Testing the motor and cognitive foundations of Paleolithic social transmission

Justin Pargeter[*] Megan Beney Kilgore[†] Cheng Liu[‡] Dietrich Stout[§]

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 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

Contents

[*] Department of Anthropology, New York University, New York, NY, USA; Palaeo-Research Institute, University of Johannesburg, Auckland Park, South Africa; justin.pargeter@nyu.edu

[†] Department of Anthropology, Emory University, Atlanta, GA, USA; megan.elizabeth.beney@emory.edu

[‡] Department of Anthropology, Emory University, Atlanta, GA, USA; raylc1996@outlook.com

[§] Department of Anthropology, Emory University, Atlanta, GA, USA; dwstout@emory.edu

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 (Gowlett, 1984; Isaac, 1976; Wynn, 1979; Wynn & Coolidge, 2004). More recently, increasing attention has been directed to the processes and demands of stone tool making skill acquisition (Cataldo et al., 2018; Duke & Pargeter, 2015; Geribàs et al., 2010; Hecht, Gutman, Khreisheh, et al., 2015; Lombao et al., 2017; Morgan et al., 2015; Nonaka et al., 2010; Pargeter et al., 2020; Pargeter et al., 2019; Putt et al., 2017, 2019; Putt et al., 2014; Roux et al., 1995; Stout et al., 2005; Stout et al., 2011; Stout, 2002; Stout & Khreisheh, 2015). 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 & Khreisheh, 2015) and by interest in the relevance of different social learning mechanisms such as imitation (Rein et al., 2014; Stout et al., 2019), emulation (Tehrani & Riede, 2008; Wilkins, 2018), and language (Cataldo et al., 2018; Lombao et al., 2017; Morgan et al., 2015; Ohnuma et al., 1997; Putt et al., 2017; Putt et al., 2014) 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 et al., 2019; Stout & 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 et al., 2019; Putt et al., 2019; Stout & Khreisheh, 2015), 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 & 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 & 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 (Barrett, 2020; Sherwood & Gómez-Robles, 2017). 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 et al., 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 & Chaminade, 2007) and coordination (Roux et al., 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 & 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 et al., 1982)) and cognitive set shifting (Wisconsin Card Sort test (Grant & 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 et al., 2010; Pargeter et al., 2020) and studies of muscle recruitment (Marzke et al., 1998) and manual pressure (Key & Dunmore, 2018; Williams-Hatala et al., 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 et al., 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, p. [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 et al., 2017; Putt et al., 2014) and simple flake production (Cataldo et al., 2018; Lombao et al., 2017; Morgan et al., 2015). This reflects recent interest in the hypothesis that language might be an adaptation for teaching (e.g., Laland, 2017; Stout & Chaminade, 2012). Teaching and learning demands of Paleolithic tool making would thus provide evidence of selective contexts favoring language evolution (Montagu, 1976; Morgan et al., 2015; Stout, 2010).

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 & Hewlett, 2017). The WEIRD (Western, educated, industrialized, rich, democratic (Henrich et al., 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 et al., 2018). While this provides experimental control, it does not allow the instructor to adjust their multi-modal (Levinson & 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 (Lombao et al., 2017; Morgan et al., 2015)) might misrepresent the relative efficacy of different teaching strategies under more realistic conditions (Stout & Khreisheh, 2015; Whiten, 2015). Even the longest training times to date (Pargeter et al., 2019; Stout & Khreisheh, 2015) 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 (Hecht, Gutman, Khreisheh, et al., 2015; Pargeter et al., 2019; Stout et al., 2011) 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 et al., 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 et al., 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 & Khreisheh, 2015) and rapid early progress (Putt et al., 2019; Stout & Khreisheh, 2015; Stout & 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

Such undertakings – based on raw material which is never standard, and with gestures of percussion that are never perfectly delivered – cannot be reduced to an elementary repetition of gestures... the realization of elaborate knapping activities necessitates a critical monitoring of the situation and of the decisions adopted all through the process. (Pelegrin, 1990: 117)

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 (Cataldo et al., 2018; Morgan et al., 2015; Nonaka et al., 2010), limestone (Stout & Semaw, 2006), porcelain (Khreisheh et al., 2013), or heat-treated chert (Putt et al., 2017, 2019; Putt et al., 2014)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 (Duke & Pargeter, 2015; Pargeter et al., 2019; Stout et al., 2011). However, raw materials vary across individual clasts within as well as between types. This has led to interest in standardizing experimental core morphology (Nonaka et al., 2010) and composition, even if this means using artificial materials such as porcelain (Khreisheh et al., 2013), brick (Geribàs et al., 2010; Lombao et al., 2017), or foam blocks (Schillinger et al., 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 et al., 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 et al., 2010) as well as control and support of the hammerstone (Williams-Hatala et al., 2018) and core (Faisal et al., 2010; Key & 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 (Pargeter et al., 2020; Stout, 2002) that management of this trade-off is critical to the accurate application of appropriate force seen in skilled knapping

(Nonaka et al., 2010; Roux et al., 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 & 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 et al., 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 & Wynn, 2005) and/or the intentional learning of new skills (Stout & Khreisheh, 2015; Unsworth & Engle, 2005). 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 et al., 2019). Such variation is a key topic for understanding social learning and cultural evolutionary processes (Heyes, 2018; Kendal et al., 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 (**Fig. 1**) 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.

