Noname manuscript No.
(will be inserted by the editor)

Testing the motor and cognitive foundations of Paleolithic social transmission

Justin Pargeter 1, 2 · Megan Beney Kilgore 3 · Cheng Liu 3 · Dietrich Stout 3 ·

Received: date / Accepted: date

Abstract Stone tools provide key evidence of human cognitive evolution but remain difficult to interpret. Toolmaking skill-learning in particular has been understudied even though: 1) the most salient cognitive demands of toolmaking should occur during learning, and 2) variation in learning aptitude would have provided the raw material for any past selection acting on tool making ability. However, we actually know very little about the cognitive prerequisites of learning under different information transmission conditions that may have prevailed during the Paleolithic. This paper presents results from a pilot experimental study to trial new experimental methods for studying the effect of learning conditions and individual differences on Oldowan flake-tool making skill acquisition. We trained 23 participants for 2 hours to make stone flakes under two different instructional conditions (observation only vs. direct active teaching) employing appropriate raw materials, practice time, and real human interaction. Participant performance was evaluated through analysis of the

Grants or other notes about the article that should go on the front page should be placed here. General acknowledgments should be placed at the end of the article.

Justin Pargeter 1, 2
1. Department of Anthropology, New York University, New York, NY, USA; 2. Palaeo-Research Institute, University of Johannesburg, Auckland Park, South Africa
E-mail: justin.pargeter@nyu.edu

Megan Beney Kilgore 3
3. Department of Anthropology, Emory University, Atlanta, GA, USA
E-mail: megan.elizabeth.beney@emory.edu

Cheng Liu 3
3. Department of Anthropology, Emory University, Atlanta, GA, USA
E-mail: raylc1996@outlook.com

Dietrich Stout 3
3. Department of Anthropology, Emory University, Atlanta, GA, USA
E-mail: dwstout@emory.edu

stone artifacts produced. Performance was compared both across experimental groups and with respect to individual participant differences in grip strength, motor accuracy, and cognitive function measured for the study. Our results show aptitude to be associated with fluid intelligence in a verbally instructed group and with a tendency to use social information in an observation-only group. These results have implications for debates surrounding the cumulative nature of human culture, the relative contributions of knowledge and know-how for stone tool making, and the role of evolved psychological mechanisms in "high fidelity" transmission of information, particularly through imitation and teaching.

Keywords Oldowan · Stone toolmaking · Social learning · Individual variation · Cognitive aptitudes · Motor skills ·

1 Introduction

Stone tools have long been seen as a key source of evidence for understanding human behavioral and cognitive evolution (Darwin 1871; Oakley 1949; Washburn 1960). Pathbreaking attempts to infer specific cognitive capacities from this evidence largely focused on the basic requirements of tool production (Isaac 1976; Wynn 1979; Gowlett 1984; Wynn and Coolidge 2004). More recently, increasing attention has been directed to the processes and demands of stone tool making skill acquisition (Roux, Bril, and Dietrich 1995; Stout 2002; Stout et al. 2005; Geribàs, Mosquera, and Vergès 2010; Nonaka, Bril, and Rein 2010; Stout et al. 2011; Putt, Woods, and Franciscus 2014; Hecht, Gutman, Khreisheh, et al. 2015; Duke and Pargeter 2015; Morgan et al. 2015; Stout and Khreisheh 2015; Lombao, Guardiola, and Mosquera 2017; Putt et al. 2017; Cataldo, Migliano, and Vinicius 2018; Putt, Wijeakumar, and Spencer 2019; Pargeter, Khreisheh, and Stout 2019; Pargeter et al. 2020). This is motivated by the expectation that the most salient cognitive demands of tool making should occur during learning rather than routine expert performance (Stout and Khreisheh 2015) and by interest in the relevance of different social learning mechanisms such as imitation (Rein, Nonaka, and Bril 2014; Stout et al. 2019), emulation (Tehrani and Riede 2008; Wilkins 2018), and language (Ohnuma, Aoki, and Akazawa 1997; Putt, Woods, and Franciscus 2014; Morgan et al. 2015; Lombao, Guardiola, and Mosquera 2017; Putt et al. 2017; Cataldo, Migliano, and Vinicius 2018) to the reproduction of Paleolithic technologies.

Studies investigating these questions have used a range of different experimental designs (e.g., varying technological goals/instructions, training times, raw materials, live vs. recorded instruction, lithic/skill assessment metrics, pseudo-knapping tasks etc.) and reached disparate conclusions regarding the neurocognitive and social foundations of skill acquisition. It is plausible that these discordant results reflect actual diversity in how humans acquire and master stone tool making skills. However, this failure of results to generalize across artificial experimental manipulations (cf. Yarkoni 2020) also raises doubts regarding the external validity (Eren et al. 2016) of conclusions with respect to real-world Paleolithic learning contexts. To address this, we conducted an exploratory study that draws on lessons from previous research in an attempt to balance the pragmatic and theoretical tradeoffs inherent in experimental studies of stone knapping skill acquisition (Pargeter, Khreisheh, and Stout 2019; Stout and Khreisheh 2015).

Learning real-world skills like stone knapping is highly demanding of time and materials and difficult to control experimentally without sacrificing generalizability to real world conditions. Prior efforts have attempted to navigate these challenges by using various combinations of 1) inauthentic raw materials that are less expensive, easier to standardize, and/or easier to knap, 2) video-recorded instruction that is uniform across participants and less demanding of experimenter time, 3) short learning periods, 4) small sample sizes, and 5) single learning conditions. The difficulty of interpreting results from this growing literature led Stout and Khreisheh (2015: 870, emphasis original) to call for “studies with sufficient sample sizes to manipulate learning conditions (e.g. instruction, motivation) and assess individual variation (e.g. performance, psychometrics, neuroanatomy) that *also* have realistic learning periods.” The current study attempts to strike a viable balance between these demands by investigating early-stage learning of a relatively simple technology (least effort, “Oldowan,” flake production (Reti 2016; Shea 2016) under two instructional conditions while collecting data on individual differences in strength, coordination, cognition, social learning, self-control, and task engagement. Unlike any previous study, this allows us to address the likelihood that group effects of training conditions might be impacted by interactions with individual participant differences in aptitude, motivation, or learning style.

We focus on early stage learning because it has been found to be relatively rapid, variable across individuals, and predictive of later outcomes (Pargeter, Khreisheh, and Stout 2019; Stout and Khreisheh 2015; Putt, Wijeakumar, and Spencer 2019), and thus provides a reasonable expectation of generating meaningful data on skill and learning variation while minimizing training costs. Moreover, understanding the minimum training times necessary to detect changes in tool making skill will help archaeologists design more realistic and cost-effective experiments. To further manage costs, we limited our study to only two learning conditions (observation only vs. active teaching). This targets a key controversy in human evolution, namely the origins of teaching and language (Gärdenfors and Högberg 2017; Morgan et al. 2015), while avoiding highly artificial manipulations of dubious relevance to real-world Paleolithic learning. These choices allowed us to invest more in other aspects of research design that we identified as theoretically important, including measurement of individual differences in cognition and behavior, inclusion of an in-person, fully interactive teaching condition, and use of naturalistic raw materials. Sample size remained small in this internally funded exploratory study but could easily be scaled up at funding levels typical of pre- and post-doctoral research grants in archaeology.

1.1 **Individual Differences**

"*The many slight differences... being observed in the individuals of the same species inhabiting the same confined locality, may be called individual differences... These individual differences are of the highest importance to us, for they are often inherited ... and they thus afford materials for natural selection to act on and accumulate...*" (Darwin 1859, Chapter 2)

Individuals vary in aptitude and learning style for particular skills (Jonassen and Grabowski 1993) but this has largely been ignored in studies of knapping skill acquisition, which have instead focused on group effects of different experimental conditions. There are good pragmatic reasons for this, as individual difference studies typically require larger sample sizes and additional data collection. However, overlooking these distinctions is not ideal since individual differences can provide valuable insight into the mechanisms, development, and evolution of cognition and behavior (Boogert et al. 2018). In particular, patterns of association between cognitive traits and behavioral performance can be used to test hypotheses about the cognitive demands of learning particular skills and the likely targets of natural selection acting on aptitude. More prosaically, individual differences can introduce an unexamined and uncontrolled source of variation in group level results. This is especially true in the relatively small "samples of convenience" typical of experimental archaeology.

While testing hypotheses in evolutionary cognitive archaeology remains a considerable challenge (Wynn 2017), investigation of individual variation in modern research participants represents one promising direction. For any particular behavior of archaeological interest, it is expected that standing variation in modern populations should remain relevant to normal variation in learning aptitude. The presence of trait variation without impact on learning aptitude would provide strong evidence against the plausibility of the proposed evolutionary relationship. An absence of variation (i.e., past fixation and rigorous developmental canalization) is not expected given the known variability of human brains and cognition (Sherwood and Gómez-Robles 2017; Barrett 2020). Any confirmatory findings of trait-aptitude correspondence would then have the testable implication that humans should be evolutionarily derived along the same dimension (e.g. Hecht, Gutman, Bradley, et al. 2015).

