Journal of Archaeological Science 48 (2014) 5–14

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

On the importance of blind testing in archaeological science: the example from lithic functional studies

Adrian Anthony Evans*

Archaeological Sciences, School of Life Sciences, University of Bradford, Bradford BD7 1DP, UK

ARTICLE INFO

Article history:
Received 16 March 2013
Received in revised form
16 October 2013
Accepted 20 October 2013
Available online 1 December 2013

Keywords:
Blind-tests
Quantification
Method improvement
Lithic microwear
Functional analysis

ABSTRACT

Blind-testing is an important tool that should be used by all analytical fields as an approach for validating method. Several fields do this well outside of archaeological science. It is unfortunate that many applied methods do not have a strong underpinning built on, what should be considered necessary, blind-testing. Historically lithic microwear analysis has been subjected to such testing, the results of which stirred considerable debate. However, putting this aside, it is argued here that the tests have not been adequately exploited. Too much attention has been focused on basic results and the implications of those rather than using the tests as a powerful tool to improve the method. Here the tests are revisited and reviewed in a new light. This approach is used to highlight specific areas of methodological weakness that can be targeted by developmental research. It illustrates the value in having a large dataset of consistently designed blind-tests in method evaluation and suggests that fields such as lithic microwear analysis would greatly benefit from such testing. Opportunity is also taken to discuss recent developments in quantitative methods within lithic functional studies and how such techniques might integrate with current practices.

1. Introduction

Blind tests are standard methodology for testing archaeological scientific method and have, to provide just a few examples, been used in faunal analysis (Blumenschine et al., 1996; Gobalet, 2001), palynology (Pearsall et al., 2003), human osteology (Donnelly et al., 1998; Hill, 2000), and radiocarbon dating (Olsen et al., 2008). The importance of such tests is hard to overstate, especially when the technique in question revolves around human ability in subjective circumstances. An example of a review in one such area, taxonomic analysis, identified relatively few such tests have occurred and this was used to argue a move towards alternative, quantitative, methods (MacLeod et al., 2010). The focus here surrounds lithic microwear analysis as an example where subjective technique, and attempts to quantify such technique, meet blind testing.

Lithic functional studies can have wide ranging impact and are crucial to help us understand the activities, behaviour, and differences between archaic human and homonin species. There are many examples of the application of functional analysis techniques which have been performed by individuals who have been trained, or have trained themselves, in the use of these techniques (e.g. Juel Jensen, 1994; Keeley, 1980; van Gijn, 2009). The results of individual analyses are useful but pale in comparison to the ability to draw trends from multiple analyses of various assemblages from multiple sites that requires multiple analysts or laboratories. With standardization of method and technique calibration, one can enable comparability of results between laboratories and individual analysts. This can ultimately lead to robust theory building due to the increased size of useful datasets. To address important questions in palaeoanthropology and general archaeology, data from different regions and temporal periods is needed in a single comparative database; a task likely to be the result of work from multiple labs and individuals. Therefore, not only do analysts need to ensure that techniques provide useful data, they also need to ensure comparability between laboratories. Such a need has already been identified in other major fields of research, the best example being radiocarbon dating where inter-laboratory comparisons and discussion surrounding calibration are commonplace (e.g. Cuzange et al., 2007; Scott et al., 2010).

Standardization implies ubiquitous use of equivalent methods across a field of analysis while calibration involves understanding the distinct capabilities of individual methodological instruments. Calibration requires simply understanding the accuracy of individually applied techniques and the associated errors. One considers calibration a higher priority to lithic functional analysis than standardisation at present because without calibration one cannot

* Tel.: +44(0)1274 235729.
E-mail address: a.a.evans@bradford.ac.uk.

http://dx.doi.org/10.1016/j.jas.2013.10.026

determine which of the various methods within functional analysis (macroscopic analysis, low or high power microscopy, scanning electron microscopy and different scoring schemes etc.) one should used as a basis for standardisation. This paper reviews prior use of blind-testing and makes the argument that such tests are the means by which individual use-wear methods can be calibrated. Moreover, it is argued that blind-tests are fundamental to the identification of problematic areas within current techniques. This allows for targeted method improvement projects. It is suggested that quantification of some form will be of use in reconciling problematic areas, so some discussion focuses on these methods and possible ways in which 'traditional' and novel approaches can be integrated.

Before continuing, one needs to make a general statement to the reader. The statistics presented here should not be used directly to form a negative opinion of *applied* microwear analysis. As used, to evaluate the method for developmental purposes, test results are biased to a negative perspective. This is because evaluation is optimised to identify weaknesses in underlying technique. In applied situations, microwear specialists behave differently (or should) to how they approach sitting a blind test. In *applied* situations analysts can/should only assign functional interpretations where confidence is high. In *applied* situations analysts also use structured categorical determinations based on confidence level (e.g. if they cannot determine specific material but are confident about contact material hardness they will record results as such). There are two types of blind test that should not be confused: 1) Tests can be used to check appropriate behaviour by analysts (and to a degree capability) by asking them to behave as if in an *applied* situation, 2) test can also be used to evaluate technique. The difference is that it may be useful to have educated guesses (i.e. antler? or bone/antler) rather than 'undetermined' when looking for improvements in technique. Therefore it should be clear at the outset of a test which of these agendas it is to serve.

The presented analysis method is fundamental to evaluating technique and underlying issues; while the technique cannot escape the implications of these data completely (generally they do not show the field in a good light), the nuances described above ought to be considered before using this to attack practitioners. It should also be noted that the data presented in the following analysis is secondary to the central purpose of this paper and need not be taken as read. The main aim is to highlight how tests can be used *if* those form a solid dataset. As remarked elsewhere the variable design, the variable marking, the room for interpretation of results and the low sample sizes, all contribute to the fact that at present the blind-test database for microwear analysis isn't useful for exploitation in the manor described below.

2. Background

Contemporary lithic functional analysis comprises multiple methods. These methods include low power edge damage analysis (stereomicroscopy) (Tringham et al., 1974), the higher power approach (reflected microscopy) (Keeley, 1980), and the use of scanning electron microscopes. These applied techniques are all autoptic methods; individuals observe the edges of tools under magnification and, via visual study, form interpretations of tool use. Analysts sometimes combine these techniques to generate an understanding of worn surface features at a wider magnification range and this along with integration of residue analysis might be considered a best practice for use-wear studies.

Technique evaluation, standardisation, and calibration requires blind-testing. Tests have been conducted in lithic microwear analysis to a limited degree on the majority of individual techniques (Gendel and Pirnay, 1982; Knutsson and Hope, 1984; Newcomer et al., 1986; Newcomer and Keeley, 1979; Odell and Odell-Vereecken, 1980; Rots et al., 2006; Shea, 1987; Unrath et al., 1986; van den Dries, 1998; Vaughan, 1985, 1981), though it should be noted that testing has never been applied to the widely applied use of scanning electron microscopy. This statement also only applies to chipped stone technology; ground stone analysis for example appears devoid of blind-testing of method.

Blind-test results, evaluated below, average at 42.7% total accuracy across all tests (Table 1). These tests have not specifically guided developmental research, but rather have been the basis to

Table 1
Summary table of results of collated data from the published lithic microwear blind-tests.

