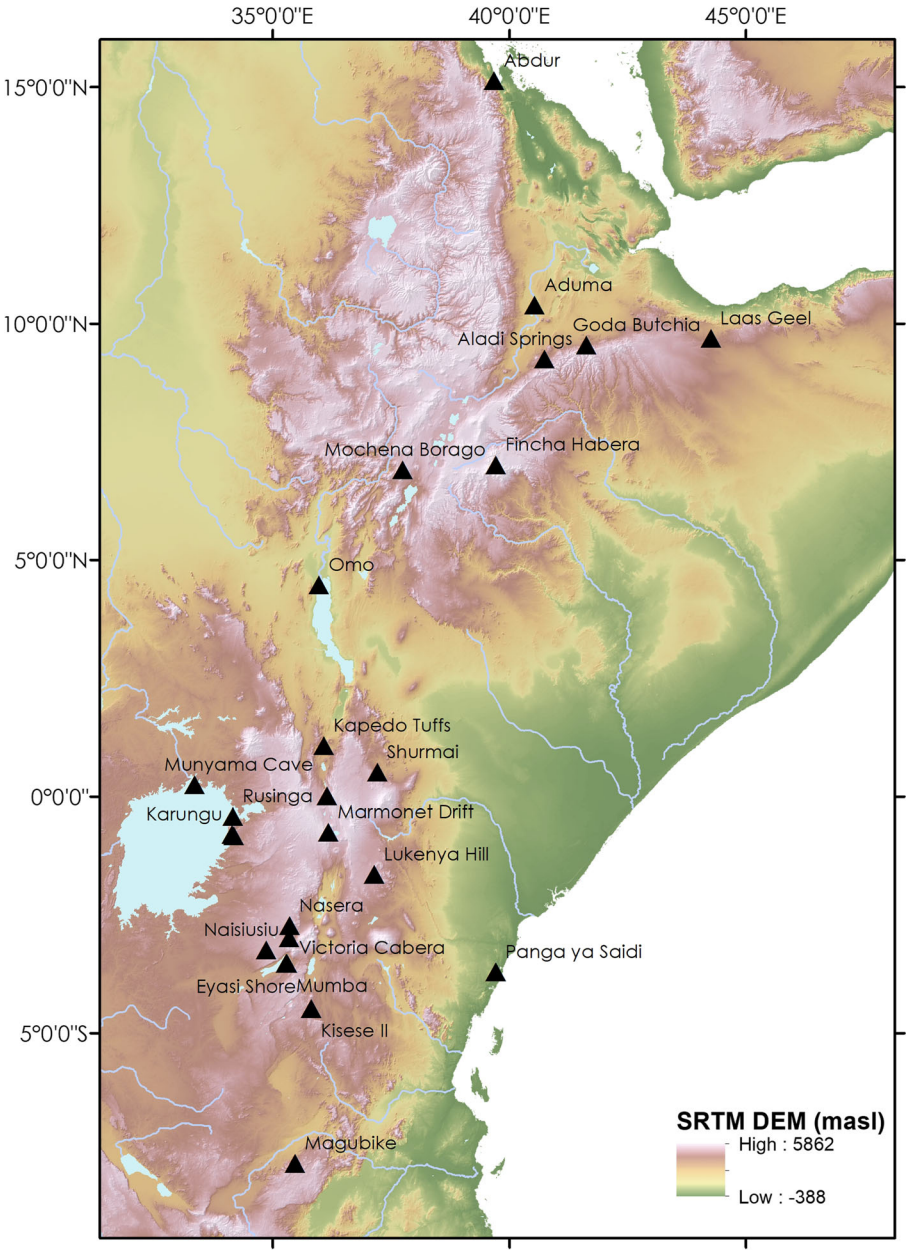

Fig. 1 The distribution of archaeological sites from which analysed assemblages derive, plotted on an SRTM (1 arc-second) DEM obtained from USGS earth explorer (https://earthexplorer.usgs.gov)

thousand permutations is normally regarded as sufficient (e.g. Ernst 2004). Since both weak and strong hypotheses can be assessed via the same permutation test (using alpha values of 0.05 and 1/(1 + the number of permutations), respectively), the results of both are discussed below.

Analyses

Weighted Binary log-F Regression

The analyses are built on the foundation of weighted binary logistic regression (henceforth, WBLR), a common statistical method for studying differences between classes. Weightings are applied both to account for differences in sample size between groups and—in the permutation study below—to accommodate the fact that there are several sub-groups of assemblages that appear identical under the typological scheme employed here. The basic weighting scheme ensures that the sums of weights for the two groups are equal; the weight for each assemblage in a given group, w_g, is given as $w_g = \frac{N}{2} n_g^{-1}$, where n_g is the number of assemblages in that group and N is the total number of assemblages in the analysis. This ensures that $\sum w_1 = \sum w_2$ and that $\sum w_1 + \sum w_2 = N$. For the example of LSA and MSA classification, LSA assemblages assigned the weight $w_{LSA} = \frac{91}{2} 30^{-1} \approx 1.517$ and MSA assemblages the weight $w_{MSA} = \frac{91}{2} 61^{-1} \approx 0.746$.

An additional weighting scheme is developed to account for the fact that certain groups of assemblages are identical under this typological scheme (see Table 1). Whilst this is not a problem for the basic WBLR analysis—and the use of this weighting scheme makes no difference to the results of that analysis—it is problematic for the permutation analysis that follows; for consistency, this additional weighting scheme is therefore employed throughout. Where a sub-group of assemblages are identical, that assemblage type is entered only once into the analyses, with a weighting that reflects the number of assemblages of that type. For example, Kisese II levels 3, 7, 9, 10, and 11 are identical; this assemblage type is entered once, with a weight five times that of a single LSA assemblage. Table 1 demonstrates that the 91 assemblages in the analysis fall into 65 types (20 LSA and 45 MSA); the table also shows a breakdown of the weighting scheme for these 65 types as used in the initial analysis.

Initial inspection of the data matrix and preliminary standard logistic regression runs demonstrated that the results suffer from a phenomenon known as 'separation' (Albert and Anderson 1984) or 'monotone likelihood' (Bryson and Johnson 1981). This is an instance of sparse data bias (Greenland et al. 2016) in which, although the likelihood appears to converge, the coefficients do not; it is immediately signalled by the presence of one or more coefficients that are effectively infinite (i.e. with an absolute value limited only by the number of iterations permitted by the analyst when minimizing the negative log-likelihood). In the current dataset, separation is caused primarily by the presence of categorical covariates with either very high or very low prevalence (i.e. tool forms that exist either in most assemblages or in very few assemblages). The output of standard logistic regression models under separation is essentially meaningless.

The issue of separation has been widely noted by statisticians (e.g. Bryson and Johnson 1981; Albert and Anderson 1984; Lesaffre and Albert 1989; Kolassa 1997; Heinze and Schemper 2002; Greenland et al. 2016; Mansournia et al. 2018), and a number of solutions have been suggested. Most of these focus on the concept of penalized logistic regression—a form of shrinkage estimation using weakly informative priors—and many derive from the initial work of Firth (1992, 1993). Though Firth's (1993) method has been reasonably widely adopted, it has been criticized on the basis