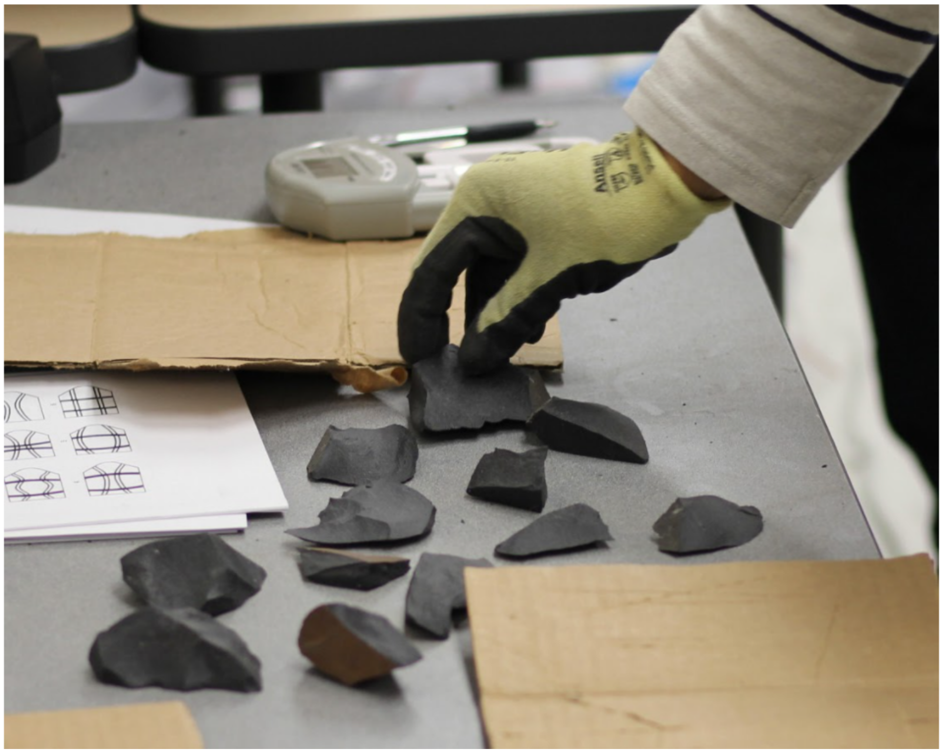

Figure 1: Subjects examining demonstration flakes prior to the experiment. The demonstration lakes were made from the same basalt as used in the experiment with the same knapping technique.

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 × 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 tool-making 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.

2.5 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 (Key & 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 et al., 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.

2.6 Statistical Analyses

To evaluate the association between psychometric, motor-skill, and training measures and technological outcomes, we adopted an information-theoretic approach (Burnham & 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 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. We were particularly interested in: 1) group level effects of experimental condition, 2) individual differences in aptitude and learning, and 3) potential interactions between learning conditions and individual differences. To address these questions, we employed data reduction (Principal Component Analysis) to derive two summary metrics of flaking performance, compared these factors across the two experimental conditions, and built multivariate models examining the relations between our various psychometric measures, subject's motor skill scores, and our two lithic performance factors.

3.1 Principal Component analyses

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

3.1.1 Detached Piece size and shape

To better understand the relationship between DP 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 DP 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 data dimensionality, we conducted a second principal component analysis examining the study's six lithic performance measures (count of large pieces [>40mm and 5g], mass of large pieces relative to total detached mass, core delta mass, and the three flake shape factors). All of these measures were summarized for each core and unique factor scores were calculated from these core-specific measures. Bartlett's Test of Sphericity was significant (χ^2 (6) =3185, 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 56% of the variance for the entire set of variables (Table). Factor 1 (hereafter "Quantity") explains 28.7% of the variance and tracks flaking quantity due to high positive loadings on large DP count and mass ratio and on core delta mass. Performance factor 2 (hereafter "Quality") covers 27% of the sample variance and measures flaking quality as reflected in high positive loadings on Shape Factors 1 (size) and 3 (thin, "flake-like" shape) and a negative loading on Shape Factor 2 ("splinter-like" thickness and elongation). High scores on Quality thus reflect production of larger, relatively thinner, and more typically flake-shaped vs. splinter-shaped DPs.

These two factors address flaking performance at the level of individual cores, however we were also interested in the overall productivity/rate of work of each participant over the entire two hours. For example, looking at a knappers average Quality and Quantity factor scores would not differentiate between a participant who spent the entire time exhaustively reducing one core vs. another participant who did the same to all nine of their allotted cores in the same time. To capture this aspect of variation we calculated a simple Total Productivity metric as the sum of all mass a participant removed from cores during the experiment.