Test	Year	Analysts/test	Unique Tools	Unique Edges	Total tests	% Accuracy Material	% Accuracy Direction	% Accuracy Total
Newcomer & Keeley	1979	1	15	16	16	43.8%	75.0%	37.5%
Odell & Odell-Vereken[2]	1980	1	31	31	31	35.5%	71.0%	32.3%
Vaughan[8]	1981	1	32	32	32			71.0%
Gendel & Pirnay	1982	1	23	23	23	65.2%	91.3%	65.2%
Knuttson & Hope	1984	1	4	4	4	75.0%	50.0%	50.0%
Newcomer et al T1[8]	1986	4	10	10	40	37.5%		
Newcomer et al T3[8]	1986	5	10	10	50	26.0 (6.0) [‡]%	46.0%	14.0%
Unrath et al	1986	4	20	28	112	42.9%	55.4%	36.6%
Bamforth et al	1990	1	20	29	29	58.6%	82.8%	58.6%
Shea T8[8, 1, 2]	1991	1	15	17	17	88.2 (64.7)[3]%	76.5%	70.6 (58.8)[3]%
Shea T2[8, 1, 2]	1991	1	18	26	26	69.2%	88.5%	61.5%
Shea T7[8, 1, 2]	1991	1	9	10	10	70.0%	80.0%	70.0%
Yamai	1992	1	9	9	9	55.6%	88.9%	55.6%
Shea & Klenck[8, 1, 2]	1993	1	60	71	71	49.3%	49.3%	38.0%
van Den Dries	1998	8	15	15	120	40.8%	76.7%	34.2%
Rots T2b[2]	2006	1	10	10	10	80.0%	90.0%	80.0%
Rots T2a[2]	2006	1	10	10	10	60.0%	100.0%	60.0%
Rots T1	2006	1	8	8	8	75.0%	87.5%	75.0%
Rots T3	2006	1	6	6	6	100.0%	83.3%	83.3%
Rots T2c	2006	1	10	10	10	90.0%	100.0%	90.0%
Stevens et al T1	2010	1	10	10	10	70.0%		
Stevens et al T1x	2010	1	10	10	10	60.0%		
Stevens et al T2	2010	1	10	10	10	60.0%		
Stevens et al T2x	2010	1	10	10	10	60.0%		
Total		40	343	383	642	49.5%	68.7%	42.7%

*Only summary data available, [1]only category based identifications, [2]low power, [‡]with/without partially correct answers, [3]variable results based on category interpretation.

justify the need for such work. With no specific direction, developmental research has taken a 'shotgun' approach, focussing on attempts to replace the method *in toto* with quantitative procedures. There are various forms of this; counting of scar types (Akoshima, 1987), simple edge damage measurement using GIS (Bird et al., 2007); reflected microscope images are analysed by eye, using expert systems – such as WAVES (van den Dries, 1998), and with the use of image analysis (Gonzalez-Urquijo and Ibanez-Estevez, 2003; Grace et al., 1985; Vila and Gallart, 1993). Other research has explored the potential of direct surface metrology applications (Anderson et al., 2006; Astruc et al., 2003; Dumont, 1982; Evans and Donahue, 2008; Evans and Macdonald, 2011; Kimball et al., 1995; Stemp and Stemp, 2003) and tribochemistry (Evans and Donahue, 2005; Šmit et al., 1999). While varied development of new approaches and advancements is expected in a research field, without an evaluation of current capability (widely practiced 'traditional' methods), one cannot identify which directions developmental research should progress, hence the variety in approaches examined. This paper reviews previous blind-tests data and outlines a strategy to advance the current situation to one in which developmental research moves in sync with methodological needs in a way that can improve the technique and the quality of the results achieved.

Blind-testing in lithic microwear analysis has become infrequent (see Fig. 1). The only published use of blind-testing in the last decade has been the testing of microwear in the context of identifying haft areas but additionally included tool use (Rots et al., 2006). Reasoning for the decline might be that the technique had reached a point where it is performing at a suitable standard. Olaussen (2005) reviewed a conference proceedings and implied that focus on applications rather than development was a sign of maturity across the field. However, there is yet to be a presentation that shows this to be the case; all evidence from blind-testing can be easily used to suggest a considerable need for improvement. This sadly could lead one to question if the many examples of application of microwear analysis have scientific merit. There are examples of novel variants of functional analysis, based on limited experimental data and no suitable blind-testing, been used to make statements of archaeological importance (e.g. Goodale et al., 2010; Wilkins et al., 2012). It is interesting that in the same broad field of lithic functional studies, the urinalysis technique known widely as the Hemastix test, used in residue analysis for identifying blood residue, scored similarly bad results through blind and semi blind-testing (Gurfinkel and Franklin, 1988; Manning, 1994) but has been all but entirely disbanded and is widely disregarded as without use (that technique has since been validated (though not blind-tested) through a comprehensive experimental review (Matheson and Veall, 2014)).

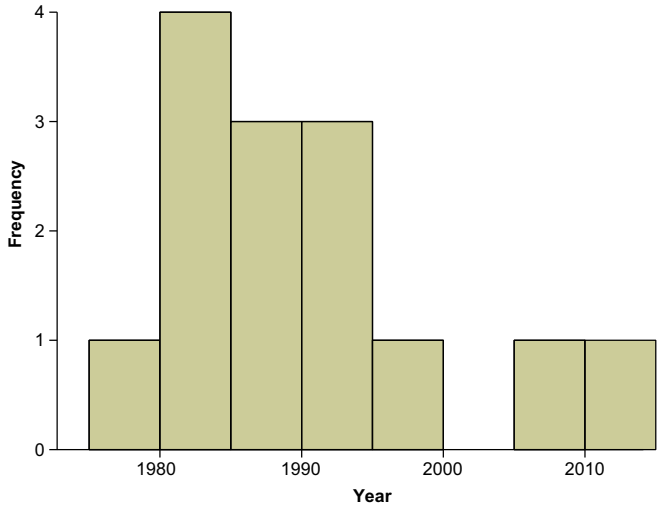

Fig. 1. Bidecadal frequency of blind-testing in lithic microwear analysis.

Table 2
Summary of the van den Dries WAVES test accuracies. Derived from van den Dries (1998: Appendix 1).

Analyst	Material	Motion	Total
I	53.3%	86.7%	53.3%
I (WAVES)	26.7%	80.0%	26.7%
II	53.3%	80.0%	46.7%
II (WAVES)	46.7%	66.7%	26.7%
IV (WAVES)	13.3%	80.0%	13.3%
III (WAVES)	26.7%	73.3%	13.3%
IV	33.3%	60.0%	26.7%
Experts (WAVES)	73.3%	86.7%	66.7%
Grand total	40.8%	76.7%	34.2%

3. Blind-tests to date

Nineteen blind-tests (from 12 sources) have previously been conducted (Bamforth et al., 1990; Gendel and Pirnay, 1982; Knutsson and Hope, 1984; Newcomer et al., 1986; Newcomer and Keeley, 1979; Odell and Odell-Vereecken, 1980; Rots et al., 2006; Shea, 1987; Shea and Klenck, 1993; Stevens et al., 2010; Unrath et al., 1986; van den Dries, 1998; Vaughan, 1981, 1985; Yamei, 1992). Tests prior to 1988 have been discussed by Bamforth (1988); tests by Yamei (1992), Shea and Klenck (1993) van den Dries (1998), Rots et al. (2006), and Stevens et al. (2010), are briefly described here. The test by Young and Bamforth (1990), though interesting, is excluded from discussion here since it refers only to identification of used edge by a test group of non-specialists.

3.1. *Bamforth et al., 1990*

This test involved the analysis of 20 tools (29 edges) used for durations between 5 min and 46 min. Analysis was conducted using the Keeley method (high power optical microscopy of edge fracturing, striations, and surface polishing (Keeley, 1980)). Seventeen out of the possible 29 edges were identified correctly for contact material (58.6%). The system used by Bamforth was designed with the intent to illustrate performance accuracy in applied situations; Bamforth only made inferences where confidence in determination was high, using 'unknown' in a number of situations. If these five pieces are excluded, then Bamforth scored 17/24 (70%).

