HOW RELIABLE ARE TRADITIONAL METHODS OF ASSESSING HANDAXES?

Bingham, P.[1] & McNabb, J.[1]

ABSTRACT

This project analysed variability in the recording of handaxe metrics, cortex/natural surface, rolling, patination and staining, by masters students and PhD candidates of Palaeolithic archaeology, compared to students of similar academic seniority from other branches of archaeology. Outlying recordings were found not to be related to specialist subject, academic seniority, experience or confidence, but rather that some individuals have a natural propensity to make outlying estimates. Overall, although the measurement of length, breadth and thickness with hand-held callipers was reasonably accurate/precise, the estimation of L1, cortex, rolling and patination/staining was highly variable. Recommendations are made for projects recruiting students to measure lithics, for candidates setting out on a PhD involving the recording of lithics and for organisations who might consider establishing reference collections to enable researchers to standardise the assessments they need to make on handaxes.

Full reference: Bingham, P. & McNabb, J. 2013. How reliable are traditional methods of assessing handaxes? *Lithics: the Journal of the Lithic Studies Society* 34: 5–13.

Keywords: Palaeolithic archaeology, biface industry, recording lithics, recruiting students, reference collection

INTRODUCTION

Comparatively little has been published on the reliability of traditional methods of assessing handaxes and other knapped stone artefacts (Gnaden & Holdaway 2000). Shaw & White (2003) remarked on ‘inevitable inter-analyst variation, especially in the tricky L1 measurement’, but without quantifying this variation. Grosman *et al* (2008) made a case for 3D digital scanning, based on their finding that an individual experienced archaeologist had an error rate greater than 5% for some values when asked to repeat metrics on 20 handaxes using hand-held callipers, attributing this to inconsistency in orienting the handaxes.

METHOD

This paper is based on a masters dissertation by one of us (Bingham 2012). The project recruited a total of 28 students so that it was possible to compare masters students and PhD candidates both of Palaeolithic archaeology and other branches of archaeology (see Table 1). The aim was to evaluate whether specialisation in the palaeolithic and/ or academic seniority increased reliability in assessing handaxes.

In an ‘assessment’ that was restricted to about an hour (not all completed the assessment), students were guided through a set protocol of analysis. At the outset, participants were instructed in the use of ‘zeroing’), and how to orient a handaxe following the method of Roe (1964). Some measurements were repeated, interspaced among other measurements (to assess test/retest reliability).

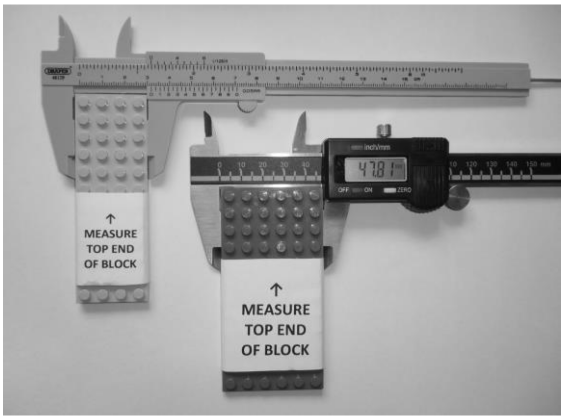

Figure 1. Manual and digital callipers used in the project, with the smaller and larger Lego blocks. Blocks labelled so same end measured each time

Participants were shown the difference between cortex/natural surface and flaked surface, and requested to estimate the unflaked proportion by eye. Rolling was assessed against written definitions, based on those suggested by Ashton (1998), but incorporating a quantitative element for arête rounding from Chambers (2003) (see Table 2). In addition, rolling was assessed against ‘reference handaxes’ that had in turn been ‘calibrated’ against the written definitions by three

[1] Department of Archaeology, University of Southampton, Southampton, SO17 1BJ, United Kingdom

handaxes only. Before starting the assessment, the students self-assessed their confidence and experience in assessing handaxes, and at the conclusion they were asked about their prefered methods of assessment.