3.2 Relationships between Performance Measures

This approach also allowed us to compare the relationship between Total Productivity, Quantity, and Quality across our two experimental groups (Figures). As might be expected, we found that per-core Quantity and Total Productivity are positively correlated in both groups (Fig. Xa), although this relationship is twice as strong in the trained (F[1, 9] = 33, p < 0.01, Adj. R^2 = 0.8) compared to untrained (F[1, 8] = 8, p = 0.02, Adj. R^2 = 0.4) group. Interestingly, we also found evidence of a negative correlation between Total Productivity and Quality in the untrained group (F[1, 8] = 28, p = <0.01, Adj. R^2 = 0.7), but no relation in the Trained group (Fig. Xb). A qualitatively similar trend with respect to Quality vs. Quantity (Fig. Xc) did not achieve significance (F[1, 17] = 0.6, p = 0.2).

Thus, it appears that Trained participants achieved higher Total Productivity by increasing average flaking Quantity across cores and without sacrificing Quality, whereas Untrained participants found other ways to vary Total Productivity (e.g., number of cores knapped rather than Quantity per core, see variance Table and Figure) and generally increased productivity at the expense of Quality. Experimental artifacts illustrating these trade-offs are presented in Figure.

3.3 Do trained, untrained, and expert knappers perform differently?

Here we compare our flaking outcomes (DP size/shape and flaking performance factors) between the trained and untrained groups. Our expert demonstrator/instructor is included as a performance benchmark.

Table summarizes the results of ANOVA tests group level difference in central tendency on various performance measures. We found no significant differences between the trained and untrained groups on our flaking Quantity and Quality factors. In contrast, three-way flake size and shape comparisons between our expert knapper and the two novice groups showed that 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 left significantly smaller finished 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). We also observed a significant difference in shape factor 2 (splinters) driven by the expert's lower values, but with a very low (<0.01) effect size. These results show that mean core reduction intensity and large flake production rates distinguish expert and novice performance whereas novices in experimental groups produced pieces of similar mean size and shape as those of the expert trainer.

While we did not find significant differences in central tendency between our two experimental groups, results (Figure and Figure) did indicate lower variance in the trained group. To test whether training reduced variability in performance outcomes between subjects, we compared variance metrics between the trained and untrained individuals using the F-test on either core-averaged or flake specific variances. Table and Figure present the results from these comparisons showing significant variance differences predominantly in flaking Quality, number of large DPs, core delta mass, and total amount of flaked mass). In most instances, variance in the untrained group exceeds that of trained individuals by 1.5 to 4.7 times. The most salient effect of instruction was thus not to shift mean performance but to reduce variability by eliminating the skew (generally toward poorer outcomes) seen in the untrained group (Figure), rather than to shift the mean.

3.4 Does performance change over time?

In addition to comparing overall performance during the two hour experiment, we also wanted to determine if groups or individuals differed in learning (i.e., performance change) over the period. For these analyses, we calculated the learning stage as the ordinal number of each core out of the total number knapped by each subject (i.e., core 2 of 4 or 4 of 8 both equal 50% complete). 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 performance factors (Quality and Quantity). 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 performance measures either as grouped data or between individuals (Figures). This result demonstrated that flaking outcomes did not change dramatically across the study interval. This lack of significant learning effects is confirmed by an inspection of individual learning curves (Figure). The one significant main training effect related to core starting mass (with a strong effect size = 0.25). On average, core starting masses start low and increase, showing that participants selected smaller nodules first. As the smaller nodules in their allotment were depleted, participants were left to knap larger, less preferred nodules. This preference for smaller cores is somewhat less pronounced in the untrained group, as indicated by a small main effect of learning condition and generally higher starting nodule masses for the untrained group (Figure).

3.5 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 perceptual-motor and cognitive variation predict success in stone flaking across different training conditions. To address this goal, we built three multivariate models examining the relations between training conditions, individual difference measures, and our three lithic performance measures (overall productivity and average per-core Quantity and Quality). 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 considered all possible interactions between five individual difference measures, core size, training condition, and the three performance measures, with each subject providing one data point. Each model's continuous predictors (highest n-back level, Raven's Progressive Matrix score, BEAST score, starting nodule mass, Fitts score, and grip strength) were centered such that zero represents the sample average, and units are standard deviations. 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, R^2 = 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.

The full models were 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). All three models follow the same complete model statement as follows:

Flaking performance variable $\sim$ Training condition + Highest n-back level + Raven's Progressive Matrix score + BEAST score + Fitt's score + Grip strength

For our two per-core performance factors (Quantity and Quality) it is also relevant to consider how individual core features may have affected performance. We found no evidence of individual or group level practice effects over the two hours, so we did not include core order in the models. We did, however, find that subjects selected progressively larger nodules throughout the experiment. It is thus important 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, R^2 = 0.54) we included nodule mass as a covariate to control for any variance in flaking performance that may be driven by nodule differences.