3.2. *Yamei, 1992*

Yamei's test was run alongside the analysis of some of the material from the Peking Man site and Ma'anshan in south China (Yamei, 1992). The test was small with just nine tools, eight of which were used in various ways. Analysis was conducted using magnifications ranging 100–500× following Keeley. Worked material was identified to an accuracy of 70% while motion was accurate to 78%.

3.3. *Shea and Klenck, 1993*

These tests represent the only sizeable attempt to qualify the effect of trampling on the ability of low-power microwear analysis. The results of these tests, which involved variable amounts of tumbling (in lieu of actual trampling) on a set of 60 tools, are

presented below in Table 4. The data serve as a useful guide and a point of debate surrounding effects of post-depositional processes. Unfortunately, detailed analysis cannot be performed due to the summary presentation of data in the source.

3.4. Rots et al., 2006

The tests presented by Rots et al. (2006) consists of a series of three short tests that include determinations of tool use but were designed with identifying hafting traces in mind. All tools of test 1were used with a minimum duration of 30 min. Test one had two errors (75% accuracy); antler use, interpreted as 'shist' and tanned leather working, interpreted as 'wood' or 'hide and ochre'. Test two is presented with results provided for three inspection techniques (macroscopic, low magnification, and high magnification). This test had ten tools and accuracies were 60% for macroscopic, 80% for low power, and 90% for high power. Test three consisted of six tools and was analysed using a combined approach which resulted in 100% accuracy for material identifications and only one minor error in use-action for a tool used for grooving but interpreted as being used for 'grooving and perforating'. The accuracies presented in this test are high relative to other blind-tests results (see Table 2) and as such are promising if they relate to some form of advancement in technique following earlier testing results. However, as they were designed for testing abilities for identifying hafting traces, and not tool use, the range of contact materials worked was very limited (the last two tests had only three classes of worked material). This may explain the relative high scoring seen here.

3.5. Vaughan (1981, 1985)

A test by Vaughan was described in his thesis (1981) and consisted of 32 tools. The tools were used by Vaughan and then subjected to methods designed to simulate sieving, handling, and trampling. Unfortunately, the specific details of the test were not reported. Analysis took place four months after tool use by which time Vaughan claimed he had forgotten how he had used them (hence why they could be used in this pseudo blind test); referring to his old notes only after analysis. Nine errors were made putting overall accuracy at 71%. Seven of these errors were ascribed to the secondary traces produced on the pieces. In later publication, Vaughan (1985) cites 28 tools and five inaccuracies, putting accuracy at 85%. The origin of this discrepancy is unclear.

3.6. van den Dries, 1998

This test was performed following the design of a computerised expert system, WAVES, to test its ability in making determinations of stone tool use. With this system, a user inputs data into the software package when prompted. This follows guided microscope observations and selecting from a list of variables such as type of edge damage, distribution of polish. The software then calculates a list of probabilities to assist in the identification of wear types. The benefit of such systems might be to assist specialists with consistency or as a training aid. The blind-test consisted of a set of 15 tools, studied by a set of three experts (who acted collectively as one individual) and three further students of varying experience. All five of these subjects used the system to identify tool use and three of them provided additional interpretations based on experience. These test results are open to interpretation due to the way that WAVES provides probabilities. For example, tool 2, used to butcher a deer carcase, scored 110 for leather, 80 for meat or fish, and 70 for hide (these numbers are probability scores, the higher the value the more likely the associated material/action was responsible for the wear). This is marked as correct in the publication, but elsewhere van den Dries states that the correct identification by the system would be meat or fish (van den Dries, 1998: 107). Since this did not score top, it is marked as incorrect in the dataset used here. Table 2 summarises the results of these tests. Scoring here is based on the author's interpretation of the meaning of such results. The lowest score for material identification accuracy was 13.3% by analyst IV using WAVES. This same analyst felt confident enough to make personal interpretations throughout the test and scored at 33%. In fact, all three analysts that provided two sets of data had higher scores without WAVES (see Table 2).

3.7. Stevens et al., 2010

This test was designed to coincide with and test a computerised analytical system involving the combined use of confocal image analysis and user observed variables. Test tools were analysed by individuals whom gave their subjective results and then by inputting their observations alongside image data from laser scanning confocal microscopy. Each analyst studied ten tools but not in replicate. The system was accurate at identifying contact material 60% of the time. Data shows analyst 1 outperformed the quantitative system by a narrow margin (70% accurate) while analyst 2 matched the system. This is a second example (alongside van den Dries (1998)) where 'traditional' microwear analysis has outperformed the advanced computed system. Those data are based on a training set of less than 40 tools and a small blind-test making it hard to draw strong conclusions. However, this is a good example of the type of shotgun approach that researchers are forced to take when there is a lack of a solid underlying blind-test dataset to inform on where method improvement can be targeted. It also points to an area where caution may be needed when designing a next-generation system (see Discussion). To digress briefly one ought to comment on the confocal system utilised by Stevens et al. They describe the use of an Olympus FV1000 for surface measurement. This is not a system designed for use in texture measurement as used by others for similar purposes (Evans and Donahue, 2008; Evans and Macdonald, 2011; Stemp and Chung, 2011). To measure correctly such systems are based on high precision stages and are calibrated using engineering standards. The FV1000 system is not considered suitable for high precision 3d measurement by Olympus due to unreliable and uncalibrated z-stages (Marcus pers comm 2013 – calibration specialist for Olympus laser and fluorescence systems). As a result, while it is not clear what steps Stevens et al. took to ensure repeatability and calibration of the system they applied, one suggests that better results could be achieved with a more appropriate system.

4. Blind-test data mining

Blind-test results were collated in a database for statistical evaluation. Variables extracted included: tool type, duration of use, material worked, motion of use, analyst experience (where possible), microscopy method used, interpreted material, and interpreted motion. Data could not be collated from all tests. Newcomer et al. (1986) described several tests but only discussed the results in terms of number of correct identifications per tool. There was no way to determine which tools were identified correctly in which test, and what the misidentifications were. The misidentifications are crucial because these show what types of wear are being confused (the utility of knowing the where and what of errors/confusions is shown below). This was also the case for tests reported by Shea (1987), Shea and Klenck (1993) and Vaughan (1985). This practice has precluded independent evaluation and combination of data with other tests here. Reasons for not simply using reported data is inconsistency in recording practices

underlying the data or potential error in original data transcription. For example in the tests conducted by Shea and Klenck (1993), their table 3 shows six identification errors but their table 4 on the same data reports seven errors.

For analysis across all tests, raw data were standardised (changing descriptions of motion, material worked so they were described consistently – e.g. some describe scraping as a motion while others use 'perpendicular'; some specify specific tree species while others simply state wood). Tests were remarked using a consistent system; partially correct and guesses are scored as incorrect and checks were made against tabulated data and raw data. This manipulation of data increased sensitivity in a way that highlights weak areas.

Together the tests consisted of 642 instances of tool analysis, 40 analyst instances, and 383 individual tool edges used in unknown ways. Table 1 shows the results of these individual blind-tests derivation from various publications. Test scores, shown in Table 1, range for contact material accuracy between 26% and 100%; averaging in total at 48.7%. These data highlight that there is likely a need to improve the technique.