Table 1. Summary of the sample of participants recruited

Branch of archaeology studied	Masters student	PhD candidate	Total
Palaeolithic archaeology	7	7	14
Other branch of archaeology	7	7	14
Total	14	14	28

Table 2. Ordinal scale of definitions of rolling

Grade	Definition
0	Undamaged edges with no evidence of abrasion on either face
1	Edges with only occasional evidence of damage and/or slight abrasion on either face
2	Edges with moderate damage and/or arêtes on faces rounded to more than 1 mm but less than 2 mm at any point
3	Edges with heavy damage and/or arêtes on faces rounded to more than 2 mm at some points

Table 3. Summary of measurements using manual callipers of length, breadth, thickness and L1 (in mm) for test handaxes B,G, & H made by all 28 participants (SD: Standard Deviation; Coefficient of Variation: (SD/mean) x 100)

Handaxe	B				G				H			
Metric	Length	Breadth	Thickness	L1	Length	Breadth	Thickness	L1	Length	Breadth	Thickness	L1
Median	112.0	81.0	30.0	29.0	114.0	69.0	33.0	46.3	105.0	57.0	30.0	34.6
Mean	111.9	80.1	30.0	27.5	113.7	68.9	32.9	47.7	104.8	56.8	29.8	37.0
SD	1.24	1.74	1.11	7.86	1.37	1.32	0.96	8.05	1.08	1.65	1.26	6.56
(SD/mean) x100	1.11	2.18	3.70	28.5	1.21	1.92	2.93	16.8	1.03	2.90	4.23	17.7
25th Percentile	111.0	79.2	29.6	23.0	112.3	68.0	32.0	42.0	104.0	56.0	29.0	32.3
75th Percentile	113.0	81.0	31.0	32.0	115.0	70.0	33.4	54.0	106.0	58.0	31.0	41.8

Table 4. Summary of measurements using digital callipers of length, breadth, thickness and L1(in mm) for test handaxes B, G & H made by all 28 participants

Handaxe	B				G				H			
Metric	Length	Breadth	Thickness	L1	Length	Breadth	Thickness	L1	Length	Breadth	Thickness	L1
Median	112.3	81.1	31.0	26.6	114.5	69.8	33.8	42.5	105.6	57.6	30.1	36.6
Mean	112.3	81.2	30.9	26.3	114.1	69.9	33.8	43.6	105.5	57.3	30.1	37.9
SD	0.66	0.65	0.59	6.0	1.05	0.75	1.21	6.73	0.71	1.38	0.77	6.46
(SD/mean) x100	0.59	0.80	1.93	22.9	0.92	1.08	3.58	15.4	0.67	2.40	2.55	17.1
25th percentile	111.9	80.8	30.6	20.4	113.7	69.6	33.4	39.2	105.1	57.1	29.6	34.3
75th percentile	112.7	81.6	31.2	30.8	114.9	70.0	34.0	45.9	106.0	58.0	30.7	41.1

RESULTS

Participant measurement of two 'standard' (Lego) blocks

At various points in the assessment, students were requested to measure two Lego blocks, (where the orientation of the distance to be measured was unambiguous), with manual and digital callipers (see Fig. 1).The same two instruments were used for all measurements. The mean (and median) estimates using manual and digital callipers for the small and larger Lego blocks respectively were similar, but students tended to be less precise in using manual than digital callipers (see Fig. 2). With manual callipers, students were more likely to under than over estimate, and those that under estimated one standard block tended to under estimate the other block by a similar proportion (data not presented).

One of the four sub-groups of students (Table 1) was found to produce more 'outlying' measurements of the standard blocks but this finding was not consistent across the other assessments (metrics, rolling, etc.), and overall this is not considered 'significant' (see below). Participants who self-scored themselves as least experienced tended to be the most accurate in using manual callipers, possibly because they took more care. Perversely, participants who favoured manual over digital callipers were less precise in using this instrument to measure the standard blocks than participants who favoured digital callipers.

Measurement of handaxe metrics

Participants were requested to measure the length, breadth, thickness and L1 (the distance from the base of a handaxe to its widest point in profile: Roe 1964) of three test handaxes with manual callipers (see Table 3). They were later asked to again measure these metrics with digital callipers, interspersed between other assessments, (see Table 4).