3.5.1 Model 1: Individual differences and quantity flaking

Our first model examined variance in overall flaking productivity measured by each subject's combined flaked mass (nodule starting mass-core final mass). This provides a basic measure of variation in individuals' success detaching pieces and reducing cores from a standardized (see Methods) raw material supply. From the same candidate pool size of 55893 possible multivariate models, the best performing model returned an AICc value of -18 (Average AIC = -13). This model comprised the following statement with two main and three interaction effects:

Total flaked mass $\sim$ Training condition + Grip strength + RPM×Highest n-back level + Fitts score×BEAST score + Grip strength×RPM

This model explains a statistically significant and substantial proportion of variance in flaking productivity (R^2 = 0.84, F (6, 14) = 12.7, p < 0.01, adj. R^2 = 0.77). A model residuals normality test shows no significant differences with the normal distribution (p = 0.72) indicating that this relationship is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (BP = 2, df = 6, p = 0.8).

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. We found significant (p< 0.05) and substantial (Standardized Estimate >= 0.50, i.e. a 50% change in variable) main effects of Grip Strength, Visuospatial nBack, and BEAST. The main effect of Grip Strength (Figure), irrespective of learning condition, indicates the basic importance of strength in generating higher production rates among naive knappers at least when efficiency and quality are not considered.

Effects of visuospatial working memory capacity and social information use are more complicated, as indicated by strong interactions with learning condition (Figures). In each case, higher scores were associated with better performance in the uninstructed group but worse performance in the instructed group. Positive effects in the uninstructed group were as expected, given the hypothesized importance of spatial cognition (Coolidge and Wynn 2005) and social learning

(Morgan et al. 2015) in the acquisition of knapping skills. Negative effects in the trained group are unexpected but presumably reflect differences in learning strategies adopted under instruction.

3.5.2 Model 2: Individual differences and quality flaking

The second full model examined the variance in average flaking Quantity per core. It thus complements our first model assessing overall productivity by testing for differences in reduction intensity at the level of individual cores. From a candidate pool of 55893 possible multivariate models, the best performing model returned an AICc value of 32 (Average AIC = 44). This model comprised the following statement with three main and four interaction effects:

Quantity $\sim$ Highest n-back level + BEAST score + Fitt's score + Grip strength + Training condition$\times$Highest n-back level + Training condition$\times$BEAST score + Training condition$\times$Grip strength + Nodule mass (as control)

This model explains a statistically significant and substantial proportion of variance in quantity flaking (R^2 = 0.7, F (8, 12) = 3.6, p = 0.02, adj. R^2 = 0.5). A model residuals normality test shows no significant differences with the normal distribution (p = 0.38) 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 = 4.4, df = 8, p = 0.8).

The Quantity model roughly paralleled results for Total Production, yielding substantial and significant interactions between training condition, n-back level, BEAST scores, and grip strength. As with Total Production, higher visuospatial n-back levels and BEAST scores were associated with lower Quantity scores in the trained group but higher or unchanged Quantity in the untrained group (Figure).

Unlike Total Productivity, the effect of Grip Strength on per-core Quantity was mediated by an interaction with learning condition (Figure). Thus, high Grip Strength enabled individuals in both groups to produce more total debitage, but only Instructed individuals translated Grip Strength into more intense reduction of individual cores, including not only delta mass, but also number and proportion of larger pieces.

3.5.3 Model 3: Individual differences and quality flaking

Our third model examining variance in Quality follows the same complete model statement we used for Quantity. From the same candidate pool size of 55893 possible multivariate models, the best performing model returned an AICc value of 32 (Average AIC = 39). This model comprised the following statement with three main and four interaction effects:

Quality flaking $\sim$ 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 Quality (R^2 = 0.75, F (8, 12) = 4.6, p < 0.01, adj. R^2 = 0.6) in the absence of any main training effects. A model residuals normality test shows no significant differences with the normal distribution (p = 0.41) indicating that this relationship is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (BP = 7, df = 8, p = 0.5).

The Quality model did produce two statistically (p< 0.05) significant interaction effects (RPM * BEAST & Fitts * n-back). However, these interactions had relatively small effects on Quality (<0.5) and we believe that interpreting these results from our small, exploratory study would be inappropriate. Figure shows the uneven distribution of data points for these interactions, which suggests vulnerability to leveraging effects of a small number of extreme value combinations.

Without a larger sample, it is not possible to determine if these are anomalous outliers or simply represent a poorly sampled part of the broader population.

3.6 Behavioral observations

We designed this exploratory study primarily to trial experimental design elements such as training time, conditions, and raw materials and to collect preliminary data on the effect of individual differences and training on knapping outcomes. We thus focused on collecting quantitative psychometric and lithic data. However, we also considered that quantifying participant knapping behaviors as well as products could be important for future studies. To support methods development in this regard, we made ad hoc notes on observed behaviors during the experiments and video-recorded all experiments to enable later, more systematic analyses yet to be completed. However, even casual behavior observation was sufficient to reveal an unexpected effect. Whereas all trained participants copied the general posture and technique of the expert (free hand knapping seated in a chair) fully half (6) of the uninstructed participants experimented with or even knapped all of their cores using the floor as a support (Figure). Three of these participants were in the same session, which is also the only group composed of just three individuals. In this group, knapping on the ground appears to have been transmitted from one participant to the other two based on appearance order and the point of gaze of participants.