To date, a small number of "neuroarchaeological" studies have reported associations between individual knapping performance and brain structure or physiological responses. Hecht et al. (2015) reported training-related changes in white matter integrity (fractional anisotropy [FA]) that correlated with individual differences in practice time and striking accuracy change. The regional patterning of FA changes also varied across individuals, with only those individuals who displayed early increases in FA under the right ventral precentral gyrus (premotor cortex involved in movement planning and guidance) showing striking accuracy improvement over the training period. Putt et al. (2019) similarly found that the proportion of flakes to shatter produced by individuals during handaxe making correlated with dorsal precentral gyrus (motor cortex) activation. Pargeter et al. (2020) used a flake prediction paradigm (modeled after Nonaka, Bril, and Rein 2010) to confirm that striking force and accuracy are important determinants of handaxe-making success. These findings all point to the central role of perceptual-motor systems (Stout and Chaminade 2007) and coordination (Roux, Bril, and Dietrich 1995) in knapping skill acquisition. In addition, Putt et al. (2019) also found successful flake production to be associated with prefrontal (working memory/cognitive control) activation and Stout et al. (2015) found that prefrontal activation correlated with success at a strategic judgement (platform selection) task which in turn was predictive of success at out-of-scanner handaxe production. Such investigations are thus starting to chart out the more specific contributions of different neural systems to particular aspects of knapping skill acquisition. To date, however, the cognitive/functional interpretation of systems identified in this manner has largely relied on informal reverse inference (reasoning backward from observed activations to inferred mental processes) from published studies of other tasks that activated the same regions, an approach which is widely regarded as problematic (Poldrack 2011).

Here we take a more direct, psychometric approach to measuring individual differences in perceptual-motor coordination and cognition. Psychometric instruments (e.g., tasks, questionnaires) are designed to assess variation in cognitive traits and states, such as fluid intelligence, working memory, attention, motivation, and personality, that have been of theoretical interest to cognitive archaeologists (e.g., Wynn and Coolidge 2016). It is thus surprising that they have been almost entirely neglected in experimental studies of knapping skill. In the only published example we are aware of, Pargeter et al. (2019) reported significant effects of variation in planning and problem solving (Tower of London test (Shallice, Broadbent, and Weiskrantz 1982)) and cognitive set shifting (Wisconsin Card Sort test (Grant and Berg 1948)) on early stage handaxe learning. Of course, cognition is not the only thing that can affect knapping performance. Flake prediction experiments highlight the importance of regulating movement speed/accuracy trade-offs (Nonaka, Bril, and Rein 2010; Pargeter et al. 2020) and studies of muscle recruitment (Marzke et al. 1998) and manual pressure (Williams-Hatala et al. 2018; Alastair J. M. Key and Dunmore 2018) during knapping highlight basic strength requirements. Along these lines, Key and Lycett (2019) found that individual differences in hand size, shape, and especially grip strength were better predictors of force loading during stone tool use than were attributes of the tools themselves. However, we are unaware of any such studies of biometric influences on variation in knapping success. Finally, the time and effort demands of knapping skill acquisition suggest that differences in personality (e.g., self-control and "grit" (Pargeter, Khreisheh, and Stout 2019), motivation (Stout 2002), and social vs. individual learning strategies (Miu et al. 2020) might also affect learning outcomes. We are again unaware of any previous studies that have assessed such effects. In this study, we assessed all participants with a battery of tests including grip strength, movement speed/accuracy, spatial working memory, fluid intelligence, self-control, tendency to use social information, and motivation/engagement with the tool making task. We were particularly interested in the possibility that these variables might not only impact learning generally, but might also have different effects under different learning conditions.

1.2 **Teaching, Language, and Tool Making**

A creature that learns to make tools to a complex pre-existing pattern. . . must have the kind of abstracting mind that would be of high selective value in facilitating the development of the ability to communicate such skills by the necessary verbal acts. (Montagu 1976: 267)

Possible links between tool making and language have been a subject of speculation for nearly 150 years (Engles 2003, [1873]), if not longer (Hewes 1993), although compelling empirical tests have remained elusive. Over 25 years ago, Toth and Schick (1993) suggested that experiments teaching modern participants to make stone tools in verbal and non-verbal conditions could test the importance of language in the social reproduction of Paleolithic technologies. Ohnuma et al. (1997) were the first to implement this suggestion in a study of Levallois flake production, followed by more recent studies of handaxe making (Putt, Woods, and Franciscus 2014; Putt et al. 2017) and simple flake production (Morgan et al. 2015; Cataldo, Migliano, and Vinicius 2018; Lombao, Guardiola, and Mosquera 2017). This reflects recent interest in the hypothesis that language might be an adaptation for teaching (e.g., Laland 2017; Stout and Chaminade 2012). Teaching and learning demands of Paleolithic tool making would thus provide evidence of selective contexts favoring language evolution (Stout 2010; Morgan et al. 2015; Montagu 1976).

Toth and Schick (1993) were, however, careful to point out that extinct hominid learning strategies and capacities might differ from modern experimental participants. Even leaving aside potential species differences in social learning (cf. Morgan et al. 2015; Stout et al. 2019), reliance on explicit verbal instruction varies widely across modern human societies (e.g., Boyette and Hewlett 2017). The WEIRD (Western, educated, industrialized, rich, democratic (Henrich, Heine, and Norenzayan 2010)) teachers and learners typical of knapping experiments arguably represent an extreme bias toward such instruction. Simply instructing such participants not to speak during an experiment (or to demonstrate but not gesture, etc. (Morgan et al. 2015)) is likely to underestimate the efficacy of non-verbal teaching and learning in cultural contexts where it is more common, let alone in a hypothetical pre-linguistic hominid species.

Such concerns are exacerbated in experiments using pre-recorded instructional videos or extremely short training periods. Video does not allow the interactive teaching that is favored even in formal academic knapping classes (e.g., Shea 2015) and is almost certainly typical of traditional learning contexts (e.g., Stout 2002). It is not known how video presentation affects the efficacy of teaching generally, or the relative effectiveness of different forms of instruction. Going further, some experiments have manipulated the presence/absence of verbal instruction by presenting the same video with and without sound (Putt et al. 2017) or the sound track without the video (Cataldo, Migliano, and Vinicius 2018). While this provides experimental control, it does not allow the instructor to adjust their multi-modal (Levinson and Holler 2014) communication strategies as they would naturally do, for example through pointing and pantomime. To simply remove a communication channel without allowing any such adaptation is highly artificial and risks generating results that cannot be generalized beyond the specific context of the experiment (Yarkoni 2020). Similarly, unnaturally short training periods (e.g., 5-15 minutes (Morgan et al. 2015; Lombao, Guardiola, and Mosquera 2017)) might misrepresent the relative efficacy of different teaching strategies under more realistic conditions (Whiten 2015; Stout and Khreisheh 2015). Even the longest training times to date (Stout and Khreisheh 2015; Pargeter, Khreisheh, and Stout 2019) have not produced knapping skills comparable to relevant archaeological examples, and were achieved by limiting sample size and using only one teaching condition.

For these reasons, we sought to explore a middle path between experimental expedience and realism by limiting our experiment to two relatively naturalistic learning conditions and a moderate learning period of two hours. As in previous experiments (Stout et al. 2011; Hecht, Gutman, Khreisheh, et al. 2015; Pargeter, Khreisheh, and Stout 2019) the first condition was unrestricted, interactive instruction in small groups, essentially reproducing the "natural" teaching/learning context familiar (cf. Shea 2015) to our WEIRD instructor and student participants. The second condition allowed observation only, with the experimenter visible making flakes but not interacting in any way with learners. This absence of teaching is again a familiar social context for our participants and did not require any novel behaviors from the instructor. It matches the "imitation/emulation" condition of Morgan et al. (2015) although we make no assumptions regarding learning mechanisms. We did not include a "reverse engineering" or "end-state emulation" condition in which only finished products were visible. This has been advocated as an important baseline or control condition (Whiten 2015) to distinguish observational from individual learning, but is not likely to model any typical Paleolithic learning context nor to stand as an adequate proxy for the cognition of hominid species with different social learning capacities. There is no reason to assume neurocognitive and behavioral processes of reverse-engineering problem solving in modern humans (e.g., Allen, Smith, and Tenenbaum 2020) approximate the social learning processes of hominids with more ape-like action observation/imitation capacities (Hecht, Gutman, et al. 2013; Hecht, Murphy, et al. 2013; Stout et al. 2019).