Inspection of Tables 3 and 4 reveals:

- Mean values obtained with digital callipers were all slightly higher than equivalent values obtained with manual callipers, (with the exception of L1 for handaxes B & G). This reflected the findings when the standard blocks were measured, with participants tending to under estimate with manual callipers.
- Standard deviations were lower for all measurements made with digital callipers (with the one exception of the thickness of handaxe G). The difference was not as great as that for the standard blocks and is likely to reflect the greater difficulty orientating and practically measuring handaxes compared to standard blocks.
- For all handaxes, the coefficient of variation, SD/mean x 100, increases in the order length<breadth<thickness<L1.
- The coefficient of variation for L1 is much larger than the coefficients for the other three handaxe dimensions, reflecting considerable variation in making this measurement.

Twenty-two out of 28 participants progressed sufficiently through the protocol to repeat the measurement of metrics with manual callipers. Figure 3 shows a scattergram of measurements made by the 22 participants of the length of test handaxe B. The Pearson correlation coefficient between the two sets of readings is 0.52, $p = 0.013$. The mean for the first set of readings was 111.95 and for the second set 111.99. Nine participants recorded the same reading.

Table 5 shows correlation coefficients for repeat measurements of length, breadth, thickness and L1 with manual callipers of test handaxes B, G, and H. Less than half are statistically significant ($p > 0.05$) and all are far below a coefficient of 0.9 for a repeatable test recommended by Hopkins (2000). However, even if the results had been repeatable, it would have been necessary to consider how many of the results were 'consistently wrong'. Unfortunately, the protocol included only one measurement of the test handaxes with digital callipers.

If 'outlying estimates', with digital callipers, of length, breadth and thickness of test handaxes B, G & H are defined as more than ± 0.5 mm of the group mean (see Table 4), of 252 estimates (28 participants x 3 handaxes x 3 dimensions), 74 (29%) estimates are 'outliers' with under estimates (43) exceeding over estimates (31). Figure 4 shows the number of outlying estimates per participant, and the total number per sub-group.

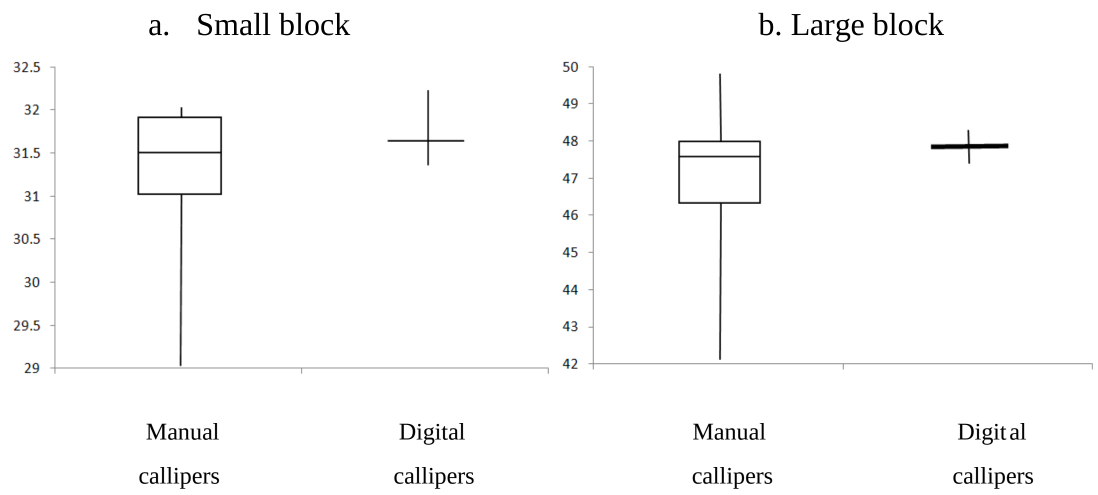

Figure 2 a & b: Box and whisker plots of the variability in students' manual and digital calliper measurements, of the smaller and larger blocks