4 Discussion

The most salient finding of our exploratory study is that the presence/absence of teaching clearly impacted knapping performance but did so in nuanced and individually variable ways that have not been explored in previous studies. In fact, we did not observe any significant differences in mean performance between our experimental conditions. Some non-significant tendencies toward enhanced Instructed group performance suggest that a larger participant sample might detect significant effects, but also that the size of any such effects would likely remain small. This could reasonably lead to the conclusion that teaching does not substantially facilitate early stage knapping skill acquisition (Ohnuma et al., 1997; cf. Putt et al., 2014). Looking closer, however, we found a number of important teaching effects.

4.1 Variance Reduction

In our experiment, the strongest effects of teaching were to reduce variance (Figure, Table) rather than shift mean values. In particular, teaching acted as a “safety net” that homogenized performance by reducing the frequency of extremely poor outcomes (i.e., learning failures). This finding provides additional support for the hypothesis that teaching would have increased the reliability of Oldowan skill reproduction (Morgan et al., 2015) while simultaneously corroborating the view that basic flaking competence can be achieved in its absence (Tennie et al., 2017). Our results thus do not imply that teaching was required or even present during Oldowan times (but see Gärdenfors & Högberg (2017)), but rather serve to reinforce the plausibility of co-evolutionary scenarios positing the cost/reliability of technological skill acquisition as a selection pressure favoring the evolution of teaching and language (Morgan et al., 2015; Stout, 2010; Stout & Hecht, 2017).

4.2 Knapping Behaviors

The current study complements the transmission chain design of Morgan et al. (2015) by finding similar effects in more naturalistic learning contexts. Our design further allowed us to examine individual variation to better understand how teaching produces its effects. Whereas transmission chains are optimized to investigate iterative learning effects (but see Caldwell et al., 2020), they necessarily involve a different instructor/model for each participant. We sacrificed this iterative component in order to consider how the presence/absence of teaching affected the behavior of individuals under otherwise standardized learning conditions.

We found that a key impact of teaching was to alter basic flake production strategies, as reflected in the relationship between Total Productivity and detached piece Quality (Fig.Xb). Whereas Untrained participants achieved greater Productivity at the expense of Quality and vice versa, these dimensions were unrelated across Trained participants. Thus, even though Untrained participants achieved the highest values on each metric, only trained individuals managed to maximize both simultaneously. Indeed, a core function of teaching is to reduce the search space that learners must explore and increase the likelihood of discovering globally as opposed to locally optimum solutions (cf. Hinton & Nowlan, 1996; Stout, 2013). In our study, Untrained individuals explored a greater range of basic behavioral variations not seen in the Trained group, including knapping on the floor, concentrating on working just a few cores (2-4, Figure) over the practice period, and showing less constrained nodule size preferences (Figure). It is notable that this variation occurred even in the presence of an observable expert example, suggesting it may be interesting for future experiments to address the impact of social context, expectations, and relationships on observational learning strategies (Kendal et al., 2018).

We also found a strong positive effect of Grip Strength on Total Productivity independent of learning condition (Figure). While it is tempting to interpret this with respect to the demands for hand strength specifically, it is important to remember that grip strength is strongly correlated with total muscle strength (Wind et al., 2010) and overall fitness (Sasaki et al., 2007). Thus, it is best taken to indicate some importance of fitness generally in increasing the rate and intensity of core reduction by naïve knappers, potentially affecting rate of work and the kinetic energy of the swing as well as the handling of core and hammerstone. It thus provides further support for hypotheses positing stone tool making as a selection pressure on the functional anatomy of hand, arm, and shoulder (e.g., Williams-Hatala et al., 2018), but initially appears orthogonal to variations in learning condition and knapping behaviors in our study. However, we also found that the effect of Grip Strength on per-core knapping Quantity is dependent on teaching (Figure). The absence of this effect in the uninstructed group reflects the weaker association between Total Productivity and per-core Quality across these participants (Fig.Xa) and shows Grip Strength increased uninstructed Productivity specifically by allowing them to knap more cores rather than to reduce individual cores more heavily. In keeping with this, uninstructed Grip Strength is positively correlated with Total Cores knapped (r2 = 0.54, p = 0.01). Conversely, strength allowed instructed participants to increase their average Quantity per core without affecting the total number of cores knapped (r2 = 0.18, p = 0.165). Thus, strength appears to have achieved its effects on core reduction rate and intensity in different ways, depending on teaching. This difference is likely related to the homogenizing effect of teaching on knapping rate (all instructed participants knapped 6 or more cores) and methods. Subjectively, knapping behaviors of uninstructed participants often appeared more physically demanding (e.g., greater number of non-productive blows, rapid and unregulated battering) which would imply different demands on both strength and aerobic fitness (Mateos et al., 2019; Williams-Hatala et al., 2021). However, this remains to be systematically investigated.