5. Collated test results

Table 3 summarises contact material occurrence in blind testing data. The inclusion of each contact material in the tests is disproportionate; wood was most commonly included, followed by bone and antler (if combined, bone/antler were the most common test material). Some materials included were unconventional e.g. plastic (whilst interesting, is only there to trick). The tooth, included in the Newcomer et al. test (1986) was the only one in any test and the interpretation of bone/antler was probably suitable. The grinding and crushing of nuts was included in only one test (Odell and Odell-Vereecken, 1980) and the interpretations of hard wood sawing and bone chopping are examples of incorrect interpretations. It could be conceived that on the study of archaeological assemblages, where wood sawing and bone working, both craft type activities, are identified that this could actually be where the processing of nuts and food preparation, was taking place. However, the low number of replicates and tests that included nuts does not allow a suitable interpretation to be made as to the degree to which this is a problem. This is a good example of why a larger blind-test dataset is needed for lithic microwear; the result described above may simply be an outlier. As an individual result it is easy to overlook, however, inability to distinguish such activity types may be a legitimate issue; it is important that this is fully understood.

Among the range of materials included in tests there are four groups with sufficient replicates to allow further interrogation of data and discussion. These groups are discussed in detail below. Figs. 2–4 shows misidentification rates for each of these groups and the figures should not be confused with total rates of identification accuracy noted in the text.

5.1. Wood

Seventy-six examples of woodworking have featured in blind-tests and 41 of these were identified correctly; an accuracy of 54%. A further eight interpretations were 'wood?' or 'wood or plant'. If these were included as correct answers (correct material is mentioned) then accuracy increases to 64.5%. However, to get the best out of using blind-tests as an evaluation process it is best to score educated guesses, i.e. 'wood?' as a fail. This is because it is useful to identify where there are certainties in the ability of the technique and where ambiguous circumstances may occur. The most prominent misidentification was those involving bone/antler at 13.2% of all identifications (Fig. 2). Interpretations where material was unknown rated at 4% whilst if those with unsure interpretations were included this raises to 19.7%. There is no relationship between the length of time wood working tools were used (and how developed the wear might have been) and ability for the analyst to make a correct assertion. The tool used for the least amount of time was used for 1 min to adze wood and the second shortest tool duration was 10 min of wood boring; both of these were interpreted correctly. The worst case of inaccuracy comes from a tool used for 30 min interpreted as 'bone/antler'. In the Newcomer et al. test (1986), woodworking featured on two of the tools. Accuracy was 20% for the tool used for 15 min to scrape and 0% for the tool used for 9 min to bore. Shea's tests (1991)

Table 3
Summary of contact materials worked by tools in the collated blind-tests.

Contact material	Unique edges	Total occurrences in testing	Correct identifications
Wood	79	116	57(49.1%)
Hide	52	94	45(47.9%)
Antler	47	70	43(61.4%)
Bone	40	82	46(56.1%)
Unused	28	44	19(43.2%)
Plant	25	37	12(32.4%)
Meat	19	33	15(45.5%)
Vegetable	10	17	10(58.8%)
Shell	8	22	6(27.3%)
Fish	6	20	4(20.0%)
Grasses	6	20	13(65.0%)
Earth	5	9	5(55.6%)
Meat/bone	5	5	3(60.0%)
Hemp rope	3	3	3(100.0%)
Ivory	2	8	5(62.5%)
Schist	2	2	2(100.0%)
Hide/meat	1	4	1(25.0%)
Plastic	1	1	0(0.0%)
Clay	1	8	2(25.0%)
Stone	1	5	0(0.0%)
Tooth	1	1	1(100.0%)
Limestone	1	1	1(100.0%)

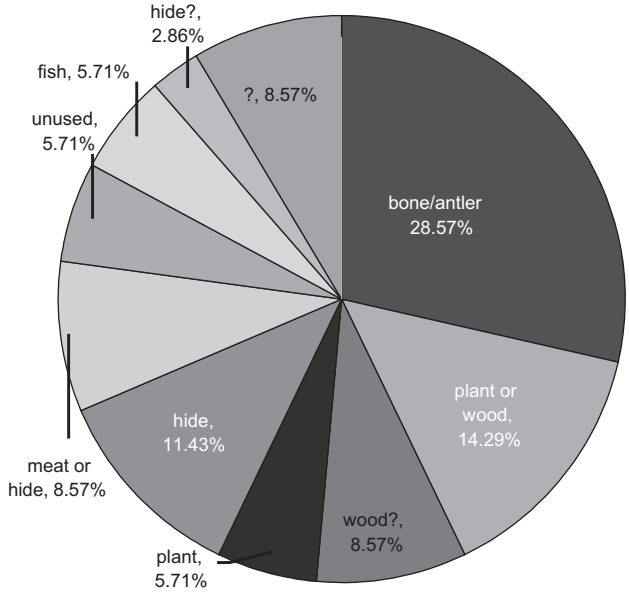

Fig. 2. Piechart showing simplified misidentification categories for test tools that were used to work wood.

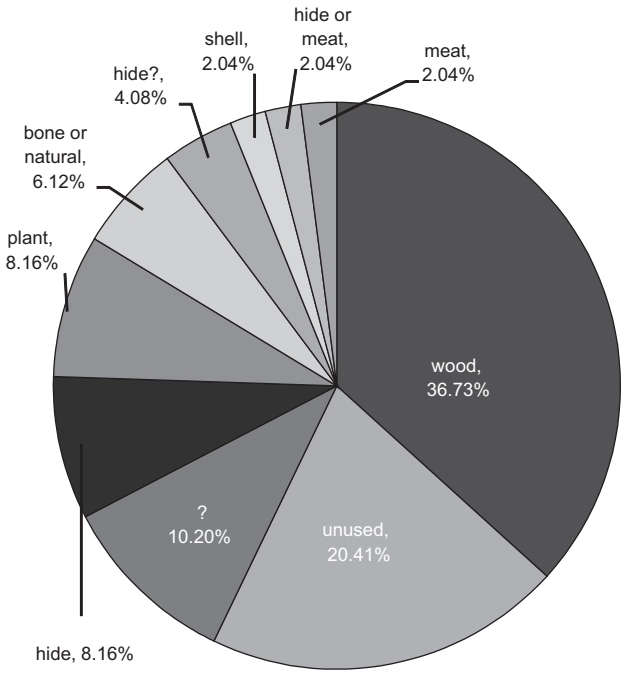

Fig. 3. Piechart showing simplified misidentification categories for test tools used to work bone and antler (combined).

categorised wood under three different types: medium soft, medium hard and hard vegetal, which allows an unusual degree of resolution and scored an overall of 71% accuracy across these groups.

5.2. Antler

Forty-three examples of antler working featured in the blind-tests. Five of these were interpreted correctly as antler but this figure should be combined with those for bone/antler interpretations, especially as the subjects' were quite clear that they were grouping these materials as one. General consensus appears to be that bone and antler can be grouped together as they produce similar wear seeing that they are essentially materially the same. Grouping these boosts accuracy to 44.2%. Thirty-five % of misidentifications are accounted for by interpretations including woodworking (Fig. 3). One misinterpretation was 'shell', one 'hide', and another 'stone'. There is no clear relationship between duration of tool use and accuracy of identification; the tool used for the shortest duration (13 min) was interpreted as 'unknown' and the tool used for the longest duration (60 min) as wood. In the Newcomer et al. (1986) tests one tool was used to work antler by cutting grooves for 12 min. Two of the five analysts had partially correct answers whilst the other three gave incorrect interpretations.