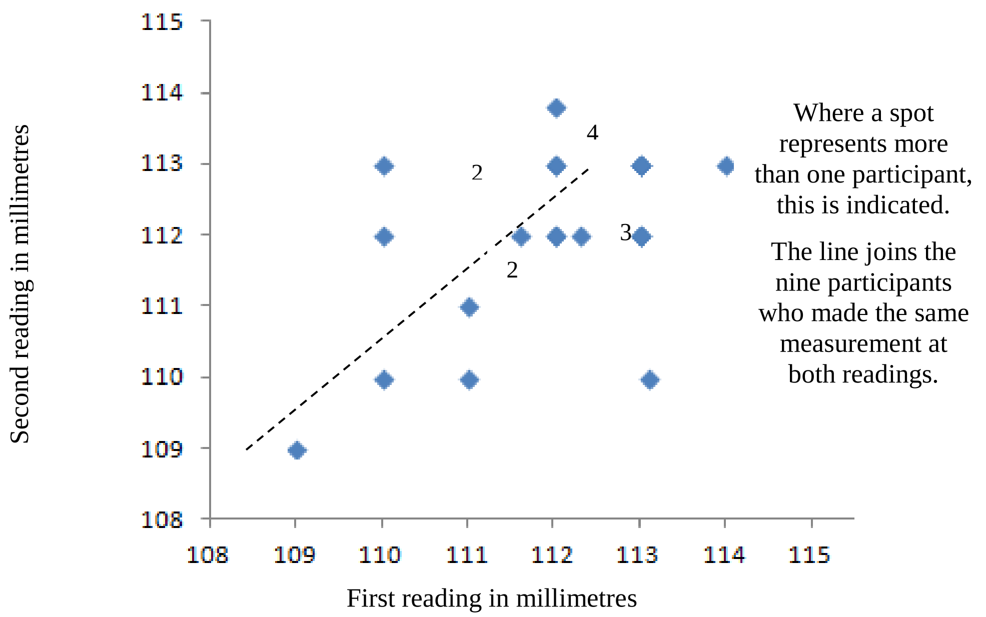

Figure 3. Correlation for test handaxe B of two readings by 22 participants, using manual callipers to measure

Table 5. Correlation of the two sets of readings of handaxe metrics, made using manual callipers by 22 out of the 28 participants

Handaxe	B				G				H			
Metric	Length	Breadth	Thickness	L1	Length	Breadth	Thickness	L1	Length	Breadth	Thickness	L1
Pearson correlation coefficient	0.52	0.79	0.70	0.02	0.38	0.07	0.19	0.65	0.37	0.50	0.41	0.28
Two tail significance	0.013	<0.001	<0.001	0.92	0.08	0.74	0.40	0.001	0.09	0.018	0.06	0.20

Figure 4 a & b: Distribution of outlying measurements of handaxe metrics (Digital callipers: Handaxes B, G & H)

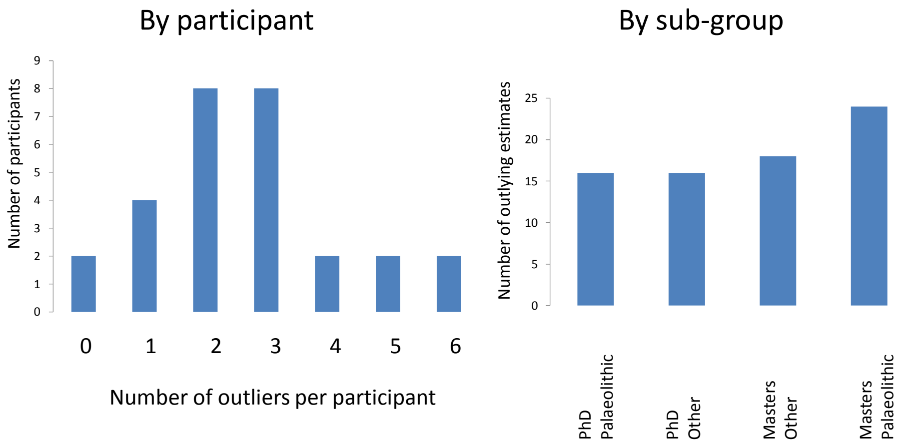

Table 6. Estimates made by all 28 participants of the percentage of the surface area of each of the nine test handaxes covered in cortex/natural surface