Table 1 Assemblages by group number with binary logistic regression weights

Assemblage	Type	Type weight	Industry	Industry weight	Final weight
Aladi Springs LSA	1	1	LSA	1.517	1.517
Kisese II 10	2		LSA	1.517	
Kisese II 3	2		LSA	1.517	
Kisese II 7	2		LSA	1.517	
Kisese II 9	2		LSA	1.517	
Kisese II 11	2	5	LSA	1.517	7.583
Kisese II 4	3		LSA	1.517	
Kisese II 5	3		LSA	1.517	
Kisese II 6	3		LSA	1.517	
Kisese II 8	3	4	LSA	1.517	6.067
Lukenya Hill GvJm16 B	4	1	LSA	1.517	1.517
Lukenya Hill GvJm22 E120 150	5	1	LSA	1.517	1.517
Mumba M III 77	6	1	LSA	1.517	1.517
Munyama Cave	7	1	LSA	1.517	1.517
Panga ya Saidi 5	8	1	LSA	1.517	1.517
Panga ya Saidi 6	9	1	LSA	1.517	1.517
Panga ya Saidi 7.5	10	1	LSA	1.517	1.517
Enkapune ya Muto DBL	11	1	LSA	1.517	1.517
Enkapune ya Muto GG	12	1	LSA	1.517	1.517
Panga ya Saidi 10	13	1	LSA	1.517	1.517
Panga ya Saidi 11	14	1	LSA	1.517	1.517
Panga ya Saidi 12	15	1	LSA	1.517	1.517
Panga ya Saidi 9	16		LSA	1.517	
Panga ya Saidi 15	16	2	LSA	1.517	3.033
Naisiusiu 1931	17		LSA	1.517	
Naisiusiu 1969 in situ	17	2	LSA	1.517	3.033
Naisiusiu 1972	18	1	LSA	1.517	1.517
Panga ya Saidi 13	19		LSA	1.517	
Panga ya Saidi 14	19	2	LSA	1.517	3.033
Panga ya Saidi 16	20	1	LSA	1.517	1.517
Lukenya Hill GvJm46	21	1	MSA	0.746	0.746
Enkapune ya Muto RBL4	22	1	MSA	0.746	0.746
Fincha Habera 8 10	23	1	MSA	0.746	0.746
Fincha Habera 8 11	24	1	MSA	0.746	0.746
Fincha Habera 8 8	25		MSA	0.746	
Fincha Habera 8 9	25		MSA	0.746	
Fincha Habera 9	25		MSA	0.746	
Panga ya Saidi 17	25	4	MSA	0.746	2.984
Goda Buticha 70 110	26	1	MSA	0.746	0.746
Karungu Kisaaka Main	27		MSA	0.746	
Karungu A3 Ex	27		MSA	0.746	
Karungu Kisaaka ZTG	27	3	MSA	0.746	2.238

Table 1 (continued)

Assemblage	Type	Type weight	Industry	Industry weight	Final weight
Kisese II 18	28		MSA	0.746	
Kisese II 21	28	2	MSA	0.746	1.492
Kisese II 19	29		MSA	0.746	
Kisese II 20	29	2	MSA	0.746	1.492
Laas Geel SU 711	30	1	MSA	0.746	0.746
Lukenya Hill GvJm22 F170 205	31	1	MSA	0.746	0.746
Magubike MSA	32	1	MSA	0.746	0.746
Mochena Borago Lower T	33	1	MSA	0.746	0.746
Mochena Borago R Group	34	1	MSA	0.746	0.746
Mochena Borago S Group	35	1	MSA	0.746	0.746
Mochena Borago Upper T	36	1	MSA	0.746	0.746
Mumba L III 38	37		MSA	0.746	
Nasera 12 17	37		MSA	0.746	
Mumba U VI A	37		MSA	0.746	
Mumba L VI A	37		MSA	0.746	
Mumba VI B	37	5	MSA	0.746	3.730
Mumba L V 81	38	1	MSA	0.746	0.746
Mumba L VI 38	39		MSA	0.746	
Nasera 6 7	39	2	MSA	0.746	1.492
Mumba MU V 81	40		MSA	0.746	
Nasera 8/9 11	40	2	MSA	0.746	1.492
Mumba U V 38	41	1	MSA	0.746	0.746
Rusinga Nyamita	42	1	MSA	0.746	0.746
Shurmai MSA	43	1	MSA	0.746	0.746
Abdur N C S	44	1	MSA	0.746	0.746
Aduma A1	45	1	MSA	0.746	0.746
Aduma A4C	46	1	MSA	0.746	0.746
Aduma A5Ex	47		MSA	0.746	
Aduma A5 Ex Surf	47	2	MSA	0.746	1.492
Aduma A8	48	1	MSA	0.746	0.746
Aduma A8AC	49		MSA	0.746	
Aduma A8AG	49	2	MSA	0.746	1.492
Aduma A8A Surf	50	1	MSA	0.746	0.746
Aduma A8B	51	1	MSA	0.746	0.746
Aduma VP1/1	52	1	MSA	0.746	0.746
Aduma VP1/3	53	1	MSA	0.746	0.746
Eyasi Shore 77 81	54	1	MSA	0.746	0.746
Eyasi Shore N surface	55	1	MSA	0.746	0.746
Eyasi Shore W in situ	56	1	MSA	0.746	0.746
Eyasi Shore W surf	57	1	MSA	0.746	0.746
Kapedo Tuffs	58	1	MSA	0.746	0.746
Marmonet Drift H4	59		MSA	0.746	

Table 1 (continued)

Assemblage	Type	Type weight	Industry	Industry weight	Final weight
Marmonet Drift H5	59	2	MSA	0.746	1.492
Omo BNS L3	60	1	MSA	0.746	0.746
Omo BNS<50m	61	1	MSA	0.746	0.746
Panga ya Saidi 18	62	1	MSA	0.746	0.746
Panga ya Saidi 19	63	1	MSA	0.746	0.746
Victoria Cabrera 2	64	1	MSA	0.746	0.746
Victoria Cabrera 2a	65	1	MSA	0.746	0.746

Group weight is the number of identical assemblages in each group, IC weight is the weight for an assemblage belonging to that industrial complex, and final weight is the product of the two previous weights

that it artificially shrinks the constant, clouds the interpretation of coefficients and odds ratios, and fails to account for possible correlations in the prior (Gelman et al. 2008; Greenland and Mansournia 2015; Rahman and Sultana 2017). The analyses below therefore employ the log-F method proposed by Greenland and Mansournia (2015), using a weakly informative prior proportional to the log of the F-distribution. This method displays all the benefits of Firth's method whilst minimizing bias; crucially, it does not include the constant in the calculation of the penalty term (Greenland and Mansournia 2015; Rahman and Sultana 2017; Mansournia et al. 2018).

Formally, the penalized log-likelihood function to be minimized in log-F regression is

$$PL(\beta) = -L(\beta) - P(\beta) \tag{1}$$

where β is the vector of coefficients (including the constant as the last coefficient). $L(\beta)$ is the standard negative weighted log-likelihood,

$$-L(\beta) = -\sum_i w_i y_i \ln(\pi_i) + w_i(1-y_i)\ln(1-\pi_i) \tag{2}$$

where w are the weights, y are the values of the dependent variable (0 for LSA or 1 for MSA), and π are the estimates of the dependent variable produced by the model with coefficients β. $P(\beta)$ is the log-F penalty term given by

$$P(\beta) = \sum_{j=1}^{n-1} \frac{m\beta_j}{2} - m\ln\left(1 + e^{\beta_j}\right) \tag{3}$$

where n is the number of coefficients in the model (i.e. the length of the vector β) and m gives the degrees of freedom of the prior. Following the recommendations of Greenland and Mansournia (Greenland and Mansournia 2015; see also Rahman and Sultana 2017), here $m = 1$. Note that the penalty is *not* applied to the nth coefficient (the constant term), as penalizing the constant can introduce exactly the form of bias for which Firth regression has been criticized (Greenland and Mansournia 2015). It is possible to carry out log-F regression via data augmentation for individual analyses

(e.g. Discacciati et al. 2015); however, given the nature of the permutation tests described below, it is computationally more efficient in this case to directly minimize the result of Eq. (1).