We selected a two-hour learning period for both pragmatic and theoretical reasons. Pargeter et al. (Pargeter, Khreisheh, and Stout 2019) found that even ˜90 hours of fully interactive instruction and practice was insufficient to achieve handaxe-making skills comparable to the later Acheulean site of Boxgrove (García-Medrano et al. 2019; Stout et al. 2014), and estimated actual time to mastery as ranging from 121 to 441 hours for different participants. However, they observed the greatest, fastest, and most individually variable skill increases during the first 20 hours of practice. In addition, initial performance was moderately correlated with later achievement. This suggests that studying early-stage learning may be a pragmatic alternative, especially for research investigating individual differences in aptitude. Studies of simple flake production similarly document large initial variation (Stout and Khreisheh 2015) and rapid early progress (Stout and Khreisheh 2015; Putt, Wijeakumar, and Spencer 2019; Stout and Semaw 2006). We designed the current study to test the utility of studying learning and variation during the first two hours of simple flaking instruction/practice, in hopes of finding a viable compromise between experimental realism and cost

1.3 Raw materials and knapping skill

Lithic raw materials vary in size, shape, and fracture mechanical properties that affect the difficulty of achieving different knapping goals (Eren et al. 2014). Unfortunately, it can be difficult and/or expensive to procure authentic raw materials. Experimental studies of knapping skill have often used proxy materials such as flint (Morgan et al. 2015; Nonaka, Bril, and Rein 2010; Cataldo, Migliano, and Vinicius 2018), limestone (Stout and Semaw 2006), porcelain (Khreisheh, Davies, and Bradley 2013), or heat-treated chert (Putt, Woods, and Franciscus 2014; Putt et al. 2017; Putt, Wijeakumar, and Spencer 2019)to model Oldowan and early Acheulean technologies executed in other materials. As well as being more readily available, these proxies are generally easier to knap. This has the benefit of reducing required practice time, but it is unclear how it might affect learning demands more generally or the efficacy of different learning conditions/strategies specifically.

To address this, some studies have attempted to more closely match experimental and archaeological raw material types (Stout et al. 2011; Duke and Pargeter 2015; Pargeter, Khreisheh, and Stout 2019). However, raw materials vary across individual clasts within as well as between types. This has led to interest in standardizing experimental core morphology (Nonaka, Bril, and Rein 2010) and composition, even if this means using artificial materials such as porcelain (Khreisheh, Davies, and Bradley 2013), brick (Geribàs, Mosquera, and Vergès 2010; Lombao, Guardiola, and Mosquera 2017), or foam blocks (Schillinger, Mesoudi, and Lycett 2014). Such manipulations enhance experimental control and internal validity (Eren et al. 2016) at the expense of external generalizability to actual archaeological conditions. Specifically, they allow more robust results from smaller samples but eliminate a core element of real-world knapping skill: the ability to produce consistent results from variable materials (Pelegrin 1990; Stout 2013). For example, Pargeter et al. (2020) found that predicting specific flaking outcomes on actual handaxe preforms was both more difficult and less technologically important than expected from previous work with standardized, frustum-shaped cores (Nonaka, Bril, and Rein 2010). The alternative to control is to incorporate raw material size, shape, and composition as experimental variables (e.g., Stout et al. 2019). This allows consideration of raw material selection and response to variation as aspects of skill but correspondingly increases the sample sizes required to identify patterning. In considering these issues, we again chose to explore a middle path between pragmatism and realism by employing commercially purchased basalt similar to that known from East African Oldowan sites, allowing clast size and shape to vary within set limits, and selecting the particular clasts provided to each participant to approximate the same distribution.

2 Materials and Methods

This research was approved by the Emory Institutional Review Board (IRB00113024). All participants provided written informed consent and completed a video release form (`https://databrary.org/support/irb/release-template.html`).

2.1 **Participants**

Twenty-four adult participants with no prior stone knapping experience were recruited from the Emory community using paper fliers and e-mail listserv advertisements. We were unable to replace one participant who failed to attend their scheduled session, resulting in a total sample of 23. Eleven participants (6 female, 5 male) completed the Untaught condition and 12 (8 female, 4 male) completed the Taught condition.

2.2 **Study Visit**

Participants were asked to visit the Paleolithic Technology Lab at Emory University to complete one three-hour session. Participants were scheduled to attend in six groups of four, however one of these groups had only three participants due to a no-show on the day of the experiment. Each visit began with the collection of individual differences measures, which took approximately one hour. After that, participants undertook 105 minutes (two hours minus a 15-minute break after 1 hour) of stone tool making practice. This session was video-recorded, and all lithic products were collected. After the tool making task, participants completed an "exit questionnaire" comprising the Intrinsic Motivation Inventory (see below).

Participants were compensated for their time with a $30 gift card. They also had the opportunity to earn a performance bonus of $5, $10, $15 or $20 on the gift card. They were told that this bonus would depend on "how well they did" on the last core of their practice session. The actual performance measure was not specified, but in order to allow on the spot payment a simple measure of the percentage of starting weight removed from the final core was used such that: > 30% earned \$5, > 40% earned \$10, > 50% earned \$15, > 75% earned \$20.

2.3 **Individual Difference Measures**

We used five individual difference measures for this study:

1) Grip strength was measured in kilograms using an electronic hand dynamometer (Camry EH101). Strength was measured twice and the higher value recorded. Grip strength is a simple measure that is well correlated with overall muscular strength (Wind et al. 2010) and a range of other health and fitness measures (Sasaki et al. 2007). It is hypothesized to be relevant to generating kinetic energy for fracture initiation (Nonaka, Bril, and Rein 2010) as well as control and support of the hammerstone (Williams-Hatala et al. 2018) and core (Faisal et al. 2010; Alastair J. M. Key and Dunmore 2015).
2) Motor accuracy was assessed using a "Fitts Law" reciprocal tapping task. Fitts Law describes the trade-off between speed and accuracy in human movement, classically measured by tapping back and forth between two targets of varying size and spacing (Fitts 1954). Archaeologists have proposed (Stout 2002; Pargeter et al. 2020) that management of this trade-off is critical to the accurate application of appropriate force seen in skilled knapping (Nonaka, Bril, and Rein 2010; Roux, Bril, and Dietrich 1995). We implemented this test on a Surface Pro tablet running free software (FittsStudy Version 4.2.8, default settings) developed by the Accessible Computing Experiences lab (Jacob O. Wobbrock, director) at the University of Washington (`depts.washington.edu/acelab/proj/fittsstudy/index.html`). Participants use a touchscreen pen to tap between ribbons on the screen, with average movement time as the performance metric.
3) Visuospatial working memory is the capacity to "hold in mind," which researchers have hypothesized to be important in stone toolmaking performance (Coolidge and Wynn 2005). It also might support a learning process known as 'chunking,' in which multiple items or operations are combined into summary chunks stored in long term memory, that is thought to be important in the acquisition of knapping and other skills (Pargeter, Khreisheh, and Stout 2019). We measured visuospatial working memory using a free n-back task (wmp.education.uci.edu/software/) developed by the Working Memory and Plasticity Laboratory at the University of California, Irvine (Susanne Jaeggi, PI) and implemented in E-Prime software on a desktop computer. In this task, participants are asked to remember the position of blue squares presented sequentially on the screen and touch

a key when the current position matches that 1, 2, 3. . . n iterations back. Progression to blocks with increasing values of n is contingent on exceeding a threshold success rate. Performance was measured as the highest n achieved.

4) Fluid intelligence (Cattell 1963) refers to the capacity to engage in abstract reasoning and problem solving in a way that is minimally dependent on prior experience. It complements "crystallized intelligence" (the ability to apply learned procedures and knowledge) as one of the two factors (gf, gc) comprising so-called "general intelligence" (g). Fluid intelligence is closely related to the executive control of attention and manipulation of information held in working memory (Engle 2018)(Engle 2018). It is hypothesized to support technological innovation (Coolidge and Wynn 2005) and/or the intentional learning of new skills (Unsworth and Engle 2005; Stout and Khreisheh 2015). We measured fluid intelligence using the short version (Bilker et al. 2012) of the classic Raven Progressive Matrices task, which requires participants to complete increasingly difficult pattern matching questions.
5) The use of social information for learning and decision making varies across individuals and societies (Molleman, Kurvers, and van den Bos 2019). Such variation is a key topic for understanding social learning and cultural evolutionary processes (Kendal et al. 2018; Heyes 2018; Miu et al. 2020) and represents a potential confound for assessing experimental effects of different social learning conditions. We measured participants' tendency to rely on social information vs. their own insights using the Berlin Estimate AdjuStment Task (BEAST) developed by Molleman et al. (2019). In this task, participants are present with large arrays of items on a screen and asked to estimate the number present. They are then provided with another person's estimate and allowed to provide a second estimate. The participants' average adjustment between first and second estimates provides a measure of their propensity to rely on social information.

2.4 **Stone Tool Making**

After individual difference testing, participants engaged in a 2-hour stone tool making session, with a 15-minute break after 1 hour. Participants were instructed not to seek out additional training or information on stone tool making (i.e., via the internet) during these breaks. Each group of participants was randomly assigned to one of two experimental conditions: no teaching or teaching. In both conditions, participants were first given an opportunity to inspect and handle examples (**Figure**) of the kind of stone tools (flakes) they are being asked to produce. They were told that their objective was to produce as many flakes as possible from the materials provided. This meant that even the untaught condition included some minimal instruction (being told the objective) , however this was considered to be unavoidable without creating a much more elaborate and naturalistic context in which participants would develop their own technological goals. Such a design would also be expected to increase behavioral variability, demanding correspondingly larger samples of participants to identify patterns and making direct comparisons with the taught condition.