Handaxe	B	A	I	G	F	H	D	C	E
Median	30.0	22.5	20.0	15.0	5.0	5.0	5.0	1.0	0
Mean	29.9	22.6	18.0	14.8	6.8	6.3	5.7	*	**
SD	7.10	5.14	6.68	7.32	3.41	2.93	3.56		
Coefficient of variation (SD/mean) x 100	23.7	22.7	37.1	49.4	50.6	46.4	62.3		
25th percentile	25	20	11.3	8.5	5	5	5		
75th percentile	34.5	25	20	20	9.5	10	5		

** 13 participants assigned an estimate of 0%, and eight assigned 1%*
*** 21 participants assigned an estimate of 0%*

Table 7. Frequency of the grading of rolling of nine test handaxes, by 28 participants against written definitions

Handaxe	E	A	I	F	B	H	C	D	G
Grade 0	-	2	2	-	-	6	5	-	-
Grade 1	-	21	21	7	7	9	12	-	14
Grade 2	2	4	5	13	16	13	8	12	11
Grade 3	26	1	-	8	5	-	3	16	3
Interpretation	Consensus on a single grade			Intermediate position				Split opinion between two grades	

A four by two Chi-squared table of the four sub-groups against 'outlying/not outlying' estimate is not significant (Chi-square 3.29, 3 degrees of freedom, p = 0.35). However, six individuals made 30 of the 74 outlying measurements and this is considered to be of potential importance in recruiting students for projects.

Assessment of percentage cortex/natural surface on handaxes

For each of nine test handaxes, taking both faces into consideration, participants were requested to estimate by eye the proportion of cortex/natural surface that remained. Seven of the nine test handaxes retained cortex/natural surface and the coefficient of variation for the estimated proportion unflaked varied between 22% and 62% (see Table 6). Three experts also differed considerably in their estimation of coverage with cortex/natural surface. Correlation of the mean estimate of cortex/natural surface against the coefficient of variation shows that participants found it more difficult to estimate this characteristic when the percentage is small – reflecting the finding made in assessing handaxe metrics (see Tables 3 & 4).

Estimate of degree of rolling of handaxes

Participants were requested to grade the degree of rolling of nine test handaxes on a scale of 0 to 3 against written definitions (see Table 2), and later against ‘reference handaxes’ that had been graded by three experts against the written definitions. The two methods of grading were interspersed with other assessments, although it is possible that some participants recalled their earlier grading. It became apparent that participants found some test handaxes more difficult to grade than others, probably because these handaxes are located at the boundary between grades (for example, handaxes D & G in Table 7). It is suggested, however, that this reflects reality, (assemblages often include handaxes of varying degrees of rolling), and the measure of a good system for grading rolling (be it written definitions, reference handaxes or some alternative) may be its ability to discriminate between handaxes at the boundary between grades.

Participants differed in their grading of rolling of the test handaxes (inter-rater variation, see Table 7) and individuals differed in their grading between readings (test-retest variation). For example, between the reading against written definitions and against reference handaxes, participants made 101 grading units of change. Where for one handaxe, a change between two readings of one grade equals one grading unit, a change between two readings of two grades equals two grading units, etc. Test-retest variation was further demonstated by 24 of the participants who progressed far enough through the protocol to repeat the grading of three of the handaxes against the written definitions. This group made 29 units of grading change compared to their first reading against the written definitions.

Twenty-two participants who progressed far enough through the protocol were asked, without reference to the written definitions or reference handaxes, to sequence the nine test handaxes from the least to the most rolled. Table 8 lists the median and mean value for each handaxe; for each participant, least rolled handaxe = 1, next least rolled handaxe = 2, most rolled handaxe = 9. This shows overlapping median values indicating considerable variation between the 22 participants in the order in which they sequenced the test handaxes, although there is broad agreement with gradings against written definitions (Table 7), giving evidence of ‘parallel form reliability’.