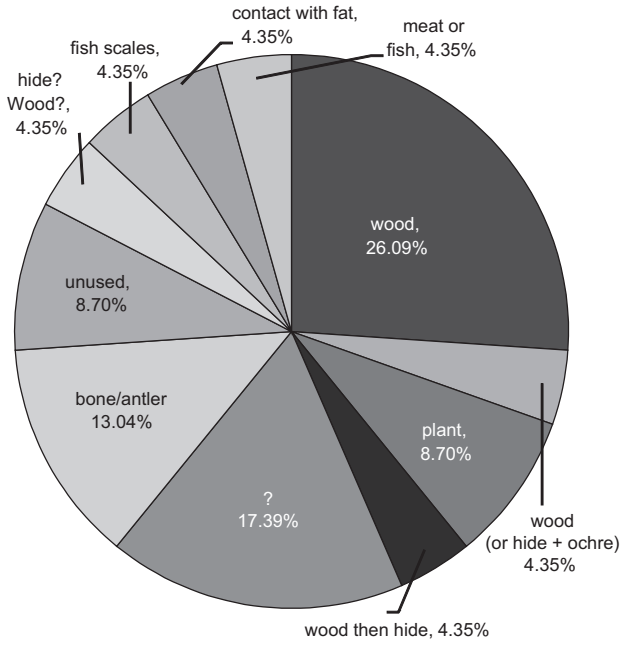

Fig. 4. Piechart showing simplified misidentification categories for test tools used to work varieties of hide.

5.3. Bone

Seventy-six bone-working identification attempts featured in the tests. Bone was the correct interpretation in 19 cases and, including 'bone/antler' as a correct result accuracy rated at 61.8%. Unused was the most frequent misidentification at 11%. This was caused primarily by the results of Tool 7 in the Unrath et al. test; a point that further emphasises the need for a much larger dataset. Wood featured in incorrect determinations 6.6% of the time. Four tools were incorrectly interpreted as being used to work hide. There is a relationship between use-duration and correct identification ($t = 2.8$, $p = 0.008$). Correctly identified tools were used on average for 34 min whereas misidentified tools averaged 22 min of use. In the Newcomer et al. (1986) tests one tool was used to scrape bone for 11 min. Three out of five analysts identified this tool as 'partially correct' (bone or antler). It is interesting to note that tools used to work bone were never misinterpreted as being used to work antler but antler working tools were misinterpreted as being bone working tools in 10% of the cases. Given the similarity of these materials some analysts have chosen to categorise as bone/antler and this also reflects the material class of hard animal used by Shea (1987). Shea successfully identified tools used on materials in this category, including partially correct answers 73% of the time.

5.4. Hide

In the combined dataset, 64 edges were used to work hide. There were variations in the types of hide and these included dry hide, fresh hide, tanned hide, wet hide, and leather. It would be useful to investigate how well these can be differentiated, but lack of clarity in interpretations, and breadth of resolution given the limited number of replicates, do not allow it. Instead, these are grouped together as 'hide'. Hide was the correctly identified worked material on 64% of tools from this group. Two tools were identified as being unused and these were used to work hide for 31 min and 46 min each. There is a relationship between use-duration and correct interpretation ($t = 2.5$, $p = 0.016$) as tools correctly interpreted were used on average for 73 min versus 46 min for misidentified tools. None of the tools in the Newcomer et al. (1986) tests were used on hide. Misinterpretations are diverse (Fig. 4), suggesting hide is hard to characterise or that the grouping of hide types into one group is not appropriate.

6. Discussion

There have been a number of discussions surrounding the tests and how they were performed. Notably, three use-wear specialists heavily critiqued the test Newcomer et al. (1986) conducted

(Bamforth, 1988; Hurcombe, 1988; Moss, 1987). Critique centred around three key areas, appropriate duration of tool use, analyst experience, and scoring of tests. In the framework of blind-test use to improve technique these are interesting areas of debate.

The value of including tools that had been used for limited durations has been criticised on the basis that such tools have not been used for a sufficient amount of time in order for diagnostic traces to have formed. This has been used to excuse why the test results were poor, and the claim is that they should not reflect on the technique as a whole. First, the duration of tool uses is not excessively short, 2 min in a single case may appear to be very low duration but when it is considered that the tool was used in a bow drill to drill 10 holes in thick shell it appears much more reasonable. Second, archaeological assemblages will contain tools that have been used for limited durations and these tools will invariably find themselves subject to microwear analysis as part of an assemblage assessment. Analysts cannot simply avoid a piece because it hasn't been used for long enough because in *application* the duration is unknown along with function. Therefore it is valid to test how robust the method is, regardless of tool use duration. From the data so far, for some contact materials, there is no clear relationship between duration of activity and correct identification. However as a full dataset, analysis of the relationship of duration of tool use and ability to make a correct identification shows that there is a trend and that correct identifications are more common on tools used for longer durations ($t = 3.8$, $p < 0.001$) (Fig. 5). Tools identified correctly were used for a mean duration of 32 min, compared with 22 min mean duration for incorrectly identified tools. The overall pattern found across the entire dataset agrees with Bamforth's (1988) analysis of a sub-set of these data. This matches expectations but duration is perhaps too simplistic as a variable. It is useful to know stroke rate and applied force and one expects that the lack of such data may be the reason why 'duration' as an encompassing term for 'amount of work' may fail to capture the significance of such variables on ability of different techniques to identify uses (Fig. 5).

Questions of analyst experience are complicated aspects of subjective techniques in general and so traditional lithic microwear methods are exposed to this complication – the analyst invariably requires training. Labs should ensure new analysts are trained to a level of competence through an internal process which should involve blind-testing. However, there needs to be differentiation between testing analysts to test competence and testing the method, as these are very different processes. Interestingly, until the technique has been blind-tested by experienced analysts on mass it is unknown to what level new students should be expected to achieve. There is obvious overlap here since at inception there is not a way to check if so called experienced analysts are actually experienced enough to take part in tests used to understand the capabilities of the method. This is one reason why a larger blind-test dataset is needed – the large dataset can account for the variability created by different analyst skill scores. With regard to user experience, it should be an aim when developing new techniques that they are robust to variable analyst experience.

A further criticism levelled against some of the tests is that analysts were guided to put an answer or a strict interpretation rather than leaving answers blank, or putting 'unknown', when they were not certain of tool use. This highlights a need to be distinct in the agenda of a particular test and best practice in approach to interpretation in archaeological analysis (what one should expect as a training exercise) and how one should approach answers in a blind-test which is used for method development. It can be discussed whether lack of an answer should be considered an incorrect result (since the analyst recognised that they did not have sufficient confidence in forming an interpretation). From the perspective of method development 'unknown' should be considered as an incorrect answer while educated guesses serve well to highlight where types of wear produced by different processes appear similar. Such a framework for answering tests is extremely useful in the highlighting of problem areas. However, in situations of archaeological analysis this is necessarily reversed, as providing an 'unknown' interpretation is greatly beneficial over any sort of guess. Tests are probably also needed to show that analysts are capable of identifying situations of uncertainty but this is distinct from primary test goals.

The suggestion that more testing is needed follows the method adopted in other fields of research. If we consider radiocarbon dating, almost annually multiple labs go through anonymous tests using the same technique, and the results are used to infer the accuracy of the method. Lithic microwear analysis specialists, and in fact all vulnerable techniques, should have a similar processes. It is the only way that fields can progress towards a series of methods that are broadly understood and accepted by a broader scientific community. To be more in line with scientific methods, when an analyst makes an interpretation of tool use based on lithic microwear analysis, there ought to be a mechanism to assign a probability to that interpretation. The ability to do this adds power to the technique and adds another dimension to the analysis of results that can make patterns of tool function easier to pick out.