2.4.1 ***Raw Materials***

Each participant was provided with 9 cores for use over the 2-hour experiment. These cores were produced from larger chunks of a fine-grained basalt purchased from neolithics.com by fracturing them with a sledgehammer. This produced irregular, angular chunks for use in the experiment, weighing between 459g - 1876g (mean = 975g). All cores were weighed, measured (Length, Width, Thickness), and painted white so that new fracture surfaces could be discriminated from those created during production. Cores were sorted by shape and weight and then distributed evenly to each participant. As a result, there were no significant difference across participants in the mean weight (ANOVA, df = 22, F=0.3, p = 0.9; Levene test of homogeneity of variance = 1.04, df1=22, df2 = 184, p = 0.4) or shape (Length x Width/Thickness: ANOVA, df = 22, F=0.4, p = 0.9; Levene statistic = .6, df1=22, df2 = 184, p = 0.9) of cores provided. This was also true comparing the two experimental conditions (Taught vs. Untaught mean weight = 1001g vs. 956g, t = 1.24, df = 205, p = 0.2, Levene's Test F = 0.6, p = 0.4; mean shape = 221.43 vs. 221.45, t = -0.003, df = 205, p = 0.9, Levene's Test F = 3.8, p = 0.05). Participants were, however, allowed to choose which cores to work on so that differences in the weight and shape of cores actually used across participants and conditions could still emerge as a result of selection bias.

Sixty pounds of 3-to-5 inch basalt "Mexican Beach Pebbles" were purchased from a landscaping supply company for use as hammerstones in the experiment. Of these, 90 were selected as suitable for use. These weighed between 213g-1360g (mean = 425) and varied in elongation (L/W = 1.01 to 2.65) and relative thickness (LxW/T = 90.48 to 283.67). Forty-five stones were placed in the middle of the knapping area (Figure 2) for participants to freely choose from during the experiment. Broken hammerstones were replaced from the reserve to maintain a consistent number and range of choices. Each hammerstone was numbered and participants' choices were recorded along with the number of the core(s) being worked on with a particular hammerstone.

2.4.2 ***Experimental Conditions***

In both conditions, three researchers were present to record activities and collect materials. Participants were seated in a circle (Shea 2015) and experiments were video recorded using two cameras (Figure 1). Participants were free to select hammerstones from the common pile and to work on any or all of their nine assigned cores in any order they preferred. However, each core and all associated debitage were collected before participants were allowed to start working on a new core, so it was not possible to partially work and then return to a particular core later. The order of cores used and associated hammerstones were recorded for each participant during the experiment.

In the untaught condition, a researcher (DS) sat with the participants and made stone tools but remained silent and made no effort to facilitate learning (e.g., through gesture, modified performance, facial expression, attention direction, or verbal instruction). Over the 2-hour period, the researcher completely reduced four cores (one every ˜30 minutes). Participants were not restricted from talking to each other, as this would create an unnatural and potentially stressful social context that might affect learning. Participants were asked to avoid any form of communication about the tool making task specifically, and they complied with this request. Participants in this condition thus had the opportunity to observe tool making by an expert and/or by other learners, should they choose to do so, but received no intentional instruction.

In the Taught condition, there were no restrictions on participant interaction and the researcher engaged in direct active teaching (Kline 2015) of toolmaking techniques through verbal instruction, demonstration, gesture, and shaping of behavior. The instructor has a moderate level of experience teaching basic knapping skills to students in undergraduate archaeology classes and to participants in previous knapping research (e.g., Stout et al. 2011). The pedagogical strategy employed was based on the instructor's own learning experiences and theoretical interpretations (e.g., Pargeter et al. 2020), and focused on coaching participants in effective body postures, movement patterns, and grips as well as the assessment of viable core morphology.

Lithic Analysis

All finished cores were weighed and measured (L, W, T). Delta weight was calculated as (Start weight-End weight)/Start weight. All detached pieces (DPs) were collected and weighed. We did not sort DPs into types (e.g., whole flakes, fragments) as this would have greatly increased processing time and it is not clear that such distinctions add relevant information regarding utility/desirability beyond that supplied by metrics (Stout et al. 2019). All DPs larger than 40mm in maximum dimension were photographed and measured. It is conventional in Early Stone Age lithic analysis to employ a 20 mm cut-off. We selected a higher threshold for both pragmatic (analysis time) and theoretical reasons. Flake use experiments have shown that flakes weighing less than 5–10 g or with a surface area below 7–10 cm2 (Prasciunas 2007) or with a maximum dimension <50-60 mm (Alastair J. M. Key and Lycett 2014) become markedly inefficient for basic cutting tasks. Similarly, data from Oldowan replication experiments (Stout et al. 2019) show that the utility index (flake cutting edge/flake mass1/3) * (1 - exp[-0.31 * (flake maximum dimension – 1.81)]) developed by Morgan et al. (2015) falls off rapidly below 40mm maximum dimension (Mean Utility < 40mm = 0.508; >=40mm = 0.946; t= 11.99, df = 707, $p < 0.000$). By including weight in our cut-off criteria we also avoid skewing the flake shape distribution by selectively retaining long, thin pieces (i.e., MD $>$ 40, weight $<$ 5g) while discarding rounder pieces of similar (or greater) weight and area.

For measurement, DP length was defined as the longest axis and width as the maximum dimension orthogonal to length. Thickness was defined as the maximum dimension orthogonal to the plane formed by L and W and was measured using calipers. L, W, and plan-view area measurements were taken from photographs captured using a Canon Rebel T3i fitted with a 60 mm macro lens and attached to a photographic stand with adjustable upper and lower light fittings. The camera was positioned directly above the flakes and kept at a constant height. DPs were positioned irrespective of any technological features so that the longest axis was vertical, and the wider end was placed toward the bottom of the photograph.

Photographs were post-processed using Equalight software to adjust for lens and lighting falloff that result from bending light through a lens and its aperture which can affect measurements taken from photographs. Each image was shot with a scale that was then used to rectify the photograph's pixel scale to a real-world measurement scale in Adobe Photoshop. Images were converted to binary black and white format and silhouettes of the tools were extracted in Adobe Photoshop. We then used a custom ImageJ (Rueden et al. 2017) script (Pargeter, Khreisheh, and Stout 2019) to measure DP length and take nine width measurements at 10% increments of length starting at the base of each DP. We used the built-in ImageJ tool to measure DP area. A "Proportion Larger DPs" was calculated per core as the combined weight of all DPs >40mm in maximum dimension and 5g in weight divided by the weight of all DPs. Higher values show cores with proportionally more large DPs.

2.5 **Statistical Analyses**

To evaluate the association between psychometric, motor-skill, and training measures and technological outcomes, we adopted an information-theoretic approach (Burnham and Anderson 2002). Information-theoretic approaches provide methods for model selection using all possible combinations of variables while avoiding problems associated with significance-threshold stepwise selection. We used the corrected Akaike information criterion (AICc) to rate each possible combination of predictors on the balance between goodness of fit (likelihood of the data given the model) and parsimony (number of parameters). The AICc consists of the log likelihood (i.e., how well does the model fit the data?) and a penalty term for the number of parameters that must be estimated in the model (i.e., how parsimonious is the model?), with a correction for small sample sizes (AICc converges to the standard AIC at large samples). A lower AICc indicates a more generalizable model and we used it to compare and rank various possible models. Each analysis begins with a full model that includes all predictors of interest. All possible combinations of predictors are then fit, and the resulting models are ranked and weighted based on their AICc. The "best" model is chosen because it has the lowest AICc score.

Continuous predictors were centered such that zero represents the sample average, and units are standard deviations. The full model was fitted with the lm function in R 3.2.3, and the glmulti package (**Bartoń?**) was used for multi-modal selection and model comparison.

3 Results

Following a recent protocol to enhance the reproducibility and data transparency of archaeological research (Marwick 2017), detailed results of all analyses and assessments of the data structure are available in our paper's supplementary materials and through Github (`https://github.com/Raylc/PaST-pilot`). Here we limit discussion to the major findings regarding flaking performance and individual differences. The purpose of this section is to determine whether training impacted subject flaking performance and if any of the subject's individual psychometric and motor-skill aptitudes predict flaking performance.

3.1 **Principal Component analyses**

The following two sections outline factor analyses designed to summarize our main study metrics tracking individual variation in flake sizes and shapes and lithic performance measures.

3.1.1 ***Flake size and shape***

To better understand the relationship between flake shape and training/individual variation, we entered our nine flake linear plan measurements along with maximum flake length and thickness into a principal component analysis (PCA) from which summary coordinates were extracted. Bartlett's Test of Sphericity was significant (χ^2 (10) =4480, p < .01) indicating that the set of variables are adequately related for factor analysis.

The analysis yielded three factors explaining a total of 90% of the variance for the entire 11 measurement variable set **(Table**). Factor 1 tracks flake size with higher scores indicating larger flakes since all 11 measures load positively on this factor. Factor 2's loadings track the increasing relationship between thickness, length, and flake width. As factor 2 scores increase, flakes get thicker, longer, and narrow, resembling irregular splinters. Factor 3 tracks the relationship between flake proximal and distal width relative to thickness. As factor 3 scores go up, flakes get thinner and narrower at the distal ends and wider at the base. Factor 3 therefore tracks flakes with a typical shape having a thin cross-section, wider base, and narrower tip. We used these three flake shape coordinates to approximate flake size and shape in the project's flake performance factor analysis.

3.1.2 ***Lithic flaking performance measures***

To better understand the relationship between our various lithic performance measurements and to reduce these data dimensionality, we conducted a second principal component analysis examining the study's seven lithic performance measures (mass of flakes relative to flaked core mass, count of large flakes [>40mm and 5g], core delta mass, three flake shape factors, and the total number of cores used). All of these measures except the total cores flaked were summarized for each core. The Bartlett's Test of Sphericity was significant (χ^2 (6) =3950, $p < .01$) indicating that the set of variables are at least adequately related for factor analysis.