Table 8. Median and mean values from the sequencing of nine test handaxes for rolling by 22 out of 28 participants

Handaxe	Median	Mean
A	1.5	1.9
I	3	3.18
H	3	3.23
C	4.5	3.9
G	4.5	4.6
B	6	5.6
F	6.5	5.9
D	8	7.8
E	9	9.0

After completing all the assessments, participants were asked for their view on whether ‘average condition’ or ‘most extreme feature of wear’ should determine the grade of rolling of a handaxe, and their preference for written definitions or reference handaxes. Twenty-one (75%) of the participants considered that the average condition of a handaxe should determine its grading, and this option was more strongly held by Palaeolithic students compared to students of other branches of archaeology. However, although the majority favoured ‘average condition’ rather than ‘the most extreme feature of wear’ in defining the grading of rolling, this view was not unanimous. Other aspects that might cause confusion in grading rolling include how much ‘weight’ to place on edge wear as opposed to rounding of arêtes.

Twenty-two (79%) participants favoured reference handaxes in preference to written definitions, with only three participants finding written definitions easier to use. Reference handaxes are, however, an ‘indirect’ standard, being based on written definitions and the interpretation of a group of experts, yet reference handaxes are clearly favoured. The practicalities of forming a collection of reference handaxes for rolling are explored below.

Assessment of patination/staining of handaxes

Participants were invited to grade the nine test handaxes for patination and staining against ‘reference handaxes’ that had been graded by the three experts. To assist the participants, the reference handaxes were arranged in a four by four grid (see Fig. 5). Selecting reference handaxes, over which the three experts had reasonable inter-expert and test-retest agreement, was difficult. Participants expressed considerable difficulty in grading the test handaxes, and there was considerable inter-rater variation in the grades assigned (data not presented).

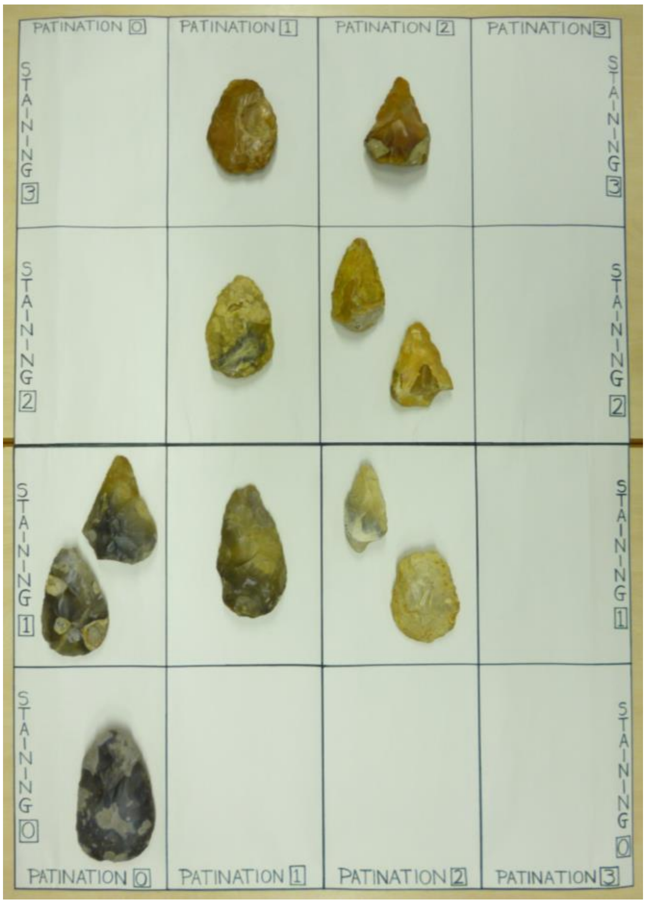

Figure 5. Four by four grid of patination and staining for reference handaxes

CONCLUSIONS AND RECOMMENDATIONS

Manual callipers are more likely to be misread than digital callipers and are thus less reliable, although digital callipers with metal jaws should be avoided because of the risk of damaging artefacts. Based on the data above it is suggested that the degree of precision shown in Table 8 should be achievable by students of Palaeolithic archaeology.