The most important overall measure of fit for a logistic regression model is the deviance (= −2 × log-likelihood); for the initial regression model, the log-likelihood, penalized log-likelihood, sample-size corrected Akaike's Information Criterion (AICc; Burnham et al. 2011) and the Cox and Snell, Nagelkerke, and count pseudo-R^2 statistics are also reported. The Cox and Snell R^2 (R^2_{CS}) is appropriate as, like the deviance, it assesses the fit of the full model relative to that of the null (intercept only) model (Maddala 1983; Cox and Snell 1989). The $R^2_{C\&S}$, however, has a maximum attainable value of less than one; Nagelkerke's (1991) correction (R^2_{Nag}) re-scales it by the null model likelihood to give it a range of possible values between zero and one. The count R^2 (R^2_{Co}) is simply the number of cases correctly classified divided by the total number of cases and is useful when assessing the classificatory ability of a model. For the initial regression model, the values of the individual coefficients and their likelihood ratio statistics are also reported; likelihood ratio statistics are preferred over the simpler Wald statistics as they are more reliable when dealing with small sample sizes (e.g. Agresti 2007:11ff.), particularly when dealing with the results of penalized logistic regression (Greenland et al. 2016). Whilst the production and assessment of regression coefficients is not the primary aim of this study, assessing the significance of the coefficients in relation to the results of Grove and Blinkhorn (2020) on significant predictors obtained via neural network analyses provides a useful comparison of these two methods.

Permutation Analysis

In order to assess the validity of LSA/MSA division, a permutation test was performed to compare this division to a random subset of other possible divisions of the data. Each permutation was carried out by randomly assigning the 65 assemblage types to two groups, performing a log-F WBLR on those two groups, and recording the resulting model deviance. Results of logistic regression can be imprecise and biased towards the larger group if the smaller group is too small; weighting goes some way to addressing this problem, but the fact remains that highly imbalanced sample sizes can lead to meaningless results. To ensure a range of sample sizes for the two groups (whilst ensuring that the size of neither group became trivially small), the sample size of the first group one was called from an integer-rounded probability distribution, with the size of the second group set equal to 65 minus the size of the first group. To minimize the possible effects of imbalanced sample sizes, two probability distributions were used in two different permutation exercises:

1. A triangular distribution with a minimum of 15, a maximum of 50, and a mean of 65/2
2. A uniform distribution with a minimum of 15 and a maximum of 50

If low deviances tend to occur more frequently in imbalanced models, 2 would be expected to produce a greater frequency of lower deviance results. This potential bias

was further tested by examining correlations between the deviance of a model and the sample size of the smaller group; if greater sample size discrepancies between the two groups lead to lower log-F WBLR deviances, these correlations will be positive and significant.

Prior to each log-F WBLR, weights were adjusted such that the sums of weights for the two groups were equal to 91/2. Sample sizes for the two groups can therefore vary from 15 to 50, but sums of weights for the two groups remain identical at 91/2 in each permutation. The second weighting procedure described above (dividing the dataset into 65 weighted assemblage types rather than 91 individual assemblages) is particularly important to the results of the permutation test. Without this procedure, identical assemblages could be permuted into different groups, automatically increasing the deviance of the resulting log-F WBLR. Assessing model fits to types of assemblages rather than individual assemblages ensures the results regarding the integrity of the LSA/MSA split are as conservative as possible. Estimated *p*-values for the significance of the LSA/MSA split are given by

$$\widehat{p} = \frac{1 + \sum_{i=1}^{R} I\left(d_i \leq d^*\right)}{1 + R} \tag{4}$$

where R is the number or permutations of the data, d_i is the deviance of the log-F WBLR model fitted to the ith permutation, d^* is the deviance of the original log-F WBLR model, and I is an indicator function that equals 1 if $d_i \leq d^*$ and 0 otherwise (Grove and Pearson 2014). R was set to 99,999 permutations, yielding a minimum attainable *p*-value of 0.00001. Both the log-F WBLR and permutation procedures were written as custom scripts in Matlab R2019b (Mathworks, Natick, MA, USA) and are included as supplementary materials.

Results

Initial Analysis

The initial WBLR model had a deviance of 40.375 and was highly significant (relative to the null (intercept-only) model, $\chi^2(16,65) = 96.471$, $p < .001$, null deviance = 136.846). The AICc value for the full model was 87.396, relative to 138.909 for the null model. The pseudo-R^2 statistics were $R^2_{C\&S} = .773$, $R^2_{\text{Nag}} = .881$, and $R^2_{\text{Co}} = .892$; the latter implies that seven of 65 assemblage types were misclassified. Of the seven misclassified assemblage types, six consisted of single assemblages whilst one consisted of two assemblages; thus, eight assemblages were misclassified in total, leading to an overall accuracy for individual assemblages of 83/91 = .912. The accuracy achieved is lower than the 91/92 = .989 achieved using neural networks by Grove and Blinkhorn (2020), but this is to be expected as WBLR is a less sophisticated classification model. The incorrectly classified assemblages were Mumba M III 77 and Panga ya Saidi 11 (LSA misclassified as MSA) and Lukenya Hill GvJm46, Enkapune ya Muto RBL4, Mumba L V 81, Marmonet Drift H4, Marmonet Drift H5, and Laas Geel SU 711 (all MSA misclassified as LSA). A graphical summary of the regression output is shown in Fig. 2.

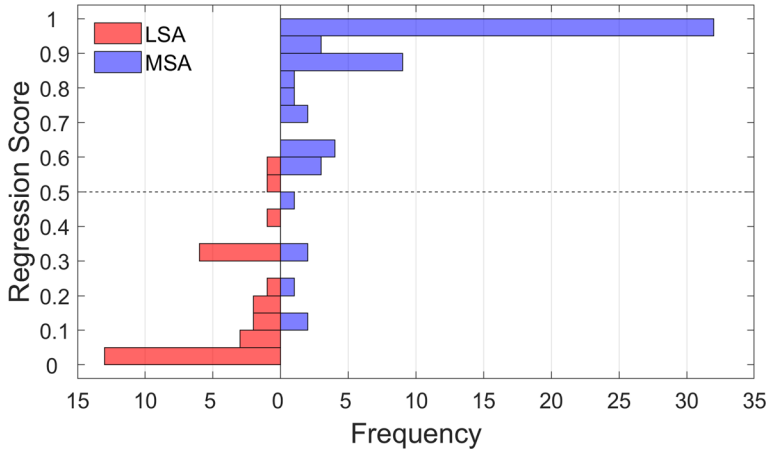

Fig. 2 Assemblage frequencies plotted at binned regression scores for LSA and MSA assemblages. A regression score of less than 0.5 indicates an LSA classification via the logistic regression, with a regression score of greater than 0.5 indicating an MSA classification; as such, blue (MSA) bars with scores less than 0.5 represent MSA assemblages misclassified as LSA and red (LSA) bars with scores greater than 0.5 represent LSA assemblages misclassified as MSA

Coefficients for individual technologies and their likelihood ratio statistics are given in Table 2. Significant coefficients were found for backed pieces, bipolar technology, blade technology, Levallois flake technology, and point technology. Signs of the coefficients demonstrate that the former three technologies are associated with LSA assemblage types whereas the latter three are associated with MSA assemblage types. These results agree with those of Grove and Blinkhorn (2020), with the exception that the latter study also suggested the presence of core tools and scrapers as predictors of MSA assemblages.

Permutation Analysis

The primary goal of this study was to assess the validity of the division of these assemblage types into the widely adopted categories of LSA and MSA. The results of the permutation test are shown in Fig. 3. Using the triangular distribution of group sizes, six of the 99,999 permuted divisions resulted in WBLR models that returned deviance values less than or equal to that of the LSA/MSA division, yielding $\widehat{p}_t$ = 0.00007. Using the uniform distribution, the equivalent figure was 18 of 99,999, yielding $\widehat{p}_u$ = 0.00019. The division of these assemblage types into LSA and MSA is thus highly significant by traditional statistical standards, suggesting that these labels provide a valid classificatory scheme for this material. More nuanced interpretations of this result are possible, however, and are discussed in detail below.