The analysis yielded two factors explaining a total of 52.3% of the variance for the entire set of variables. Factor 1 tracks flaking quantity due to high positive loadings on mass flakes/flaked mass, core delta mass, total cores used, and flake shape factor 2. Negative loadings on flake factor 1 (flake size) and 3 (basal/tip width shape) suggest that flaking quantity comes at the expense of producing lots of small and thick stone splinters/chunks. This first factor explained 28% of the variance. Higher factor 1 values reflect higher flaking quantities. The second factor covers 24% of the sample variance. Factor 2 measures one's ability to carefully flake cores due to high positive loadings on large flake count and lower or neutral loadings on mass flakes/flaked mass, core delta mass, and total cores used. The resulting flakes are larger, relatively thinner, and more typically shaped due to high positive loadings on flake factor 1 (size) and flake factor 2 (relative thickness) contrasted with a high negative loading on flake shape factor 3 (basal/tip width). Higher factor two values show increased quality flaking performance.

Figure shows the covariance between our two flake performance principal components. While there is no significant relationship between the two factors, and the two study groups do not show significantly different slopes ($p = 0.3$), there are never-the-less differences between the two groups. Trained individuals are able to increase flake quality while increasing flake quality, whereas untrained individuals increase their flaking quantity at the expense of flake quality.

3.2 **Do trained, untrained, and expert knappers perform differently?**

Here we compare our flaking outcomes (flake size/shape and flaking performance factors) between the trained and untrained groups to see if training in any way impacted their flaking performance. We added a performance comparison (individual flake and core knapping performance metrics) between these two novice groups and our expert knapper to test for differences at different points in the stone flaking skill spectrum.

Table summarizes the group level performance tests. The results show no significant differences in flaking performance between the trained and untrained groups measured by our two flaking performance factors. Trained and untrained individuals overall performed equally as well in terms of flaking quality and quantity. Three-way flake size and shape comparisons between our expert knapper and the two novice groups show significant differences in flake shape factor 2 (relative thickness), but with a very low (<0.01) effect size. This difference is driven by the expert's lower overall relative flake thickness. Regardless of training, our experimental subjects produced flakes that were on average the same size and shape as those of the expert trainer. We did, however, find several significant differences in the three-way comparisons of our individual flaking performance measures. The expert knapper made significantly more large flakes (effect size = 0.14), had a significantly higher core delta mass signal than either of the novice groups (effect size = 0.26), and on average had significantly smaller cores (effect size = 0.27) (**Figure**). All three of these results show either medium or large effect sizes. In all three comparisons, the trained group's data distributions tended towards the expert sample (although they were not significantly different from the untrained group) (**Figures-Core examples too**). These results show that core reduction intensity and large flake production track this experiment's greatest differences between expert and novice performance.

3.3 Does training/practice time impact flaking performance?

Here we use the relative order in which each subject flaked their cores to test for changes in flaking performance across the 2hr experiment. For these analyses, we calculated the relative percentage for each core relative to the total cores knapped by each subject. These relative core use-order percentages were then binned into 20 percent brackets for core-order and group-level comparisons. Flaking outcomes were tracked using the two flake performance factors (quality and quantity flaking). We added the nodule starting mass to track whether training/practice times impacted raw material selection.

Table shows no significant training effects across the two flaking performance measures either as grouped data or between individuals (**Figures)**. This result demonstrated that flaking outcomes did not change dramatically across the study interval. The one significant main training effect related to core starting mass (with a strong main training effect size = 0.25). On average, core starting masses increase as subjects flake more cores, showing that subjects started with smaller nodules first that were smaller and easier to hold/flake. As the experiment wore on, they were left with larger and more challenging nodules. The small main effect of training condition is driven by higher starting nodule masses in the untrained group at the beginning, half way through, and at the end of the experiment.

3.4 Do individual differences in motor skill and psychometric measures predict flaking performance?

One of the experiment's primary goals was to test if measures of individual variation in motor skill and intelligence/motivation predict success in stone flaking. To address this goal, we built two multivariate models examining the relations between our various psychometric measures, subject's motor skill scores, and our two lithic performance factors (quantity flaking and quality flaking). These models enabled us to determine which of the psychometric and motor skill factors are better predictors of a participant's flaking performance in the study.

We have already demonstrated that subjects selected progressively larger nodules throughout the experiment. It is important now to understand whether nodule variability had any impact on our flaking results. Because starting nodule size (mass) and shape were strongly correlated (F [1,157] = 186, $p < 0.01$, R2 = 0.54) we included nodule mass as a covariate to control for any variance in flaking performance that may be driven by nodule differences. Our two motor skill and strength measures (grip strength and Fitt's performance scores) are also strongly correlated (F [1,19] = 15, $p < 0.01$, R2 = 0.41). However, these two measures track complementary components of athleticism (strength vs. speed/accuracy tradeoffs) and so we decided to include both in the model selection process.

We considered all possible interactions between five individual difference measures, core size, training condition, and the two lithic flaking performance factors (quantity and quality flaking) (wherein each subject provides one data point). Each model's continuous predictors (highest n-back level, Raven's Progressive Matrix score, BEAST score, starting nodule mass, Fitt's score, and grip strength) were centered such that zero represents the sample average, and units are standard deviations.

3.4.1 ***Model 1: Individual differences and quantity flaking***

The first full model examined variance in the quantity of flaking tracked by our first performance factor explaining increases in the number of cores used, the degree of reduction on each core, and large/relatively thick flake production. The full model was fitted with the lm function in R 3.2.3, and we used the Glmulti package's automated model selection algorithm to select the best performing model (lowest AICc score) (see methods for further details on the multimodal selection process). The complete model statement is as follows:

Quantity flaking ∼ Training condition + Highest n-back level + Raven's Progressive Matrix score + BEAST score + Fitt's score + Grip strength

From a candidate pool of 55893 possible multivariate models, the best performing model returned an AICc value of 36 (Average AIC = 52). This model comprised the following statement with three main and four interaction effects:

Quantity flaking ∼ Training condition + Highest n-back level + BEAST score + Grip strength + Fitt's score·Raven's Progressive Matrix score + Training condition·Highest n-back level + Training condition·BEAST score + Nodule mass (as control)

This model explains a statistically significant and substantial proportion of variance in quantity flaking (R2 = 0.84, F (8, 12) = 8.2, p < 0.01, adj. R2 = 0.74). A model evaluation comparing training vs. test data subsets shows no evidence for overfitting (Difference in R2 between training and test models = 0.02). A model residuals normality test shows no significant differences with the normal distribution (p = 0.35) indicating that this relationship (as required) is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (whether variance for all observations in our data set are the same) (BP = 8.2, df = 7, p = 0.3).

Table presents this model's coefficients and summary outputs, wherein baseline refers to the untrained condition with all continuous predictors at the sample average. The parameter estimates for the continuous predictors reflect the expected change in utility for 1 standard deviation change in the predictor variable. Significant increases in quantity flaking were found for subjects with training (Est. = 0.63 [0.15, 1.11]), higher n-back levels (Est. = 0.71 [0.29, 1.13]), grip strength (Est. = 0.66 [0.36, 0.95]), and BEAST scores (Est. = 0.41 [0.04, 0.78]) regardless of starting nodule sizes (**Figure**). This suggests that several independent factors tracking the effects of training, social information use, visuo-spatial working memory, and strength improve an individual's ability to flake in greater quantities.

The model produced two significant interactions between training condition and n-back level and BEAST scores. Subjects in the trained group with higher n-back levels show lower quantity flaking scores (Est. = -1.35 [-1.85, -0.84]) while subjects in the trained group with higher BEAST scores also show lower quantity flaking scores (Est. = -0.83 [-1.37, -0.28]) when all the other variables are held constant. At first glance, these results appear to contrast with expectations regarding the effects of visuo-spatial working memory and social information use on success in technological tasks. However, looking at the matter more holistically it is clear that untrained subjects produce flaking quantities in a different way that relies on visuo-spatial working memory rather than taught information. When teaching is provided, being aware of one's surroundings (i.e. copying from other novices) can have a negative effect on flaking quantity. The interaction between BEAST scores and training shows that social information use acts to level the effects of an individual's propensity to "muscle" through the flaking task (both groups converge on an average quantity flaking score with higher BEAST scores). However, the direction of this change is different depending on training with untrained and socially inclined individuals increasing flaking quantities in contrast to trained and socially inclined individuals decreasing flaking quantities.

3.4.2 Model 2: Individual differences and quality flaking

Our second model examining variance in quality flaking follows the same complete model statement we used for the quantity flaking with six covariates. From the same candidate pool size of 55893 possible multivariate models, the best performing model returned an AICc value of 32 (Average AIC = 41). This model comprised the following statement with three main and four interaction effects:

Quality flaking ∼ Highest n-back level + Fitt's score + Grip strength + Fitt's score·BEAST score + Grip strength·BEAST score + Grip strength·Fitt's score + Training condition·Grip strength + Nodule mass (as control)

This model explains a statistically significant and substantial proportion of variance in careful flaking (R2 = 0.78, F (8, 12) = 5.5, $p < 0.01$, adj. R2 = 0.64) in the absence of any main training effects. A model evaluation comparing training vs. test data subsets shows some evidence for overfitting (Difference in R2 between training and test models = 0.17), which is unfortunately unavoidable with our small sample sizes. A model residuals normality test shows no significant differences with the normal distribution (p = 0.13) indicating that this relationship is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (BP = 3.4, df = 8, p = 0.9).