A significant result from this research is the identification of individuals who have a propensity to make outlying estimates of the various handaxe characteristics. Such individuals may not be identified by subject (Palaeolithic archaeology/other branch of archaeology), academic seniority (PhD candidate/masters student) or self-scoring of experience and confidence. If, for example, participants were to be recruited for a project where an important assemblage was to be catalogued, the project team should consider selecting individuals on the basis of their assessment of a selection of test handaxes. In addition, it is recommended that PhD candidates should be clear at an early stage of the accuracy they need to achieve for their project. A precision of ±5 mm for the measurement of L1, for example, may not be sufficient to answer some research questions (Pope *et al*. 2006). PhD candidates and others might be assisted if a collection of calibrated handaxes were established at a central location, such as the British Museum outstation at Franks House. A 'metrics' collection could include 12 handaxes that vary in basic dimensions and shape. It should be possible for a PhD candidate to complete an assessment of the test handaxes within an hour, and then enter their findings into a programme that prints off their results compared to other candidates/the experts. Anonymised data obtained from calibrating this assemblage, (experts and 3D scanning), might for example be used to further investigate the relative difficulty of measuring length/breadth/ thickness. In addition, a reference 'rolling' collection might be formed. Such an undertaking would not be easy and would depend on reconciling the views of a range of experts, for example over whether the average condition or the most extreme feature of wear should determine grading. This might be achieved by using a Delphi method where individual experts make an initial grading, and without revealing the authorship of individual gradings, results are fed back to the experts to see if this ‘moderates’ their views towards a consensus grading (Linstone & Turoff 1975). The practicability of a reference rolling collection in grading handaxes, especially those of intermediate grades of rolling, would then need to be assessed against a range of assemblages where rolling is considered important in interpretation. It may be that separate reference collections might be required for assemblages that are predominantly lightly rolled in contrast to those that are predominantly heavily rolled. Once formed, however, a reference collection for rolling should enable researchers to calibrate assemblages against a nationally accepted scale.

ACKNOWLEDGEMENTS

We are grateful to Gill Woolrich of the Southampton Museum Service, Robin Iles of the Winchester Museum Service and Delian Backhouse-Fry of the Isle of Wight Natural History and Archaeology Society for the loan of handaxes used in this project.

REFERENCES

Ashton, N.M. 1998. Flint analysis methodology. In N.M. Ashton, S.G. Lewis & S.A. Parfitt (eds) *Excavations at the Lower Palaeolithic Site at East Farm, Barnham, Suffolk, 1989–94*. British Museum Press, London.

Bingham, P. 2012. How reliable are traditional methods of assessing handaxes? Unpublished MA thesis, University of Southampton.

Chambers, J.C. 2003. Like a rolling stone? The identification of fluvial transportation damage signatures on secondary context bifaces. *Lithics* 24: 66–77.

Gnaden, D. & Holdaway, S. 2000. Understanding observer variation when recording stone artefacts. *American Antiquity* 65: 739–747.

Grosman, L., Smikt, O. & Smilansky, U. 2008. On the application of 3-D scanning technology for the documentation and typology of lithic artifacts. *Journal of Archaeological Science* 35: 3101–3110.

Hopkins, W.G. 2000. Measures of Reliability in Sports Medicine and Science. *Sports Medicine* 30: 1–15.

Linstone, H.A. & Turoff, M. 1975. *Delphi Method: Techniques and Applications*. Addistone-Wesley, Reading.

Pope, M., Russel, K. & Watson, K. 2006. Biface form and structured behaviour in the Acheulean. *Lithics* 27: 44–57.

Roe, D.A. 1964. The British Lower and Middle Palaeolithic: some problems, methods of study and preliminary results. *Proceedings of the Prehistoric Society* 30: 245–267.

Shaw, A.D. & White, M.J. 2003. Another look at the Cuxton handaxe assemblage. *Proceedings of the Prehistoric Society* 69: 305–313.