Correlations between the sample size of the smallest group and WBLR model deviance were positive and significant in both cases (triangular, $r(99{,}997) = 0.192$, $p < 0.001$; uniform, $r(99{,}997) = 0.294$, $p < 0.001$), demonstrating that models with greater sample size imbalance produce lower deviances. Overall, 3.29% of permutation models in the triangular analysis and 12.94% of permutation models in the uniform analysis were more imbalanced than the empirical model. Of the permutation models

Table 2 Logistic regression coefficients for individual technologies and associated likelihood ratio statistics; * denotes significance at $\alpha = 0.05$

Technology	*B*	log(L)	$\chi 2$	*p*
Backed pieces	−2.955	−24.042	7.709	0.005*
Bipolar technology	−2.474	−22.847	5.320	0.021*
Blade technology	−1.961	−22.164	3.953	0.047*
Borers	0.135	−20.297	0.219	0.640
Burins	0.473	−20.515	0.655	0.418
Centripetal technology	0.532	−20.586	0.798	0.372
Core tools	1.779	−21.511	2.646	0.104
Denticulates	−0.415	−20.284	0.193	0.661
Levallois blade technology	0.221	−20.195	0.016	0.900
Levallois flake technology	2.099	−24.558	8.741	0.003*
Levallois point technology	−0.486	−20.450	0.524	0.469
Notches	0.901	−20.863	1.351	0.245
Platform cores	−1.425	−21.716	3.057	0.080
Point technology	3.053	−28.517	16.658	0.000*
Bifacial retouch	0.982	−20.420	0.464	0.496
Scrapers	1.165	−21.083	1.791	0.181
{Constant}	1.855	−21.590	2.804	0.094

Log-likelihoods given are those for reduced models in which each technology in turn is omitted.

demonstrating lower deviance than the empirical model, 77.78% were more imbalanced than the empirical model when using the uniform distribution, but none were more imbalanced than the empirical model when using the triangular distribution. These results suggest that 14 of the permutation models that returned lower deviances than the empirical model when using the uniform distribution may have done so simply because they were more imbalanced; overall, however, there were at least ten models (four generated by the uniform distribution and six by the triangular distribution) that were better than the empirical model and could not be explained by statistical artefacts.

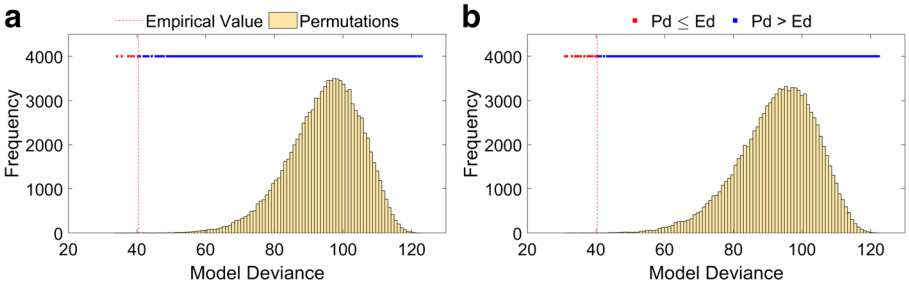

Fig. 3 Results of permutation tests using (**a**) a triangular distribution and (**b**) a uniform distribution for group sample size. Red squares show permutations producing WBLR model deviances less than or equal to the empirical model deviance; blue squares show permutations producing WBLR model deviances greater than the empirical model deviance

Discussion

The results of the weighted binary logistic regression reported above agree substantively with those of Grove and Blinkhorn (2020) in that backed pieces, blades, and bipolar reduction are seen as indicative of LSA assemblages whilst Levallois flakes and points are seen as indicative of MSA assemblages. Grove and Blinkhorn (2020) also found core tools and scrapers to be indicative of the MSA; in the current study, both are found to be more associated with the MSA than the LSA, but not significantly so. In relation to scrapers, it is worth noting Tryon's (2019:267) finding that end scrapers are found more often in LSA contexts, with side scrapers more prevalent during the MSA.

These results also broadly agree with the intuitions of previous researchers regarding the associations of these technologies with the respective industrial complexes. Technologies that are indicative of each industrial complex, however, also occasionally appear in the other, recalling Goodwin's (1946) point that there is considerable overlap between them and agreeing with Tryon's (2019) recent description of the eastern African transition as a prolonged process with varying regional trajectories. It is therefore important, as per Grove and Blinkhorn (2020), to recognize constellations of indicative technologies rather than individual tool forms when discussing the dynamics of the transition.

The analyses undertaken here aimed to assess the validity of the MSA/LSA division, but did not assess whether each individual assemblage was 'correctly' classified to one of these two industrial complexes; without detailed examination of each and every assemblage, the policy of adopting the designation provided by the excavators in each case is clearly the only sensible one to follow. Similarly, the analyses reported here are dependent upon the excavators' use of terminology for identification of the different technologies. Whilst only further archaeological study can robustly re-assign assemblages to alternative industrial complexes, statistical results can be informative concerning which assemblages might be prioritized for re-examination. An experiment in which each of the eight misclassified assemblages in turn was reclassified and the models re-calculated—with appropriate changes to all weightings—led to the results shown in Table 3.

As expected, the above experiment suggests that, were any of the eight misclassified assemblages reclassified, reductions relative to the original model deviance of 40.375 could be achieved. Most of these reductions are relatively minor, however, and it is important to note that at this scale the deviance does not necessarily correlate with the number of assemblages misclassified. Whilst reclassification of Mumba M III 77 would lead to the greatest reduction in deviance, reclassification of either Lukenya Hill GvJm46 or Mumba L V 81 would lead to the greatest improvement in the number of correct assemblage classifications. Any reclassifications could only take place, of course, after careful archaeological examination of the assemblages in question.

There are numerous cultural, stratigraphic, taphonomic, chronological, and methodological factors that might either prompt re-investigation or suggest why a given assemblage is not fully indicative of the industrial complex to which it is attributed. To take Mumba as an example, Mehlman's (1977) excavations (Mehlman 1979, 1989) were intended to address issues with the original excavations by Kohl-Larsen (Kohl-Larsen 1943). Nonetheless, he was only able to retrieve relatively limited samples (Mehlman 1989:78), and many of these remain unstudied (Prendergast et al. 2007).

Table 3 Statistics obtained by reclassifying the assemblages misclassified by the logistic regression analysis reported above and re-calculating the model

Assemblage	IC	log(L)	Deviance	N MisC	AICc	Δ	Weight	Prob
Mumba M III 77	LSA	−16.454	32.908	7	75.291	0.000	1.000	0.467
Lukenya Hill GvJm46	MSA	−17.339	34.677	5	77.061	1.770	0.413	0.193
Panga ya Saidi 11	LSA	−17.676	35.353	6	77.736	2.445	0.294	0.138
Enkapune ya Muto RBL4	MSA	−18.065	36.129	7	78.513	3.222	0.200	0.093
Mumba L V 81	MSA	−18.674	37.348	5	79.732	4.440	0.109	0.051
Marmonet Drift H4	MSA	−19.630	39.260	6	81.644	6.352	0.042	0.020
Marmonet Drift H5	MSA	−19.630	39.260	6	81.644	6.352	0.042	0.020
Laas Geel SU 711	MSA	−19.670	39.341	8	81.724	6.433	0.040	0.019

IC industrial complex, *NMisC* number misclassified, *Δ* AICc difference (i.e. AICci – AICcmin), *RL* relative likelihood, *Prob* relative probability of each model given the data and the set of models considered. Assemblages are ranked by AICc

Subsequent studies have focused on the transitional nature of Mehlman's (1989) Mumba Industry, located primarily in the Bed V horizons of the site, and on more comprehensive dating of the deposits so as to recognize patterns of change and innovation (Mabulla 2007; Prendergast et al. 2007; Diez-Martin et al. 2009; Gliganic et al. 2012; Bushozi et al. 2020). If the Middle Bed III samples recovered by Mehlman in 1977 are MSA, they would be stratigraphically and chronologically anomalous, particularly given the results of Gliganic and colleagues (Gliganic et al. 2012; see also Diez-Martin et al. 2009; Eren et al. 2013) who argue for a relatively early LSA associated with abundant ostrich eggshell beads beginning in Upper Bed V at 49.1 ± 4.3 ka. A realistic explanation for the effect of the Mumba M III 77 assemblage on the above analyses, therefore, is that it is a relatively small assemblage that is not fully indicative of its LSA provenance.

he misclassifications of some other assemblages, such as Lukenya Hill GvJm46 and Enkapune ya Muto RBL4, may be due to the fact that previously published inventories rely on partial samples from selected test pits, from depositional contexts that are not clearly established, or from sparse occupations that may not have resulted in extensive or indicative lithic accumulations (e.g. Miller & Willoughby 2014; Kelly 1996:271; Ambrose 1998:384). Ideally, future analyses would consider individually the various processes that act in combination to generate archaeological samples; in practice this is rarely possible, but a valid (albeit post hoc) alternative would be to subject those assemblages that have been misclassified in the above analyses to further investigation in relation to such processes.

