Journal of Archaeological Science xxx (2013) 1–10

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

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.

0305-4403/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
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.

Please cite this article in press as: Evans, A.A., On the importance of blind testing in archaeological science: the example from lithic functional studies, Journal of Archaeological Science (2013), http://dx.doi.org/10.1016/j.jas.2013.10.026