A caveat of the presented review is that these tests are not strictly suitable for amalgamation; they include different techniques, different ranges of raw contact materials and other varying constraints such as different test designs including marking criteria and how analysts were briefed prior to the tests. However it does serve to illustrate the value of large blind-test datasets for technique review. The review identifies one problematic area, namely that it is common in the tests for tools used on wood to be misinterpreted as bone/antler processing tools and vice versa.

Finally, these tests were generally conducted using freshly produced material and the complications of post-depositional modifications are almost completely ignored at this point. While this is not inappropriate, since it makes sense to evaluate how the technique performs in ideal situations, it would be useful to understand potential issues. Some tests have tools that have undergone some kind of post-depositional modification or non-use related wear. Unrath et al. (1986) have examples of tools that were trampled and kept in a leather bag, and Vaughan's (1981) tools were all subject to a mix of damage. Tests by Shea and Klenck (1993) did include trampling prior to blind-test analysis to study the effect of post-depositional wear on the method's capability. A summary of their data is presented in Table 4. The trend identified there is that any amount of trampling drastically reduces interpretive ability but the amount of trampling is less important. These data resulted from low power analysis and category based answers. Consequently, how trampling effects ability to identify specific worked materials remains untested.

7. Building on a blind-test framework

With a robust blind-test dataset, developmental research could target weak areas. Data presented here show identification

Table 4
Summary data of the Shea & Klenck Blind test where tools were tumbled for different durations prior to blind analysis. Data is for contact material identification errors and accuracy.

Trampling (min)	Attempts	Errors accuracy	Accuracy
0	16	5	68.8%
15	18	10	44.4%
30	20	12	40.0%
45	17	9	47.0%

accuracy for woodworking tools is 54%, with the highest level of inaccuracy resulting from misinterpretation as antler or bone contact, accounting for almost 30% of misinterpretations. In terms of improving the accuracy, one envisages a method, which would allow differentiation between wood and bone/antler wear. That is at the point where an answer like 'wood?' or 'bone?' were not final but followed up by a technique to ratify the interpretation either way. If such an approach were available and had been available for these tests, interpretive accuracy may have risen at minimum from 51% to 75%. In the categories of bone and antler, accuracy in the tests rates at 61% and 44% respectively. If one were to count the results from the grouped tests in this way, accuracy at identifying bone/antler is 60% (62% with partial). Twenty-two per cent of overall misinterpretations involve wood contact, so an additional step of analysis that allowed differentiation between these two types (wood and bone/antler) would increase accuracy to 82%; a substantial gain. The power of identifying weak areas and using this to progress additional assistive methods is clear. For example, we can see that the primary misinterpretation of a tool used to work wood is that it was used to work bone/antler. This observation was anticipated; Keeley (1980) mentioned the problems of differentiating the worn surfaces resulting from use against these materials. The identification of an expected pattern shows that despite methodological differences the combine test dataset is of some use. This kind of analysis provides a quantified understanding of the problem: the scale of the misidentification is known and an estimate can be made on the benefit of reconciling any specific issue to the method as a whole. It also enables the assignment of statistical confidence with which an interpretation is made. This is only achieved because of the underlying blind-test system on which the method is founded.

From this point forward, upgrading methodological practices can progress in several ways. The first could involve a revaluation of the criteria used to differentiate these specific, hard to distinguish, material classes. Targeted research in this area might highlight a better set of textural or edge damage markers that can be used to make differentiations between contact materials. The solution could come from the use of approaches with improved microscopic resolution. While there are currently no data from the use of scanning electron microscopy, limited application of laser scanning confocal microscopy does reveal that there may be nanoscale textural features that can be used to differentiate these materials (Evans and Donahue, 2008). That research also indicated that quantitative approaches to analysis by the application of surface metrology can be used to aid interpretation. A situation may arise where the primary step in analysis is conducted using the traditional methods and then, where wood or bone/antler are suspected tool uses the interpretation can be substantiated through the use of this type of secondary technique. A further technique that may be suitable in this situation is the application of trace-element analysis. Using current tribological theory and experimental data, a case can be made for adhesive processes in wear models (Kato, 2002). While bone/antler and wood have a similar hardness – which is likely why the traces of wear are similar – they have very different structure and chemistry. Bone/antler is comprised primarily of apatite and collagen while wood is comprised primarily of cellulose. The mineral constituent of bone has calcium, phosphorus, and magnesium in levels that are far higher than those seen in wood. Two independent research groups have presented results that imply an ability to differentiate between wood and bone/antler by inter-surface chemical analysis (Evans and Donahue, 2005; Śmit et al., 1999).

Another area of interest highlighted by reviewing the blind-tests is the case for accuracy in the identification of wear produced from working hide in different preparatory states. There are no

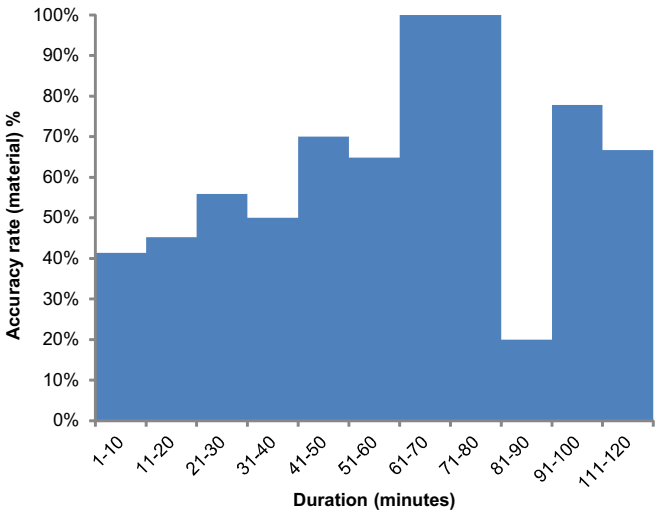

Fig. 5. % Histogram showing material identification accuracy by duration of tool use from the combined test dataset.

indications from the combined tests that the method has the ability to make distinctions between tools used to work hide in fresh, greasy, or dry states. Surface metrology (statistical analysis of surface texture) may be useful in this case. Fig. 6 shows some surface roughness data gathered from a set of experimental tools used on different contact materials (Evans and Donahue, 2008). This data was collected by measuring surface texture using confocal microscopy. The figure also shows some data collected from a scraper presented as a blind sample. The data overlaps substantially with dry hide scrapers and greasy hide scrapers. Statistical analysis (post ANOVA Tukey) groups the unknown with 'greasy hide' more frequently than any other and within these groups it is always the greasy hide that is matched closest (Evans and Macdonald, 2011). This suggests an interpretation that the unknown tool was used on

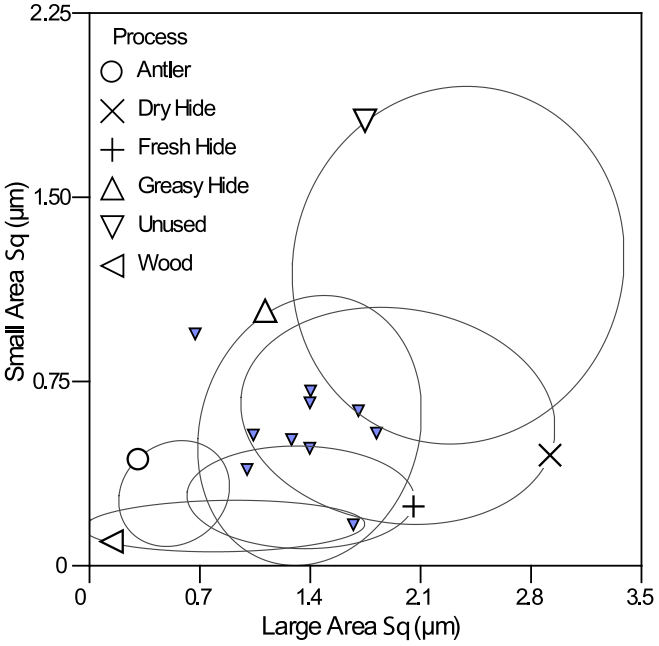

Fig. 6. Data from laser scanning confocal microscopy showing experimental data overlain by data from a tool used to work greasy hide – small filled triangles. Figure modified from Evans and Macdonald (2011).

greasy hide; which was revealed as a correct interpretation in this case. While this is by no means considered an extensive blind-test of the technique, it non-the-less serves as an example of potential capability. Such an approach of texture measurement could equally be applied to assist with the bone/antler-wood issue.