As highlighted above, these analyses and their results depend upon the collation of data previously published by numerous researchers. The database therefore inevitably encompasses differences not only in the terminology used to describe individual lithics but also in the techniques employed in excavation and recording. Excavation by context, for example—where 'context' is defined as a homogenous unit of the matrix, regardless of its vertical or horizontal extent—leads to a different concept of 'assemblage' than does excavation by regular, arbitrary spits. Ideally, an assemblage—however, defined in terms of the sedimentary matrix—would equate to a discrete

occupation horizon, but of course this is rarely the case. The amalgamation of synonyms into a broad typology of 16 technologies largely removes concerns about the inconsistent use of terminology for individual lithics, but inconsistencies in the delineation of assemblages remain. Such inconsistencies are unavoidable in a study of this kind—after all, the material cannot be excavated again—and in the current study, they do not appear to introduce any systematic bias in terms of the results. For example, assemblages defined by arbitrary spits (or groups thereof) are no more likely to be misclassified than those defined by archaeological contexts. This issue does, however, starkly reveal the fact that the problems facing archaeological taxonomies act at multiple scales.

The taxonomy of individual lithics has been criticized on the basis that it discretizes the continuous variation produced either by a reduction continuum or by spatio-temporal variation in the cultural production of functionally equivalent tools (e.g. Davidson and Noble 1993; Davidson 2002). The process of dividing excavated material into assemblages—except in those rare cases where such assemblages are bracketed by sterile layers—is a second process by which continuous variation is discretized. Finally, assemblages are categorized by technocomplex, which further masks the continuity between them. Most archaeological analyses, therefore, depend on various, cumulative methods of discretization; comparisons between periods or between regions cannot be accomplished without the application of such methods. The resulting analyses are often genuinely valuable, but archaeologists must also remain cognizant of the limitations the underlying methods impose.

The results of the permutation analyses reported above suggest that the LSA/MSA division is valid based on a standard statistical criterion (i.e. $\alpha = 0.05$, 0.01, or even 0.001), but that it is not the single best division of the data; thus, the weak form of the hypothesis is supported, but the strong form is not. The history of archaeology as a largely descriptive discipline, with quantitative hypothesis testing emerging as a significant component long after the establishment of our cultural taxonomies, leads to a situation in which statistical analyses are being used as post hoc tests of those taxonomies (see also Ivanovaite et al. 2020). On the one hand, this is regrettable, but on the other, it is important to stress that the meaning—and therefore the usefulness—of our taxonomies must emerge from archaeological rather than from statistical reasoning. Ideally the two would be complementary, but the complexity and paucity of the archaeological record often stifle this alliance.

As an example of why archaeological reasoning must take precedence in such cases, Fig. 4 shows a (purely theoretical) series of assemblages plotted in two dimensions; these dimensions could be counts of two tool forms, or more realistically the first two axes of a principal components analysis. The two dashed lines in the figure are both examples of complete separation in this two-dimensional space; that is, a binary logistic regression or linear discriminant analysis could perfectly separate the data into two groups to either side of either of these lines based purely on the two axes shown (vertical or horizontal divisions would only need one axis to do so). Yet, there are any number of additional lines that could also achieve such separation; all would be statistically equivalent, but would any be archaeologically meaningful?

The above is an example of an unsupervised analysis, in which patterns are sought in the data without prior labelling; a complete overhaul of archaeological cultural taxonomy would necessarily be built on the results of such analyses. The WBLR reported

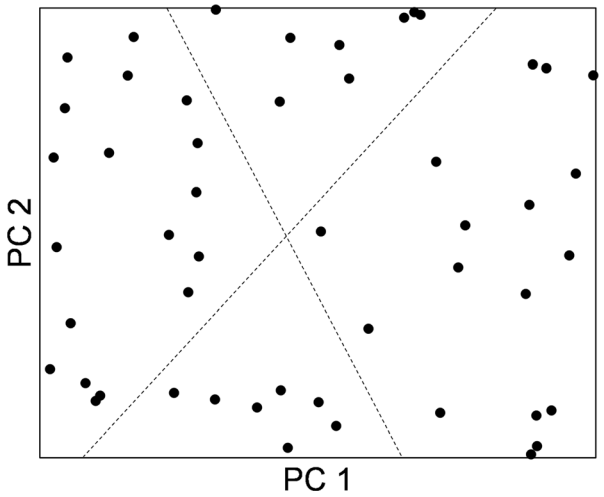

Fig. 4 Theoretical plot of the first two axes of a principal components analysis on a group of archaeological assemblages. Dashed lines represent two possible examples of complete separation

above is a supervised analysis, in which coefficients are sought that divide the data as well as possible into categories to which they have been assigned a priori. The permutation analyses undertaken here exist in the space between supervised and unsupervised analyses, as each permutation returns the result of a supervised analysis in which membership of the a priori categories is assigned at random. Archaeology is currently in a position whereby revision of existing cultural taxonomies is likely to be more beneficial than building those taxonomies anew from the ground up; as Reynolds and Riede (2019b:1369) state, 'there is structure in the archaeological record, and abandoning taxonomies altogether would limit… the types of questions that we can ask'. In this context, further exploration of the space between supervised and unsupervised analyses is likely to prove useful.

The problems of cultural taxonomy discussed here are certainly not limited to archaeological endeavours. For example, some British architectural styles correspond broadly to chronological periods, but these styles frequently overlap, and their specific durations are disputed, even when their labels derive from correspondence to the reigns of particular monarchs (e.g. Victorian, Edwardian). The differences between Victorian and Edwardian domestic architecture (fewer storeys, higher ceilings, broader hallways behind wood-framed porches in the latter) are fewer than their similarities; as such, much like lithic industries or biological species, they grade into one another when viewed from a sufficient chronological distance. With the exception of architectural styles that consciously derive inspiration from previous periods (e.g. Neoclassical), labels are applied post hoc in much the same way that they are in archaeological systematics. 'Edwardian' architects did not set out to create a distinctly 'Edwardian' style as a counterpoint to the previous 'Victorian' style; instead, differences can only be discerned in hindsight by scholars working in later periods. Nonetheless, these labels act as useful heuristic devices for the discussion of changing architectural styles through time, serving much the same purpose as our archaeological nomenclature. If the labels did not exist, the discussion could not proceed, and this would be detrimental

not only to systematics itself but also to broader understanding. Attempts to simply *abandon* existing cultural taxonomies—in archaeology as in any other discipline—are therefore entirely without value; attempts to *revise* existing taxonomies must be grounded in first-hand re-examination and logical assessment of affinities between large numbers of assemblages (see Shea 2020 for a recent example). Current archaeological taxonomy may resemble a 'house of cards' (Reynolds and Riede 2019a), but it would be premature to pull this house down before a new one has been built.