Table presents this model's coefficients and summary outputs following the same data format as **Table**. The results show three significant main effects between our measures of motor-skill, strength, visuo-spatial working memory, and quality flaking. A one unit increase in Fitt's score (Est. = -0.7 [-1.08, -0.31]), grip strength (Est. = -0.9 [-1.32, -0.53]), visuo-spatial working memory (n-back level) (Est. = -0.3 [-0.58, -0.07]), and nodule starting mass (Est. = -0.3 [-0.58, -0.04]) results in decreased flaking quality regardless of training. Collectively, these results show that quality flaking comes at a cost in terms of a subject's motor abilities and that quality flaking decreases as subjects are faced with flaking larger and more challenging nodules.

The quality flaking model shows two significant interaction effects involving grip strength, BEAST scores, and training. The interaction between grip strength and BEAST scores has a significant positive effect on careful flaking (Est. = 0.5 [0.22, 0.86]). **Figure** illustrates this interaction whereby a subject's propensity to use social information mitigates against individual differences in grip strength. Strong subjects with a low propensity to use social information tend to perform worst in terms of quality flaking. This result makes sense if one considers that our quantity flaking model showed how social information equalizes performance levels in trained and untrained subjects.

The second interaction involves grip strength and training. A one unit increase in this interaction results in a significant positive effect on careful flaking (Est. = 1 [0.51, 1.47]), but only in the untrained group. **Figure** illustrates this interaction showing how training helps even out flaking performance differences between stronger and weaker individuals. Stronger individuals perform worse in terms of quality flaking when they are untrained. Trained subjects, regardless of grip strength, perform equally well in terms of quality flaking.

4 Discussion

we have little basis other than personal experience and/or tradition (Callahan 1979; Whittaker 1994; Shea 2015) and theoretical speculation (Stout 2013; Whiten 2015) from which to assess what pedagogical techniques are most effective even in WEIRD contexts. For example, no study to date has considered how variation in teacher skill (Shea 2015) or social relationship to participants might impact learning under different conditions. To properly address these questions would require a major research program, including both cross-cultural comparative studies (Barrett 2020) (Barrett 2020) and more naturalistic study designs. While costly, such research would produce results of broad relevance to anthropologists, biologists, psychologists, and sociologists interested in teaching and learning, well beyond any particular implications for language evolution.

5 Conclusions

6 Acknowledgments

References

Allen, Kelsey R., Kevin A. Smith, and Joshua B. Tenenbaum. 2020. "Rapid Trial-and-Error Learning with Simulation Supports Flexible Tool Use and Physical Reasoning." *Proceedings of the National Academy of Sciences* 117 (47): 29302–10. https://doi.org/10.1073/pnas.1912341117.

Barrett, H. Clark. 2020. "Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency." *Trends in Cognitive Sciences* 24 (8): 620–38. https://doi.org/10.1016/j.tics.2020.05.007.

Bilker, Warren B., John A. Hansen, Colleen M. Brensinger, Jan Richard, Raquel E. Gur, and Ruben C. Gur. 2012. "Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test." *Assessment* 19 (3): 354–69. https://doi.org/10.1177/1073191112446655.

Boogert, Neeltje J., Joah R. Madden, Julie Morand-Ferron, and Alex Thornton. 2018. "Measuring and Understanding Individual Differences in Cognition." *Philosophical Transactions of the Royal Society B: Biological Sciences* 373 (1756): 20170280. https://doi.org/10.1098/rstb.2017.0280.

Boyette, Adam H., and Barry S. Hewlett. 2017. "Autonomy, Equality, and Teaching Among Aka Foragers and Ngandu Farmers of the Congo Basin." *Human Nature* 28 (3): 289–322. https://doi.org/10.1007/s12110-017-9294-y.

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* 2nd ed. New York: Springer-Verlag. https://doi.org/10.1007/b97636.

Callahan, Errett. 1979. "THE BASICS OF BIFACE KNAPPING IN THE EASTERN FLUTED POINT TRADITION: A MANUAL FOR FLINTKNAPPERS AND LITHIC ANALYSTS." *Archaeology of Eastern North America* 7 (1): 1–180. https://www.jstor.org/stable/40914177.

Cataldo, Dana Michelle, Andrea Bamberg Migliano, and Lucio Vinicius. 2018. "Speech, Stone Tool-Making and the Evolution of Language." *PLOS ONE* 13 (1): e0191071. https://doi.org/10.1371/journal.pone.0191071.

Cattell, Raymond B. 1963. "Theory of Fluid and Crystallized Intelligence: A Critical Experiment." *Journal of Educational Psychology* 54 (1): 1–22. https://doi.org/10.1037/h0046743.

Coolidge, Frederick L., and Thomas Wynn. 2005. "Working Memory, Its Executive Functions, and the Emergence of Modern Thinking." *Cambridge Archaeological Journal* 15 (1): 5–26. https://doi.org/10.1017/S0959774305000016.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life.* 1st ed. London: John Murray.

———. 1871. *The Descent of Man, and Selection in Relation to Sex.* 1st ed. London: John Murray.

Duke, Hilary, and Justin Pargeter. 2015. "Weaving Simple Solutions to Complex Problems: An Experimental Study of Skill in Bipolar Cobble-Splitting." *Lithic Technology* 40 (4): 349–65. https://doi.org/10.1179/2051618515Y.0000000016.

Engle, Randall W. 2018. "Working Memory and Executive Attention: A Revisit." *Perspectives on Psychological Science* 13 (2): 190–93. https://doi.org/10.1177/1745691617720478.

Engles, Friedrich. 2003. "The Part Played by Labour in the Transition from Ape to Man." In, edited by Robert C. Scharff and Val Dusek, 71–77. London: Blackwell.

Eren, Metin I., Stephen J. Lycett, Robert J. Patten, Briggs Buchanan, Justin Pargeter, and Michael J. O'Brien. 2016. "Test, Model, and Method Validation: The Role of Experimental Stone Artifact Replication in Hypothesis-Driven Archaeology." *Ethnoarchaeology: Journal of Archaeological, Ethnographic and Experimental Studies* 8 (2): 103–36. https://doi.org/10.1080/19442890.2016.1213972.

Eren, Metin I., Christopher I. Roos, Brett A. Story, Noreen von Cramon-Taubadel, and Stephen J. Lycett. 2014. "The Role of Raw Material Differences in Stone Tool Shape Variation: An Experimental Assessment." *Journal of Archaeological Science* 49: 472–87. https://doi.org/10.1016/j.jas.2014.05.034.

Faisal, Aldo, Dietrich Stout, Jan Apel, and Bruce Bradley. 2010. "The Manipulative Complexity of Lower Paleolithic Stone Toolmaking." *PLOS ONE* 5 (11): e13718. https://doi.org/10.1371/journal.pone.0013718.

Fitts, Paul M. 1954. "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement." *Journal of Experimental Psychology* 47 (6): 381–91. https://doi.org/10.1037/h0055392.

García-Medrano, Paula, Andreu Ollé, Nick Ashton, and Mark B. Roberts. 2019. "The Mental Template in Handaxe Manufacture: New Insights into Acheulean Lithic Technological Behavior at Boxgrove, Sussex, UK." *Journal of Archaeological Method and Theory* 26 (1): 396–422. https://doi.org/10.1007/s10816-018-9376-0.

Gärdenfors, Peter, and Anders Högberg. 2017. "The Archaeology of Teaching and the Evolution of Homo Docens." *Current Anthropology* 58 (2): 188–208. https://doi.org/10.1086/691178.

Geribàs, Núria, Marina Mosquera, and Josep Maria Vergès. 2010. "What Novice Knappers Have to Learn to Become Expert Stone Toolmakers." *Journal of Archaeological Science* 37 (11): 2857–70. https://doi.org/10.1016/j.jas.2010.06.026.

Gowlett, John A. J. 1984. "Mental Abilities of Early Man: A Look at Some Hard Evidence." *Higher Education Quarterly* 38 (3): 199–220. https://doi.org/10.1111/j.1468-2273.1984.tb01387.x.

Grant, David A., and Esta Berg. 1948. "A Behavioral Analysis of Degree of Reinforcement and Ease of Shifting to New Responses in a Weigl-Type Card-Sorting Problem." *Journal of Experimental Psychology* 38 (4): 404–11. https://doi.org/10.1037/h0059831.

Hecht, Erin E., David A. Gutman, Bruce A. Bradley, Todd M. Preuss, and Dietrich Stout. 2015. "Virtual dissection and comparative connectivity of the superior longitudinal fasciculus in chimpanzees and humans." *NeuroImage* 108 (March): 124–37. https://doi.org/10.1016/j.neuroimage.2014.12.039.