In this respect, it is also important to note that we do not know how well the knapping objectives and strategies communicated by the expert in our experiment correspond to actual Oldowan goals and behaviors. The instructor has successfully replicated assemblage-level patterning at Gona (Stout et al., 2019) but Oldowan behavior is variable across space and time (e.g., Braun et al., 2019) and alternative knapping methods might maximize different values (productivity, quality, effort), especially in novices (Putt, 2015) (Putt 2015). Nevertheless, the effect of instruction to constrain behavioral exploration and homogenize outcomes is clear. We expect that this effect would generalize to the teaching of alternative knapping goals and behaviors, although this remains to be tested.

4.3 Learning Strategies

One major goal of this experiment was to test the viability of a moderate, two-hour, learning period for studies of skill acquisition. Unfortunately, we found that this duration was insufficient to capture learning effects for Oldowan-like flake production. The lack of performance change over the period (Figure) cannot be attributed to a ceiling effect (i.e., rapid task mastery at the outset of the practice period) as participants remained well below expert levels and continued to display the high within-individual variability typical of naïve/novice knapping (Eren et al., 2011; Pargeter et al., 2019). This negative result was unexpected but is broadly consistent with evidence that Early Stone Age flaking, while conceptually simple, requires substantial practice for perceptual-motor skill development (Nonaka et al., 2010; Pargeter et al., 2020; Stout & Khreisheh, 2015). Future investigations of learning variation across individuals and/or experimental conditions may thus need to incorporate longer practice periods to capture skill acquisition processes. In theory, much shorter knapping trials might be used to assess the variation in initial performance across individuals and under different conditions that is captured in our study. However, the presence of substantial core-to-core variation within individuals cautions against overly brief experiments that might not provide a representative sample. Greater durations also allow for the expression of different learning strategies over time, even in the absence of directional performance change.

At a basic level, learners of any new task must balance investment in task exploration vs. exploitation of knowledge and skills already in hand (Sutton & Barto, 2018). Premature exploitation risks settling for a sub-optimal local solution whereas continued exploration sacrifices more immediate payoffs. Managing this trade-off is especially challenging for complex, real-world tasks like stone knapping, and is thought to depend on the interplay of uncertainty and reward expectation (Wilson et al., 2021). Teaching and social learning generally have the potential to provide low-cost information about task structure and payoffs (Kendal et al., 2018; Rendell et al., 2010), which if adopted, would be expected to affect exploration/exploitation decisions. Such adoption is itself known to be influenced by individual cognitive differences, for example if higher fluid intelligence allows observers to better understand observed tasks (Vostroknutov et al., 2018) or if individuals vary in their tendency to use and value social information (Molleman et al., 2019; Toelch et al., 2014).

In our study, we did not observe any effect of fluid intelligence (RPM) on knapping outcomes but did find strong interactions of learning condition with participant visuospatial working memory and social information use tendency (Figure). As expected, uninstructed individuals with higher scores on these dimensions displayed higher Total Productivity and average per-core flaking Quantity (although the effect on n-Back on Quantity did not achieve significance). We attribute these effects to increased ability to hold relevant morphological/spatial information in mind and a tendency to benefit from observing successful strategies of others, including the expert model. In contrast, instructed individuals with higher scores tended to have lower Productivity and Quantity. We interpret this unexpected effect to an increased tendency to privilege exploratory learning behavior over exploitation. In particular, we suggest that trained participants might knap more slowly and less productively if higher working memory capacity inclined them to experiment more with morphological/spatial variables highlighted by the instructor or if a predisposition to use social information use caused them to invest greater time and effort attending to and trying out observed actions and/or instructions. These suggestions remain to be tested by further work.

Unfortunately, the training period in our current experiment was insufficient to capture learning effects and so we have no evidence of the effects of these individual differences and putative exploration/exploitation tradeoffs to the ultimate achievement of expertise. A similar negative effect of instruction on knapping outcomes during early stage learning was reported by Putt et al. (2014), and has been interpreted to reflect learners experimenting with advanced techniques before they have the perceptual-motor skill to execute them (Stout & Khreisheh, 2015; Whiten, 2015). Such effects might be further explored with more detailed behavioral data, as opposed to purely lithic data, and with longer learning periods.