The analyses carried out above examine assemblages at the scale of industrial complexes and do so by recording the presence or absence of 16 technologies within each assemblage. This is a relatively common approach to the African Stone Age record (e.g. Tryon and Faith 2013; Blinkhorn and Grove 2018; Grove and Blinkhorn 2020; Shea 2020) but clearly operates at a very different scale to analyses of, for example, metric attributes of individual tool forms (e.g. O'Brien et al. 2014; Ivanovaite et al. 2020). Different questions demand different scales of analysis, and it is often the case that analyses at finer scales can only proceed by deliberately ignoring patterning at coarser scales. To employ a biological example, traits that distinguish genus one from genus two are unlikely to be useful in distinguishing between two species that both belong to genus two because the attribution of those species to genus two necessarily implies that they both display those traits. These shared or 'primitive' traits are of no use in pursuing the finer-scale division between species. In much the same way, it may be feasible to support broad scale archaeological cultural taxonomies (e.g. MSA, LSA) whilst simultaneously questioning their subdivisions (e.g. Nubian; see Groucutt 2020).

Perhaps the most substantive problem with existing archaeological cultural taxonomy stems from the way in which it is interpreted and used. This stems from a lasting culture-historical legacy that equates particular groups of artefacts or artefact types with particular groups of people; the 'culture' of a people is explicitly manifest in the material culture assemblages those people produce, and the assemblages therefore indicate the people. In this regard, Kleindienst (2006:17) argues that the Burg Wartenstein recommendations were 'fatally flawed' because 'those in favour of such a system could not persuade their colleagues… to leave the "group of prehistoric people" out of the definition of the "Basic Unit"' (i.e. the 'Industry' as defined in Bishop and Clark (1967:893)). The idea that 'cultures' in this sense are immutable and inextricably linked to groups of people permits migration and diffusion but ignores both the ability of hominin actors to flexibly respond to changing circumstances and the possibility of convergent responses of different temporally or geographically distant groups to similar circumstances.

A stark alternative to the 'group of prehistoric people' perspective sees the production of material culture primarily as a functional reaction to ecological circumstances and provides markedly different interpretations of the same datasets (e.g. Bordes 1961; Binford and Binford 1966; Bordes and De Sonneville-Bordes 1970; Binford 1973). If a recurring assemblage of archaeological material is the physical manifestation of a distinct set of ideas belonging to a distinct group of people, then archaeological analyses tell us about the spatio-temporal history of that group of people; but if the same recurring assemblage represents just a subset of a group's material repertoire, and if different groups employ the similar subsets when encountering similar circumstances, then analyses tell us more about the circumstances these groups encountered than about their social norms or cultural values. If one adopts the latter position, a

further difficulty arises in the need to disaggregate those aspects of the assemblage (and of individual artefact form) that serve a direct subsistence function from those that serve a social, symbolic, or otherwise cultural function (e.g. Dunnell 1978; Brantingham 2007). Any taxonomy—cultural or otherwise—is constructed in reference to a particular analytical goal, with results interpreted in relation to a particular theoretical position.

Conclusions

The analyses presented above sought to test the integrity of the cultural taxonomic division between MSA and LSA assemblages in eastern Africa by comparing that division to a large sample of arbitrary divisions of the same data. Results suggest that the division is valid on the basis of any routinely employed statistical criterion, but that it is not the single best division of the data. These results invite questions about what archaeologists seek to achieve via cultural taxonomy and about the analytical methods that should be employed when attempting to revise existing nomenclature. Quantitative analyses are necessarily more robust than their purely descriptive counterparts but will only prove truly useful if their results can be interpreted in archaeologically meaningful ways. Archaeologists seek information about similarities and differences that characterize assemblages that originated in different periods and regions or that were produced under different environmental regimes. Such similarities and differences—where they occur—are often highly complex, existing at different scales and along multiple axes of variation. The sheer variety of hominin behaviour precludes simple classification, but classification is essential to discussion. Archaeological cultural taxonomies are largely heuristic devices, but they remain valuable, and—at least in the case of the eastern African MSA and LSA—they map onto important differences in the stone tool assemblages created by our ancestors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41982-021-00087-4.

Availability of Data and Material All data used are provided as Supplementary Materials.

Code Availability All code used is provided as Supplementary Materials.

Author Contribution The authors contributed equally to this work.

Funding This study was financially supported by the Natural Environment Research Council Grant NE/K014560/1.

Declarations

Competing Interests The authors declare no competing interests.

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken: Wiley.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum-likelihood estimates in logistic regression models. *Biometrika, 71*(1), 1–10.

Ambrose, S. H. (1998). Chronology of the later Stone Age and food production in East Africa. *Journal of Archaeological Science, 25*(4), 377–392. https://doi.org/10.1006/jasc.1997.0277.

Assefa, Z., Asrat, A., Hovers, E., Lam, Y., Pearson, O., & Pleurdeau, D. (2018). Engraved ostrich eggshell from the Middle Stone Age contexts of Goda Buticha, Ethiopia. *Journal of Archaeological Science-Reports, 17*, 723–729. https://doi.org/10.1016/j.jasrep.2017.12.035.

Binford, L. R. (1973). Interassemblage variability: The Mousterian and the 'functional' argument. In C. Renfrew (Ed.), *The Explanation of Culture Change: Models in Prehistory* (pp. 227–254). London: Duckworth.

Binford, L. R., & Binford, S. R. (1966). A preliminary analysis of functional variability in the Mousterian of Levallois facies. *American Anthropologist, 68*, 238–295.

Bishop, W. W., & Clark, J. D. (Eds.). (1967). *Background to Evolution in Africa*. Chicago: University of Chicago Press.

Bisson, M. S. (2000). Nineteenth century tools for twenty-first century archaeology? Why the Middle Paleolithic typology of Francois Bordes must be replaced. *Journal of Archaeological Method and Theory, 7*(1), 1–48. https://doi.org/10.1023/a:1009578011590.

Blinkhorn, J., & Grove, M. (2018). The structure of the Middle Stone Age of eastern Africa. *Quaternary Science Reviews, 195*, 1–20. https://doi.org/10.1016/j.quascirev.2018.07.011.

Bordes, F. (1961). Mousterian Cultures in France. *Science, 134*, 803–810.

Bordes, F., & De Sonneville-Bordes, D. (1970). The significance of variability in Paleolithic assemblages. *World Archaeology, 2*, 61–73.

Brantingham, P. J. (2007). A unified evolutionary model of archaeological style and function based on the price equation. *American Antiquity, 72*(3), 395–416. https://doi.org/10.2307/40035853.

Brooks, A. S., Yellen, J. E., Potts, R., Behrensmeyer, A. K., Deino, A. L., Leslie, D. E., et al. (2018). Long-distance stone transport and pigment use in the earliest Middle Stone Age. *Science, 360*(6384), 90–94. https://doi.org/10.1126/science.aao2646.

Bryson, M. C., & Johnson, M. E. (1981). The incidence of monotone likelihood in the Cox model. *Technometrics, 23*(4), 381–383. https://doi.org/10.2307/1268228.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology, 65*(1), 23–35. https://doi.org/10.1007/s00265-010-1029-6.

Bushozi, P. M., Skinner, A., & de Luque, L. (2020). The Middle Stone Age (MSA) technological patterns, innovations, and behavioral changes at bed VIA of Mumba rockshelter, northern Tanzania. *African Archaeological Review, 37*(2), 293–310. https://doi.org/10.1007/s10437-019-09360-y.

Clark, G. (1969). *World Prehistory*. Cambridge: Cambridge University Press.

Clark, G., & Lindly, J. M. (1991). On paradigmatic biases and Paleolithic research traditions. *Current Anthropology, 32*, 577–587.

Clark, J. D. (Ed.). (1957). *Preceedings of the Third Pan-African Congress on Prehistory, Livingstone, 1955*. London: Chatto and Windus.

Cole, S. (1955). Third Pan-African Congress on Prehistory. *Antiquity, 29*, 203–208.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman and Hall.

Davidson, I. (2002). The `Finished Artefact Fallacy': Acheulean handaxes and language origins. In A. Wray (Ed.), *Transitions to Language* (pp. 180–203). Oxford: Oxford University Press.

Davidson, I., & Noble, W. (1993). Tools and Language in Human Evolution. In K. Gibson & T. Ingold (Eds.), *Tools, Language and Cognition in Human Evolution* (pp. 363–388). Cambridge: Cambridge University Press.