Hecht, Erin E., David. A. Gutman, Nada Khreisheh, S. V. Taylor, J. Kilner, A. A. Faisal, Bruce A. Bradley, T. Chaminade, and D. Stout. 2015. "Acquisition of Paleolithic toolmaking abilities involves structural remodeling to inferior frontoparietal regions." *Brain Structure & Function* 220 (4): 2315–31. https://doi.org/10.1007/s00429-014-0789-6.

Hecht, Erin E., David A. Gutman, Todd M. Preuss, Mar M. Sanchez, Lisa A. Parr, and James K. Rilling. 2013. "Process Versus Product in Social Learning: Comparative Diffusion Tensor Imaging of Neural Systems for Action Executionobservation Matching in Macaques, Chimpanzees, and Humans." *Cerebral Cortex* 23 (5): 1014–24. https://doi.org/10.1093/cercor/bhs097.

Hecht, Erin E., Lauren E. Murphy, David A. Gutman, John R. Votaw, David M. Schuster, Todd M. Preuss, Guy A. Orban, Dietrich Stout, and Lisa A. Parr. 2013. "Differences in Neural Activation for Object-Directed Grasping in Chimpanzees and Humans." *The Journal of Neuroscience* 33 (35): 14117–34. https://doi.org/10.1523/JNEUROSCI.2172-13.2013.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "Most People Are Not WEIRD." *Nature* 466 (7302): 29–29. https://doi.org/10.1038/466029a.

Hewes, Gordon W. 1993. "A History of Speculation on the Relation Between Tools and Language." In, edited by Kathleen R. Gibson and Tim Ingold, 20–31. Cambridge: Cambridge University Press.

Heyes, Cecilia. 2018. "Enquire Within: Cultural Evolution and Cognitive Science." *Philosophical Transactions of the Royal Society B: Biological Sciences* 373 (1743): 20170051. `https://doi.org/10.1098/rstb.2017.0051`.

Isaac, Glynn L. 1976. "Stages of Cultural Elaboration in the Pleistocene: Possible Archaeological Indicators of the Development of Language Capabilities." *Annals of the New York Academy of Sciences* 280 (1): 275–88. `https://doi.org/10.1111/j.1749-6632.1976.tb25494.x`.

Jonassen, David H., and Barbara L. Grabowski. 1993. *Handbook of Individual Differences, Learning, and Instruction.* Hillsdale, NJ: Lawrence Erlbaum,.

Kendal, Rachel L., Neeltje J. Boogert, Luke Rendell, Kevin N. Laland, Mike Webster, and Patricia L. Jones. 2018. "Social Learning Strategies: Bridge-Building Between Fields." *Trends in Cognitive Sciences* 22 (7): 651–65. `https://doi.org/10.1016/j.tics.2018.04.003`.

Key, A. J. M., and S. J. Lycett. 2019. "Biometric Variables Predict Stone Tool Functional Performance More Effectively Than Tool-Form Attributes: A Case Study in Handaxe Loading Capabilities." *Archaeometry* 61 (3): 539–55. `https://doi.org/10.1111/arcm.12439`.

Key, Alastair J. M., and Christopher J. Dunmore. 2015. "The Evolution of the Hominin Thumb and the Influence Exerted by the Non-Dominant Hand During Stone Tool Production." *Journal of Human Evolution* 78 (January): 60–69. `https://doi.org/10.1016/j.jhevol.2014.08.006`.

———. 2018. "Manual Restrictions on Palaeolithic Technological Behaviours." *PeerJ* 6 (August): e5399. `https://doi.org/10.7717/peerj.5399`.

Key, Alastair J. M., and Stephen J. Lycett. 2014. "Are Bigger Flakes Always Better? An Experimental Assessment of Flake Size Variation on Cutting Efficiency and Loading." *Journal of Archaeological Science* 41 (January): 140–46. `https://doi.org/10.1016/j.jas.2013.07.033`.

Khreisheh, Nada N., Danielle Davies, and Bruce A. Bradley. 2013. "Extending Experimental Control: The Use of Porcelain in Flaked Stone Experimentation." *Advances in Archaeological Practice* 1 (1): 38–46. `https://doi.org/10.7183/2326-3768.1.1.37`.

Kline, Michelle Ann. 2015. "How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals." *The Behavioral and Brain Sciences* 38: e31. `https://doi.org/10.1017/S0140525X14000090`.

Laland, Kevin N. 2017. "The Origins of Language in Teaching." *Psychonomic Bulletin & Review* 24 (1): 225–31. `https://doi.org/10.3758/s13423-016-1077-7`.

Levinson, Stephen C., and Judith Holler. 2014. "The Origin of Human Multi-Modal Communication." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1651): 20130302. `https://doi.org/10.1098/rstb.2013.0302`.

Lombao, D., M. Guardiola, and M. Mosquera. 2017. "Teaching to Make Stone Tools: New Experimental Evidence Supporting a Technological Hypothesis for the Origins of Language." *Scientific Reports* 7 (1): 1–14. https://doi.org/10.1038/s41598-017-14322-y.

Marwick, Ben. 2017. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." *Journal of Archaeological Method and Theory* 24 (2): 424–50. https://doi.org/10.1007/s10816-015-9272-9.

Marzke, Mary W., N. Toth, K. Schick, S. Reece, B. Steinberg, K. Hunt, R. L. Linscheid, and K.-N. An. 1998. "EMG Study of Hand Muscle Recruitment During Hard Hammer Percussion Manufacture of Oldowan Tools." *American Journal of Physical Anthropology* 105 (3): 315–32. https://doi.org/10.1002/(SICI)1096-8644(199803)105:3%3C315::AID-AJPA3%3E3.0.CO;2-Q.

Miu, Elena, Ned Gulley, Kevin N. Laland, and Luke Rendell. 2020. "Flexible Learning, Rather Than Inveterate Innovation or Copying, Drives Cumulative Knowledge Gain." *Science Advances* 6 (23): eaaz0286. https://doi.org/10.1126/sciadv.aaz0286.

Molleman, Lucas, Ralf H. J. M. Kurvers, and Wouter van den Bos. 2019. "Unleashing the BEAST: A Brief Measure of Human Social Information Use." *Evolution and Human Behavior* 40 (5): 492–99. https://doi.org/10.1016/j.evolhumbehav.2019.06.005.

Montagu, Ashley. 1976. "Toolmaking, Hunting, and the Origin of Language." *Annals of the New York Academy of Sciences* 280 (1): 266–74. https://doi.org/10.1111/j.1749-6632.1976.tb25493.x.

Morgan, T. J. H., N. T. Uomini, L. E. Rendell, L. Chouinard-Thuly, S. E. Street, H. M. Lewis, C. P. Cross, et al. 2015. "Experimental Evidence for the Co-Evolution of Hominin Tool-Making Teaching and Language." *Nature Communications* 6 (1): 6029. https://doi.org/10.1038/ncomms7029.

Nonaka, Tetsushi, Blandine Bril, and Robert Rein. 2010. "How Do Stone Knappers Predict and Control the Outcome of Flaking? Implications for Understanding Early Stone Tool Technology." *Journal of Human Evolution* 59 (2): 155–67. https://doi.org/10.1016/j.jhevol.2010.04.006.

Oakley, Kenneth P. 1949. *Man the Toolmaker*. London: Trustees of the British Museum.

Ohnuma, Katsuhiko, Kenichi Aoki, and And Takeru Akazawa. 1997. "Transmission of Tool-Making Through Verbal and Non-Verbal Commu-Nication: Preliminary Experiments in Levallois Flake Production." *Anthropological Science* 105 (3): 159–68. https://doi.org/10.1537/ase.105.159.

Pargeter, Justin, Nada Khreisheh, John J. Shea, and Dietrich Stout. 2020. "Knowledge Vs. Know-How? Dissecting the Foundations of Stone Knapping Skill." *Journal of Human Evolution* 145 (August): 102807. https://doi.org/10.1016/j.jhevol.2020.102807.

Pargeter, Justin, Nada Khreisheh, and Dietrich Stout. 2019. "Understanding Stone Tool-Making Skill Acquisition: Experimental Methods and Evolutionary Implications." *Journal of Human Evolution* 133 (August): 146–66. https://doi.org/10.1016/j.jhevol.2019.05.010.

Pelegrin, Jacques. 1990. "Prehistoric Lithic Technology : Some Aspects of Research." *Archaeological Review from Cambridge* 9 (1): 116–25. /paper/Prehistoric-Lithic-Technology-.

Poldrack, Russell A. 2011. "Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding." *Neuron* 72 (5): 692–97. https://doi.org/10.1016/j.neuron.2011.11.001.

Prasciunas, Mary M. 2007. "Bifacial Cores and Flake Production Efficiency: An Experimental Test of Technological Assumptions." *American Antiquity* 72 (2): 334–48. https://doi.org/10.2307/40035817.

Putt, Shelby S., Sobanawartiny Wijeakumar, Robert G. Franciscus, and John P. Spencer. 2017. "The Functional Brain Networks That Underlie Early Stone Age Tool Manufacture." *Nature Human Behaviour* 1 (6): 1–8. https://doi.org/10.1038/s41562-017-0102.

Putt, Shelby S., Sobanawartiny Wijeakumar, and John P. Spencer. 2019. "Prefrontal Cortex Activation Supports the Emergence of Early Stone Age Toolmaking Skill." *NeuroImage* 199 (October): 57–69. https://doi.org/10.1016/j.neuroimage.2019.05.056.