4.4 Limitations and Prospects

Although our exploratory study produced a number of robust results with respect to the effects of instruction and individual differences on lithic products, it is clearly limited by a small sample size, short training duration, and lack of detailed quantification of observed behaviors. These are limitations that can hopefully be addressed in future studies building on the methods and evidence presented here. For example, it is notable that our study failed to document any reliable effects on knapping Quality. Obviously, this might reflect an actual lack of such effects, but it may also indicate a need for more sensitive measures and/or increased sample size and training duration to identify subtle or delayed effects. One aspect of our attempt to balance pragmatic costs and benefits in our study was to test the efficacy of relatively limited lithic analysis. More detailed ongoing analyses of core morphology and debitage features (e.g., typology, cutting edge length, platform dimensions) may yet reveal a more reliable signal of knapping quality. Results of the Quality model in particular also seem to suffer from the uneven distribution and discrete rather than continuous nature of scores on our RPM and n-Back tests. Concerns about the sampling of variation on these dimensions could be addressed with larger samples or by pre-screening participants to ensure more even representation. Alternative psychometric tests (e.g., full rather than short version of the RPM) might also provide more sensitive and continuous measures.

Another major limitation that our study shares with all other published experiments on knapping skill acquisition is that we do not address variation in social and cultural context or in teaching style. Currently, we have little basis other than personal experience/tradition (Callahan, 1979; Shea, 2015; Whittaker, 1994) and theoretical speculation (Stout, 2013; Whiten, 2015) from which to assess which pedagogical techniques are most effective even in WEIRD contexts. 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) 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.

5 Conclusions

6 Acknowledgments

7 Figures

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. https://doi.org/10.1073/pnas.1912341117

Barrett, H. C. (2020). Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency. *Trends in Cognitive Sciences*, *24*(8), 620–638. https://doi.org/10.1016/j.tics.2020.05.007

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test.

Assessment, *19*(3), 354–369. https://doi.org/10.1177/1073191112446655

Boogert, N. J., Madden, J. R., Morand-Ferron, J., & Thornton, A. (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, A. H., & Hewlett, B. S. (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

Braun, D. R., Aldeias, V., Archer, W., Arrowsmith, J. R., Baraki, N., Campisano, C. J., Deino, A. L., DiMaggio, E. N., Dupont-Nivet, G., Engda, B., Feary, D. A., Garello, D. I., Kerfelew, Z., McPherron, S. P., Patterson, D. B., Reeves, J. S., Thompson, J. C., & Reed, K. E. (2019). Earliest known Oldowan artifacts at >2.58 Ma from Ledi-Geraru, Ethiopia, highlight early technological diversity. *Proceedings of the National Academy of Sciences*, *116*(24), 11712–11717. https://doi.org/10.1073/pnas.1820177116

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/b97636

Caldwell, C. A., Atkinson, M., Blakey, K. H., Dunstone, J., Kean, D., Mackintosh, G., Renner, E., & Wilks, C. E. H. (2020). Experimental assessment of capacities for cumulative culture: Review and evaluation of methods. *WIREs Cognitive Science*, *11*(1), e1516. https://doi.org/10.1002/wcs.1516

Callahan, E. (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, D. M., Migliano, A. B., & Vinicius, L. (2018). Speech, stone tool-making and the evolution of language. *PLOS ONE*, *13*(1), e0191071. https://doi.org/10.1371/journal.pone.0191071

Cattell, R. 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, F. L., & Wynn, T. (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, C. (1859). *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (1st ed.). John Murray.

Darwin, C. (1871). *The descent of man, and selection in relation to sex* (1st ed.). John Murray.

Duke, H., & Pargeter, J. (2015). Weaving simple solutions to complex problems: An experimental study of skill in bipolar cobble-splitting. *Lithic Technology*, *40*(4), 349–365. https://doi.org/10.1179/2051618515Y.0000000016

Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science*, *13*(2), 190–193. https://doi.org/10.1177/1745691617720478

Engles, F. (2003). The part played by labour in the transition from ape to man. In R. C. Scharff & V. Dusek (Eds.), *Philosophy of Technology – The Technological Condition: An Anthology* (pp. 71–77). Blackwell.

Eren, M. I., Bradley, B. A., & Sampson, C. G. (2011). Middle Paleolithic Skill Level and the Individual Knapper: An Experiment. *American Antiquity*, *76*(2), 229–251. https://doi.org/10.7183/0002-7316.76.2.229

Eren, M. I., Lycett, S. J., Patten, R. J., Buchanan, B., Pargeter, J., & O'Brien, M. J. (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), 103–136. https://doi.org/10.1080/19442890.2016.1213972

Eren, M. I., Roos, C. I., Story, B. A., von Cramon-Taubadel, N., & Lycett, S. J. (2014). The role of raw material differences in stone tool shape variation: an experimental assessment. *Journal of Archaeological Science*, *49*, 472–487. https://doi.org/10.1016/j.jas.2014.05.034

Faisal, A., Stout, D., Apel, J., & Bradley, B. (2010). The Manipulative Complexity of Lower Paleolithic Stone Toolmaking. *PLOS ONE*, *5*(11), e13718. https://doi.org/10.1371/journal.pone.0013718

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*(6), 381–391. https://doi.org/10.1037/h0055392

García-Medrano, P., Ollé, A., Ashton, N., & Roberts, M. B. (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, P., & Högberg, A. (2017). The archaeology of teaching and the evolution of homo docens. *Current Anthropology*, *58*(2), 188–208. https://doi.org/10.1086/691178