Using advanced quantitative methods as secondary approaches that 'add-on' to the traditional method serves to improve technique in a way that should not drive a wedge between practitioners of different approaches to analysis. The suggested gains above are hypothetical (because the existing blind-test dataset is not robust) and the next stage along is to reassess the method including the developed 'add-on' approaches using blind-tests to arrive at a new determination of method accuracy. Failings of attempts to build expert systems or fully quantitative systems (van den Dries, 1998; Stevens et al., 2010) spring from attempts to rebuild method from the ground up. While both efforts are good examples of system design and are considered a move in an appropriate direction, if they had a blind-test dataset, such as the one described here, on which to build new methods, they would have better understood what probabilities to assign to identification attempts of polish, edge damage, and interpreted material, when considering input variables.

Outside of strengthening material identification, this review has also highlighted duration of tool use as a problematic area. Research should focus on two areas. The first might be testing if current analytical frameworks are capable of identifying underdeveloped traces and not overreaching interpretation. The second is to develop add on techniques that can increase sensitivity at low durations of tool use.

This paper has focused on the lithic microwear analysis as a specific example of how blind-testing frameworks can be used to improve rather than simply critique technique; the principles apply to all areas of approach in archaeological science. Residue analysis, in particular structural residue studies, is becoming popular in lithic studies. Unfortunately, there are very few blind tests currently available for the technique (Hardy and Garufi, 1998; Wadley et al., 2004) the construction and analysis of which has been called into question (Crowther and Haslam, 2007). The argument put forward for lithic microwear analysis is equally valid here; much more effort needs to be placed on understanding the techniques in detail and then using this to find ways to improve scores (Monnier et al., 2012).

8. Conclusion

Methods in lithic microwear analysis (and functional analysis in general) have a limited blind-test dataset from which to form an understanding of accuracy. These limitations stem from highly diverse practices and low number of replicates. It is hard to make developmental decisions surrounding methodological research based on such a limited dataset. However, there is potential capability to be gained by the use of large blind-test datasets. Examples of secondary 'add-on' techniques have been presented as solutions to possible issues. This provides a useful example to other areas of development in archaeological science where should be considered as a means to evaluate advances.

There are clear benefits from understanding the current accuracy of our methods, and benefits in understanding which areas of the method can be improved upon. The current level of accuracy can only be derived from the tests at hand and these show a level of accuracy that should be regarded as unacceptable. If specialists aspire for microwear analysis to become an accepted archaeological science, real steps in understanding and methodological development need to be made. It is hoped that this review and discussion of this historical data reinvigorates discussion on the quality of the technique and the means for its improvement.

Future research effort should be placed on the production of a new large set of blind-test data. Such test can serve the field in a way as presented above and as a tool to help the individual analysts involved judge their capabilities against the results across various laboratories.

Acknowledgements

The analysts who have taken time to conduct and publish blind-test data are greatly thanked for their efforts to qualify the nature of our methods. I would like to thank Angel Jan for help include data from Yamei's test. Danielle Macdonald, Randy Donahue, Veerle Rots, James Stemp, and Harry Lerner for useful discussions, comments on a draft of this paper and continual collaboration. I'd also like to thank a number of specialists for useful discussions following conference presentation of versions of this paper. This research was funded by NERC (NER/S/A/2004/12213) and the AHRC (AH/J007935/1).

References

Akoshima, K., 1987. Microflaking quantification. In: Sieveking, G.D.G., Newcomer, M.H. (Eds.), The Human Uses of Flint and Chert: Papers Presented at the Fourth International Flint Symposium. Cambridge University Press, Cambridge, pp. 71–80.
Anderson, P.C., Georges, J.M., Vargiolu, R., Zahouani, H., 2006. Insights from a tribological analysis of the tribulum. J. Archaeol. Sci. 33, 1559–1568.
Astruc, L., Vargiolu, R., Zahouani, H., 2003. Wear assessments of prehistoric instruments. Wear 255, 341–347.
Bamforth, D.B., 1988. Investigating microwear polishes with blind tests: the institute results in context. J. Archaeol. Sci. 15, 11–23.
Bamforth, D.B., Burns, G.R., Woodman, C., 1990. Ambiguous use traces and blind test results: new data. J. Archaeol. Sci. 17, 413–430.
Bird, C., Minichillo, T., Marean, C.W., 2007. Edge damage distribution at the assemblage level on Middle Stone Age lithics: an image-based GIS approach. J. Archaeol. Sci. 34, 771–780.
Blumenschine, R.J., Marean, C.W., Capaldo, S.D., 1996. Blind tests of inter-analyst correspondence and accuracy in the identification of cut marks, percussion marks, and carnivore tooth marks on bone surfaces. J. Archaeol. Sci. 23, 493–507.
Crowther, A., Haslam, M., 2007. Blind tests in microscopic residue analysis: comments on Wadley et al. (2004). J. Archaeol. Sci. 34, 997–1000.
Cuzange, M.-T., Delque-Kolic, E., Goslar, T., Grootes, P.M., Higham, T., Kaltnecker, E., Nadeau, M.-J., Oberlin, C., Paterne, M., van der Plicht, J., Ramsey, C.B., Valladas, H., Clottes, J., Geneste, J.-M., 2007. Radiocarbon intercomparison program for Chauvet Cave. Radiocarbon 49, 339–347.
Donnelly, S.M., Hens, S.M., Rogers, N.L., Schneider, K.L., 1998. Technical note: a blind test of mandibular ramus flexure as a morphologic indicator of sexual dimorphism in the human skeleton. Am. J. Phys. Anthropol. 107, 363–366.
Dumont, J., 1982. The quantification of microwear traces: a new use for interferometry. World Archaeol. 14, 206–217.
Evans, A.A., Donahue, R.E., 2005. The elemental chemistry of microwear traces: an experiment. J. Archaeol. Sci. 32, 1733–1740.
Evans, A.A., Donahue, R.E., 2008. Laser scanning confocal microscopy: a potential technique for the study of lithic microwear. J. Archaeol. Sci. 35, 2223–2230.
Evans, A.A., Macdonald, D., 2011. Using metrology in early prehistoric stone tool research: further work and a brief instrument comparison. Scanning 33, 294–303.
Gendel, P.A., Pirnay, L., 1982. Microwear analysis of experimental stone tools: further test results. Stud. Praehist. Belg. 2, 251–265.
Gobalet, K.W., 2001. A critique of faunal analysis: inconsistency among experts in blind tests. J. Archaeol. Sci. 28, 377–386.
Gonzalez-Urquijo, J.E., Ibanez-Estevez, J.J., 2003. The quantification of use-wear polish using image analysis. First results. J. Archaeol. Sci. 30, 481–489.
Goodale, N., Otis, H., Andrefsky Jr., W., Kuijt, I., Finlayson, B., Bart, K., 2010. Sickle blade life-history and the transition to agriculture: an early Neolithic case study from Southwest Asia. J. Archaeol. Sci. 37, 1192–1201.
Grace, R., Graham, I.D.G., Newcomer, M.H., 1985. The quantification of microwear polishes. World Archaeol. 17, 112–120.
Gurfinkel, D.M., Franklin, U.M., 1988. A study of the feasibility of detecting blood residue on artifacts. J. Archaeol. Sci. 15, 83–97.
Hardy, B.L., Garufi, G.T., 1998. Identification of woodworking on stone tools through residue and use-wear analyses: experimental results. J. Archaeol. Sci. 25, 177–184.
Hill, C.A., 2000. Technical note: evaluating mandibular ramus flexure as a morphological indicator of sex. Am. J. Phys. Anthropol. 111, 573–577.
Hurcombe, L., 1988. Some criticisms and suggestions in response to Newcomer et al (1986). J. Archaeol. Sci. 15, 1–10.