Putt, Shelby S., Alexander D. Woods, and Robert G. Franciscus. 2014. "The Role of Verbal Interaction During Experimental Bifacial Stone Tool Manufacture." *Lithic Technology* 39 (2): 96–112. https://doi.org/10.1179/0197726114Z.00000000036.

Rein, Robert, Tetsushi Nonaka, and Blandine Bril. 2014. "Movement Pattern Variability in Stone Knapping: Implications for the Development of Percussive Traditions." *PLOS ONE* 9 (11): e113567. https://doi.org/10.1371/journal.pone.0113567.

Reti, Jay S. 2016. "Quantifying Oldowan Stone Tool Production at Olduvai Gorge, Tanzania." *PLOS ONE* 11 (1): e0147352. https://doi.org/10.1371/journal.pone.0147352.

Roux, Valentine, Blandine Bril, and Gilles Dietrich. 1995. "Skills and Learning Difficulties Involved in Stone Knapping: The Case of Stone-Bead Knapping in Khambhat, India." *World Archaeology* 27 (1): 63–87. https://doi.org/10.1080/00438243.1995.9980293.

Rueden, Curtis T., Johannes Schindelin, Mark C. Hiner, Barry E. DeZonia, Alison E. Walter, Ellen T. Arena, and Kevin W. Eliceiri. 2017. "ImageJ2: ImageJ for the Next Generation of Scientific Image Data." *BMC Bioinformatics* 18 (1): 529. https://doi.org/10.1186/s12859-017-1934-z.

Sasaki, Hideo, Fumiyoshi Kasagi, Michiko Yamada, and Shoichiro Fujita. 2007. "Grip Strength Predicts Cause-Specific Mortality in Middle-Aged and Elderly Persons." *The American Journal of Medicine* 120 (4): 337–42. https://doi.org/10.1016/j.amjmed.2006.04.018.

Schillinger, Kerstin, Alex Mesoudi, and Stephen J. Lycett. 2014. "Copying Error and the Cultural Evolution of "Additive" Vs. "Reductive" Material Traditions: An Experimental Assessment." *American Antiquity* 79 (1): 128–43. https://doi.org/10.7183/0002-7316.79.1.128.

Shallice, Timothy, Donald Eric Broadbent, and Lawrence Weiskrantz. 1982. "Specific Impairments of Planning." *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 298 (1089): 199–209. https://doi.org/10.1098/rstb.1982.0082.

Shea, John J. 2015. "Making and Using Stone Tools: Advice for Learners and Teachers and Insights for Archaeologists." *Lithic Technology* 40 (3): 231–48. https://doi.org/10.1179/2051618515Y.0000000011.

———. 2016. *Stone Tools in Human Evolution: Behavioral Differences Among Technological Primates.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316389355.

Sherwood, Chet C., and Aida Gómez-Robles. 2017. "Brain Plasticity and Human Evolution." *Annual Review of Anthropology* 46 (1): 399–419. https://doi.org/10.1146/annurev-anthro-102215-100009.

Stout, Dietrich. 2002. "Skill and Cognition in Stone Tool Production: An Ethnographic Case Study from Irian Jaya." *Current Anthropology* 43 (5): 693–722. https://doi.org/10.1086/342638.

———. 2010. "Possible Relations Between Language and Technology in Human Evolution." In, edited by April Nowell and Iain Davidson, 159184. Boulder, CO: University Press of Colorado.

———. 2013. "Neuroscience of Technology." In, edited by Peter J. Richerson and Morten H. Christiansen, 157–73. Cambridge, MA: The MIT Press.

Stout, Dietrich, Jan Apel, Julia Commander, and Mark Roberts. 2014. "Late Acheulean Technology and Cognition at Boxgrove, UK." *Journal of Archaeological Science* 41 (January): 576–90. https://doi.org/10.1016/j.jas.2013.10.001.

Stout, Dietrich, and Thierry Chaminade. 2007. "The Evolutionary Neuroscience of Tool Making." *Neuropsychologia* 45 (5): 1091–1100. https://doi.org/10.1016/j.neuropsychologia.2006.09.014.

———. 2012. "Stone Tools, Language and the Brain in Human Evolution." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1585): 75–87. https://doi.org/10.1098/rstb.2011.0099.

Stout, Dietrich, Erin Hecht, Nada Khreisheh, Bruce Bradley, and Thierry Chaminade. 2015. "Cognitive Demands of Lower Paleolithic Toolmaking." *PLOS ONE* 10 (4): e0121804. https://doi.org/10.1371/journal.pone.0121804.

Stout, Dietrich, and Nada Khreisheh. 2015. "Skill Learning and Human Brain Evolution: An Experimental Approach." *Cambridge Archaeological Journal* 25 (4): 867–75. https://doi.org/10.1017/S0959774315000359.

Stout, Dietrich, Richard Passingham, Christopher Frith, Jan Apel, and Thierry Chaminade. 2011. "Technology, expertise and social cognition in human evolution." *The European Journal of Neuroscience* 33 (7): 1328–38. https://doi.org/10.1111/j.1460-9568.2011.07619.x.

Stout, Dietrich, Jay Quade, Sileshi Semaw, Michael J. Rogers, and Naomi E. Levin. 2005. "Raw Material Selectivity of the Earliest Stone Toolmakers at Gona, Afar, Ethiopia." *Journal of Human Evolution* 48 (4): 365–80. https://doi.org/10.1016/j.jhevol.2004.10.006.

Stout, Dietrich, Michael J. Rogers, Adrian V. Jaeggi, and Sileshi Semaw. 2019. "Archaeology and the Origins of Human Cumulative Culture: A Case Study from the Earliest Oldowan at Gona, Ethiopia." *Current Anthropology* 60 (3): 309–40. https://doi.org/10.1086/703173.

Stout, Dietrich, and Sileshi Semaw. 2006. "Knapping Skill of the Earliest Stone Toolmakers: Insights from the Study of Modern Human Novices." In *The Oldowan: Case Studies into the Earliest Stone Age*, edited by Nicholas Toth and Kathy Schick, 307–20. Gosport, IN: Stone Age Institute Press.

Tehrani, Jamshid J., and Felix Riede. 2008. "Towards an Archaeology of Pedagogy: Learning, Teaching and the Generation of Material Culture Traditions." *World Archaeology* 40 (3): 316–31. https://doi.org/10.1080/00438240802261267.

Toth, Nicholas, and Kathy Schick. 1993. "Early Stone Industries and Inferences Regarding Language and Cognition." In, edited by Kathleen R. Gibson and Tim Ingold, 346362. Cambridge: Cambridge University Press.

Unsworth, Nash, and Randall W. Engle. 2005. "Individual Differences in Working Memory Capacity and Learning: Evidence from the Serial Reaction Time Task." *Memory & Cognition* 33 (2): 213–20. https://doi.org/10.3758/BF03195310.

Washburn, Sherwood L. 1960. "Tools and Human Evolution." *Scientific American* 203 (3): 62–75. https://doi.org/10.1038/scientificamerican0960-62.

Whiten, Andrew. 2015. "Experimental Studies Illuminate the Cultural Transmission of Percussive Technologies in Homo and Pan." *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1682): 20140359. https://doi.org/10.1098/rstb.2014.0359.

Whittaker, John C. 1994. *Flintknapping Making and Understanding Stone Tools By John C. Whittaker*. Austin, TX: University of Texas Press.

Wilkins, Jayne. 2018. "The Point Is the Point: Emulative Social Learning and Weapon Manufacture in the Middle Stone Age of South Africa." In, edited by Michael J. O'Brien, Briggs Buchanan, and Metin I. Eren, 153–74. Cambridge, MA: The MIT Press.

Williams-Hatala, Erin Marie, Kevin G. Hatala, McKenzie Gordon, Alastair Key, Margaret Kasper, and Tracy L. Kivell. 2018. "The Manual Pressures of Stone Tool Behaviors and Their Implications for the Evolution of the Human Hand." *Journal of Human Evolution* 119 (June): 14–26. https://doi.org/10.1016/j.jhevol.2018.02.008.

Wind, Anne E., Tim Takken, Paul J. M. Helders, and Raoul H. H. Engelbert. 2010. "Is Grip Strength a Predictor for Total Muscle Strength in Healthy Children, Adolescents, and Young Adults?" *European Journal of Pediatrics* 169 (3): 281–87. https://doi.org/10.1007/s00431-009-1010-4.

Wynn, Thomas. 1979. "The Intelligence of Later Acheulean Hominids." *Man* 14 (3): 371–91. https://doi.org/10.2307/2801865.

———. 2017. "Evolutionary Cognitive Archaeology." In, edited by Thomas Wynn and Frederick Coolidge, 120. Oxford: Oxford University Press.

Wynn, Thomas, and Frederick L. Coolidge. 2004. “The expert Neandertal mind.” *Journal of Human Evolution* 46 (4): 467–87. https://doi.org/10.1016/j.jhevol.2004.01.005.

———. 2016. “Archeological Insights into Hominin Cognitive Evolution.” *Evolutionary Anthropology: Issues, News, and Reviews* 25 (4): 200–213. https://doi.org/10.1002/evan.21496.

Yarkoni, Tal. 2020. “The Generalizability Crisis.” *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685.