D'Errico, F., Marti, A. P., Shipton, C., Le Vraux, E., Ndiema, E., Goldstein, S., et al. (2020). Trajectories of cultural innovation from the Middle to Later Stone Age in Eastern Africa: Personal ornaments, bone artifacts, and ocher from Panga ya Saidi, Kenya. *Journal of Human Evolution, 141*. https://doi.org/10.1016/j.jhevol.2019.102737.

Diez-Martin, F., Dominguez-Rodrigo, M., Sanchez, P., Mabulla, A. Z. P., Tarrino, A., Barba, R., et al. (2009). The Middle to Later Stone Age technological transition in East Africa: New data from Mumba Rockshelter Bed V (Tanzania) and their implications for the origin of modern behavior. *Journal of African Archaeology, 7*(2), 147–173. https://doi.org/10.3213/1612-1651-10136.

Discacciati, A., Orsini, N., & Greenland, S. (2015). Approximate Bayesian logistic regression via penalized likelihood by data augmentation. *Stata Journal, 15*(3), 712–736. https://doi.org/10.1177/1536867x1501500306.

Dunnell, R. C. (1971). *Systematics in Prehistory*. New York: MacMillan.

Dunnell, R. C. (1978). Style and function: a fundamental dichotomy. *American Antiquity, 43*(2), 192–202. https://doi.org/10.2307/279244.

Eren, M. I., Diez-Martin, F., & Dominguez-Rodrigo, M. (2013). An empirical test of the relative frequency of bipolar reduction in Beds VI, V, and III at Mumba Rockshelter, Tanzania: Implications for the East African Middle to Late Stone Age transition. *Journal of Archaeological Science, 40*(1), 248–256. https://doi.org/10.1016/j.jas.2012.08.012.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19*(4), 676–685. https://doi.org/10.1214/088342304000000396.

Firth, D. (1992). Generalized linear models and Jeffreys priors: An iterative weighted least-squares approach. In Y. Dodge & J. Whittaker (Eds.), *Comutational Statistics Volume One* (pp. 553–557). Heidelberg: Physica Verlag.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics, 2*(4), 1360–1383. https://doi.org/10.1214/08-aoas191.

Gliganic, L. A., Jacobs, Z., Roberts, R. G., Dominguez-Rodrigo, M., & Mabulla, A. Z. P. (2012). New ages for Middle and Later Stone Age deposits at Mumba rockshelter, Tanzania: Optically stimulated luminescence dating of quartz and feldspar grains. *Journal of Human Evolution, 62*(4), 533–547. https://doi.org/10.1016/j.jhevol.2012.02.004.

Goodwin, A. J. H. (1926). South African stone implement industries. *South African Journal of Science, 23*, 784–788.

Goodwin, A. J. H. (1928). Sir Langham Dale's collection of stone implements. *South African Journal of Science, 25*, 419–426.

Goodwin, A. J. H. (1946). Earlier, middle, and later. *South African Archaeological Bulletin, 1*, 74–76.

Goodwin, A. J. H. (1958). Formative years of our prehistoric terminology. *South African Archaeological Bulletin, 13*, 25–33.

Goodwin, A. J. H., & Van Riet Lowe, C. (1929). *The Stone Age cultures of South Africa (Annals of the South African Museum Vol 27)*. Edinburgh: Neill and Company.

Gossa, T., Sahle, Y., & Negash, A. (2012). A reassessment of the Middle and Later Stone Age lithic assemblages from Aladi Springs, Southern Afar Rift, Ethiopia. *Azania-Archaeological Research in Africa, 47*(2), 210–222. https://doi.org/10.1080/0067270x.2012.676314.

Greenland, S., & Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine, 34*(23), 3133–3143. https://doi.org/10.1002/sim.6537.

Greenland, S., Mansournia, M. A., & Altman, D. G. (2016). Sparse data bias: A problem hiding in plain sight. *British Medical Journal, 353*. https://doi.org/10.1136/bmj.i1981.

Groucutt, H. S. (2020). Culture and convergence: The curious case of the Nubian complex. In H. S. Groucutt (Ed.), *Culture History and Convergent Evolution* (pp. 55–86). Cham: Springer Nature Switzerland AG.

Grove, M., & Blinkhorn, J. (2020). Neural networks differentiate between Middle and Later Stone Age lithic assemblages in eastern Africa. *Plos One, 15*(8). https://doi.org/10.1371/journal.pone.0237528.

Grove, M., & Pearson, J. (2014). Visualisation and permutation methods for archaeological data analysis. *Archaeological and Anthropological Sciences, 6*(4), 319–328. https://doi.org/10.1007/s12520-013-0158-z.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine, 21*(16), 2409–2419. https://doi.org/10.1002/sim.1047.

Ivanovaite, L., Serwatka, K., Hoggard, C. S., Sauer, F., & Riede, F. (2020). All these Fantastic cultures? research history and regionalization in the Late Palaeolithic tanged point cultures of Eastern Europe. *European Journal of Archaeology, 23*(2), 162–185. https://doi.org/10.1017/eaa.2019.59.

Kelly, A. J. (1996). *Intra-regional and inter-regional variability in the East Turkana (Kenya) and Kenyan Middle Stone Age*. Rutgers: State University of New Jersey, New Brunswick, New Jersey.

Kleindienst, M. R. (1967). Questions of terminology in regard to the study of Stone Age industries in eastern Africa: "Cultural Stratigraphic Units". In W. W. Bishop & J. D. Clark (Eds.), *Background to Evolution in Africa* (pp. 821–859). Chicago: University of Chicago Press.

Kleindienst, M. R. (2006). On naming things: Behavioral changes in the Later Middle to Earlier Late Pleistocene, viewed from the Eastern Sahara. In E. Hovers & S. L. Kuhn (Eds.), *Transitions Before the Transition* (pp. 13–28). Dordrecht: Springer.

Kohl-Larsen, L. (1943). *Auf den Spuren des Vormenshen*. Stuttgart: Strecker und Schroeder Verlag.

Kolassa, J. E. (1997). Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scandinavian Journal of Statistics, 24*(4), 523–530. https://doi.org/10.1111/1467-9469.00078.

Lahr, M. M., & Foley, R. A. (2016). Human evolution in Late Quaternary eastern Africa. In S. C. Jones & B. A. Stewart (Eds.), *Africa from MIS 2-6: Population Dynamics and Palaeoenvironments* (pp. 215–231). Dordrecht: Springer.

Langley, M. C., Prendergast, M. E., Shipton, C., Morales, E. M. Q., Crowther, A., & Boivin, N. (2016). Poison arrows and bone utensils in late Pleistoceneeastern Africa: evidence from Kuumbi Cave. *Zanzibar. Azania-Archaeological Research in Africa, 51*(2), 155–177. https://doi.org/10.1080/0067270x.2016.1173302.

Leakey, M. D., Hay, R. L., Thurber, D. L., Protsch, R., & Berger, R. (1972). Stratigraphy, archaeology, and age of Ndutu and Naisiusiu beds, Olduvai Gorge, Tanzania. *World Archaeology, 3*(3), 328–341. https://doi.org/10.1080/00438243.1972.9979514.

Leplongeon, A., Pleurdeau, D., & Hovers, E. (2017). Late Pleistocene and Holocene Lithic Variability at Goda Buticha (Southeastern Ethiopia): Implications for the understanding of the Middle and Late Stone Age of the horn of Africa. *Journal of African Archaeology, 15*(2), 202–233. https://doi.org/10.1163/21915784-12340010.

Lesaffre, E., & Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society Series B-Methodological, 51*(1), 109–116.

Mabulla, A. Z. P. (2007). Hunting and foraging in the Eyasi Basin, Northern Tanzania: Past, present and future prospects. *African Archaeological Review, 24*(1-2), 15–33. https://doi.org/10.1007/s10437-007-9013-x.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Mansournia, M. A., Geroldinger, A., Greenland, S., & Heinze, G. (2018). Separation in logistic regression: Causes, consequences, and control. *American Journal of Epidemiology, 187*(4), 864–870. https://doi.org/10.1093/aje/kwx299.