Juel Jensen, H., 1994. Flint Tools Plant Working: Hidden Traces of Stone Age Technology. Aarhus University Press, Denmark.
Kato, K., 2002. Classification of wear mechanisms/models. Proc. Inst. Mech. Eng. J – J. Eng. Tribol. 216, 349–355.
Keeley, L.H., 1980. Experimental Determination of Stone Tool Uses. A Microwear Analysis. Chicago University Press, Chicago.
Kimball, L.R., Kimball, J.F., Allen, P.E., 1995. Microwear polishes as viewed through the atomic force microscope. Lithic Technol. 20, 6–28.
Knutsson, K., Hope, R., 1984. The Application of acetate peels in lithic usewear analysis. Archaeometry 26, 49–61.
MacLeod, N., Benfield, M., Culverhouse, P., 2010. Time to automate identification. Nature 467, 154–155.
Manning, A.P., 1994. A cautionary note on the use of hemastix and dot-blot assays for the detection and confirmation of archaeological blood residues. J. Archaeol. Sci. 21, 159–162.
Matheson, C.D., Veall, M., 2014. Presumptive blood test using Hemastix® with \EDTA\ in archaeology. J. Archaeol. Sci. 41, 230–241.
Monnier, G.F., Ladwig, J.L., Porter, S.T., 2012. Swept under the rug: the problem of unacknowledged ambiguity in lithic residue identification. J. Archaeol. Sci. 39, 3284–3300.
Moss, E.H., 1987. A review of investigating microwear polishes with blind tests. J. Archaeol. Sci. 14, 473–481.
Newcomer, M.H., Keeley, L.H., 1979. Testing a method of microwear analysis with experimental flint tools. In: Hayden, B. (Ed.), Lithic Use-wear Analysis. Academic Press, New York.
Newcomer, M., Grace, R., Unger-Hamilton, R., 1986. Investigating microwear polishes with blind tests. J. Archaeol. Sci. 13, 203–217.
Odell, G.H., Odell-Vereecken, F., 1980. Verifying the reliability of lithic use-wear assessments by "Blind Tests": the low power approach. J. Field Archaeol. 7, 87–120.
Olaussen, D., 2005. 'Traceology' then and now. Eur. J. Archaeol. 8, 295–297.
Olsen, J., Heinerneier, J., Bennike, P., Krause, C., Hornstrup, K.M., Thrane, H., 2008. Characterisation and blind testing of radiocarbon dating of cremated bone. J. Archaeol. Sci. 35, 791–800.
Pearsall, D.M., Chandler-Ezell, K., Chandler-Ezell, A., 2003. Identifying maize in neotropical sediments and soils using cob phytoliths. J. Archaeol. Sci. 30, 611–627.
Rots, V., Pirnay, L., Pirson, P., Baudoux, O., 2006. Blind tests shed light on possibilities and limitations for identifying stone tool prehension and hafting. J. Archaeol. Sci. 33, 935–952.
Scott, E.M., Cook, G.T., Naysmith, P., 2010. A report on Phase 2 of the Fifth International Radiocarbon Intercomparison (VIRI). Radiocarbon 52, 846–858.
Shea, J.J., 1987. On accuracy and relevance in lithic use-wear analysis. Lithic Technol. 16, 44–50.
Shea, J., 1991. The Behavioral Significance of Levantine Mousterian Industrial Variability (Doctoral dissertation). Harvard University. University Microfilms, Ann Arbor.
Shea, J.J., Klenck, J.D., 1993. An experimental investigation of the effects of trampling on the results of lithic microwear analysis. J. Archaeol. Sci. 20.
Šmit, Ž., Grime, G.W., Petru, S., Rajta, I., 1999. Microdistribution and composition of usewear polish on prehistoric stone tools. Nucl. Instr. Meth. Phys. Res. B – Beam Interact. Mater. Atoms 150, 565–570.
Stemp, W.J., Chung, S., 2011. Discrimination of surface wear on obsidian tools using LSCM and RelA: pilot study results (area-scale analysis of obsidian tool surfaces). Scanning 33, 279–293.
Stemp, W.J., Stemp, M., 2003. Documenting stages of polish development on experimental stone tools: surface characterization by fractal geometry using UBM laser profilometry. J. Archaeol. Sci. 30, 287–296.
Stevens, N.E., Harro, D.R., Hicklin, A., 2010. Practical quantitative lithic use-wear analysis using multiple classifiers. J. Archaeol. Sci. 37, 2671–2678.
Tringham, R., Cooper, G., Odell, G., Voytek, B., Whitman, A., 1974. Experimentation in the formation of edge damage: a new approach to lithic analysis. J. Field Archaeol. 1, 171–196.
Unrath, G., Owen, L., van Gijn, A., Moss, E.H., Plisson, H., Vaughan, P., 1986. An evaluation of microwear studies: a multi-analyst approach. In: Owen, L., Unrath, G. (Eds.), Technical Aspects of Microwear Studies on Stone Tools. Early Man News 9/10/11, pp. 51–68.
van den Dries, M.H., 1998. Archaeology and the Application of Artificial Intelligence. Leiden University, Leiden.
van Gijn, A., 2009. Flint in Focus: The Meaning of Flint in the Neolithic and Bronze Age. Sidestone Press.
Vaughan, P.C., 1981. Lithic Microwear Experimentation and the Functional Analysis of a Lower Magdalenian Stone Tool Assemblage (PhD thesis). University of Pennsylvania.
Vaughan, P., 1985. Use-wear Analysis of Flaked Stone Tools. The University of Arizona Press, Tucson.
Vila, A., Gallart, F., 1993. Caracterización de los micropulidos de uso: ejemplo de aplicación del análisis de imágenes digitalizadas. In: Anderson, P.C., Beyries, S., Otte, M., Plisson, H. (Eds.), Traces et fonction: les gestes retrovés, vol. 50. ERAUL, Belgium, pp. 459–465.
Wadley, L., Lombard, M., Williamson, B., 2004. The first residue analysis blind tests: results and lessons learnt. J. Archaeol. Sci. 31, 1491–1501.
Wilkins, J., Schoville, B.J., Brown, K.S., Chazan, M., 2012. Evidence for early hafted hunting technology. Science 338, 942–946.
Yamei, H., 1992. Experimental studies of microwear analysis on stone artifacts. Acta Anthropol. Sin. 11, 202–215.
Young, D., Bamforth, D.B., 1990. On the macroscopic identification of used flakes. Am. Antiq. 55, 403–409.