Masao, F. T. (2015). Characterizing archaeological assemblages from eastern Lake Natron, Tanzania: results of fieldwork conducted in the area. *African Archaeological Review, 32*(1), 137–162. https://doi.org/10.1007/s10437-014-9170-7.

Mehlman, M. J. (1979). Mumba-Hohle revisited: Relevance of a forgotten excavation to some current issues in East African prehistory. *World Archaeology, 11*(1), 80–94. https://doi.org/10.1080/00438243.1979.9979751.

Mehlman, M. J. (1989). Later Quaternary archaeological sequences in northern Tanzania. PhD Thesis, Department of Anthropology, University of Illinois at Urbana-Champaign.

Miller, J. M., & Willoughby, P. R. (2014). Radiometrically dated ostrich eggshell beads from the Middle and Later Stone Age of Magubike Rockshelter, southern Tanzania. *Journal of Human Evolution, 74,* 118–122. https://doi.org/10.1016/j.jhevol.2013.12.011.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

O'Brien, M. J., & Lyman, R. L. (2002). The epistemological nature of archaeological units. *Anthropological Theory, 2*, 37–56.

O'Brien, M. J., Boulanger, M. T., Buchanan, B., Collard, M., Lyman, R. L., & Darwent, J. (2014). Innovation and cultural transmission in the American Paleolithic: Phylogenetic analysis of eastern Paleoindian projectile-point classes. *Journal of Anthropological Archaeology, 34*, 100–119. https://doi.org/10.1016/j.jaa.2014.03.001.

Ossendorf, G., Groos, A. R., Bromm, T., Tekelemariam, M. G., Glaser, B., Lesur, J., et al. (2019). Middle Stone Age foragers resided in high elevations of the glaciated Bale Mountains, Ethiopia. *Science, 365*(6453), 583–587. https://doi.org/10.1126/science.aaw8942.

Pante, M., de la Torre, I., d'Errico, F., Njau, J., & Blumenschine, R. (2020). Bone tools from Beds II-IV, Olduvai Gorge, Tanzania, and implications for theorigins and evolution of bone technology. *Journal of Human Evolution, 148,* 102885. https://doi.org/10.1016/j.jhevol.2020.102885.

Pargeter, J., & Shea, J. J. (2019). Going big versus going small: Lithic miniaturization in hominin lithic technology. *Evolutionary Anthropology, 28*(2), 72–85. https://doi.org/10.1002/evan.21775.

Parkington, J. (1993). The neglected alternative: Historical narrative rather than cultural labelling. *South African Archaeological Bulletin, 48*, 94–97.

Pleurdeau, D., Hovers, E., Assefa, Z., Asrat, A., Pearson, O., Bahain, J. J., et al. (2014). Cultural change or continuity in the late MSA/Early LSA of southeastern Ethiopia? The site of Goda Buticha, Dire Dawa area. *Quaternary International, 343*, 117–135. https://doi.org/10.1016/j.quaint.2014.02.001.

Prendergast, M. E., Luque, L., Dominguez-Rodrigo, M., Diez-Martin, F., Mabulla, A. Z. P., & Barba, R. (2007). New excavations at Mumba Rockshelter, Tanzania. *Journal of African Archaeology, 5*(2), 217–243. https://doi.org/10.3213/1612-1651-10093.

Rahman, M. S., & Sultana, M. (2017). Performance of Firth-and log F-type penalized methods in risk prediction for small or sparse binary data. *Bmc Medical Research Methodology, 17*. https://doi.org/10.1186/s12874-017-0313-9.

Ranhorn, K., & Tryon, C. A. (2018). New radiocarbon dates from Nasera Rockshelter (Tanzania): Implications for studying spatial patterns in Late Pleistocene technology. *Journal of African Archaeology, 16*(2), 211–222. https://doi.org/10.1163/21915784-20180011.

Reynolds, N., & Riede, F. (2019a). House of cards: Cultural taxonomy and the study of the European Upper Palaeolithic. *Antiquity, 93*(371), 1350-1358, doi:10.15184/aqy.2019.49.

Reynolds, N., & Riede, F. (2019b). Reject or revive? The crisis of cultural taxonomy in the European Upper Palaeolithic and beyond. *Antiquity, 93*(371), 1368–1370. https://doi.org/10.15184/aqy.2019.156.

Riede, F., Araujo, A. G. M., Barton, M. C., Bergsvik, K. A., Groucutt, H. S., Hussain, S. T., et al. (2020). Cultural taxonomies in the Paleolithic-Old questions, novel perspectives. *Evolutionary Anthropology, 29*(2), 49–52. https://doi.org/10.1002/evan.21819.

Robertshaw, P. (1990). A history of African archaeology: An introduction. In P. Robertshaw (Ed.), *A History of African Archaeology* (pp. 3–12). London: James Currey.

Sampson, C. G. (1974). *The Stone Age Archaeology of Southern Africa*. New York: Academic Press.

Sauer, F., & Riede, F. (2019). A critical reassessment of cultural taxonomies in the Central European Late Palaeolithic. *Journal of Archaeological Method and Theory, 26*(1), 155–184. https://doi.org/10.1007/s10816-018-9368-0.

Scerri, E. M. L., Niang, K., Candy, I., Blinkhorn, J., Mills, W., Cerasoni, J. N., et al. (2021). Continuity of the Middle Stone Age into the Holocene. *Scientific Reports, 11*(1), 70. https://doi.org/10.1038/s41598-020-79418-4.

Shea, J. J. (2014). Sink the Mousterian? Named stone tool industries (NASTIES) as obstacles to investigating hominin evolutionary relationships in the Later Middle Paleolithic Levant. *Quaternary International, 350*, 169–179. https://doi.org/10.1016/j.quaint.2014.01.024.

Shea, J. J. (2019). European Upper Palaeolithic cultural taxa: Better off without them? *Antiquity, 93*(371), 1359–1361. https://doi.org/10.15184/aqy.2019.117.

Shea, J. J. (2020). *Prehistoric stone tools of Eastern Africa: A Guide*. Cambridge: Cambridge University Press.

Shipton, C., Roberts, P., Archer, W., Armitage, S. J., Bita, C., Blinkhorn, J., et al. (2018). 78,000-year-old record of Middle and Later stone age innovation in an East African tropical forest. *Nature Communications, 9*. https://doi.org/10.1038/s41467-018-04057-3.

Tryon, C. A. (2019). The Middle/Later Stone Age transition and cultural dynamics of late Pleistocene East Africa. *Evolutionary Anthropology, 28*(5), 267–282. https://doi.org/10.1002/evan.21802.

Tryon, C. A., & Faith, J. T. (2013). Variability in the Middle Stone Age of Eastern Africa. *Current Anthropology, 54*, S234–S254. https://doi.org/10.1086/673752.

Tryon, C. A., & Faith, J. T. (2016). A demographic perspective on the Middle to Later Stone Age transition from Nasera rockshelter, Tanzania. *Philosophical Transactions of the Royal Society B-Biological Sciences, 371*(1698). https://doi.org/10.1098/rstb.2015.0238.

Tryon, C. A., Lewis, J. E., Ranhorn, K. L., Kwekason, A., Alex, B., Laird, M. F., et al. (2018). Middle and Later Stone Age chronology of Kisese II rockshelter (UNESCO World Heritage Kondoa Rock-Art Sites), Tanzania. *Plos One, 13*(2). https://doi.org/10.1371/journal.pone.0192029.

Underhill, D. (2011). A history of Stone Age archaeological study in South Africa. *South African Archaeological Bulletin, 66*, 3–14.

Will, M., Tryon, C., Shaw, M., Scerri, E. M. L., Ranhorn, K., Pargeter, J., et al. (2019). Comparative analysis of Middle Stone Age artifacts in Africa (CoMSAfrica). *Evolutionary Anthropology, 28*(2), 57–59. https://doi.org/10.1002/evan.21772.