Geribàs, N., Mosquera, M., & Vergès, J. M. (2010). What novice knappers have to learn to become expert stone toolmakers. *Journal of Archaeological Science*, *37*(11), 2857–2870. https://doi.org/10.1016/j.jas.2010.06.026

Gowlett, J. 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, D. A., & Berg, E. (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), 404–411. https://doi.org/10.1037/h0059831

Hecht, E. E., Gutman, D. A., Bradley, B. A., Preuss, T. M., & Stout, D. (2015). Virtual dissection and comparative connectivity of the superior longitudinal fasciculus in chimpanzees and humans. *NeuroImage*, *108*, 124–137. https://doi.org/10.1016/j.neuroimage.2014.12.039

Hecht, E. E., Gutman, D. A., Khreisheh, N., Taylor, S. V., Kilner, J. M., Faisal, A. A., Bradley, B. A., Chaminade, T., & Stout, D. (2015). Acquisition of Paleolithic toolmaking abilities involves structural remodeling to inferior frontoparietal regions. *Brain Structure & Function*, *220*(4), 2315–2331. https://doi.org/10.1007/s00429-014-0789-6

Hecht, E. E., Gutman, D. A., Preuss, T. M., Sanchez, M. M., Parr, L. A., & Rilling, J. K. (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), 1014–1024. https://doi.org/10.1093/cercor/bhs097

Hecht, E. E., Murphy, L. E., Gutman, D. A., Votaw, J. R., Schuster, D. M., Preuss, T. M., Orban, G. A., Stout, D., & Parr, L. A. (2013). Differences in neural activation for object-directed grasping in chimpanzees and humans. *The Journal of Neuroscience*, *33*(35), 14117–14134. https://doi.org/10.1523/JNEUROSCI.2172-13.2013

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Hewes, G. W. (1993). A history of speculation on the relation between tools and language. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution* (pp. 20–31). Cambridge University Press.

Heyes, C. (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

Hinton, G. E., & Nowlan, S. J. (1996). How learning can guide evolution. In R. K. Belew & M. Mitchell (Eds.), *Adaptive individuals in evolving populations: Models and algorithms* (pp. 447–454). Addison-Wesley Publishing Company.

Isaac, G. 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), 275–288. https://doi.org/10.1111/j.1749-6632.1976.tb25494.x

Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and instruction*. Lawrence Erlbaum.

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003

Key, A. J. M., & Dunmore, C. J. (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*, 60–69. https://doi.org/10.1016/j.jhevol.2014.08.006

Key, A. J. M., & Dunmore, C. J. (2018). Manual restrictions on Palaeolithic technological behaviours. *PeerJ*, *6*, e5399. https://doi.org/10.7717/peerj.5399

Key, A. J. M., & Lycett, S. J. (2014). Are bigger flakes always better? An experimental assessment of flake size variation on cutting efficiency and loading. *Journal of Archaeological Science*, *41*, 140–146. https://doi.org/10.1016/j.jas.2013.07.033

Key, A. J. M., & Lycett, S. J. (2019). Biometric variables predict stone tool functional performance more effectively than tool-form attributes: a case study in handaxe loading capabilities. *Archaeometry*, *61*(3), 539–555. https://doi.org/10.1111/arcm.12439

Khreisheh, N. N., Davies, D., & Bradley, B. A. (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, M. A. (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, K. N. (2017). The origins of language in teaching. *Psychonomic Bulletin & Review*, *24*(1), 225–231. https://doi.org/10.3758/s13423-016-1077-7

Levinson, S. C., & Holler, J. (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., Guardiola, M., & Mosquera, M. (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, B. (2017). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*, *24*(2), 424–450. https://doi.org/10.1007/s10816-015-9272-9

Marzke, M. W., Toth, N., Schick, K., Reece, S., Steinberg, B., Hunt, K., Linscheid, R. L., & An, K.-N. (1998). EMG study of hand muscle recruitment during hard hammer percussion manufacture of Oldowan tools. *American Journal of Physical Anthropology*, *105*(3), 315–332. https://doi.org/10.1002/(SICI)1096-8644(199803)105:3%3C315::AID-AJPA3%3E3.0.CO;2-Q

Mateos, A., Terradillos-Bernal, M., & Rodríguez, J. (2019). Energy Cost of Stone Knapping. *Journal of Archaeological Method and Theory*, *26*(2), 561–580. https://doi.org/10.1007/s10816-018-9382-2

Miu, E., Gulley, N., Laland, K. N., & Rendell, L. (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, L., Kurvers, R. H. J. M., & van den Bos, W. (2019). Unleashing the BEAST: a brief measure of human social information use. *Evolution and Human Behavior*, *40*(5), 492–499. https://doi.org/10.1016/j.evolhumbehav.2019.06.005

Montagu, A. (1976). Toolmaking, Hunting, and the Origin of Language. *Annals of the New York Academy of Sciences*, *280*(1), 266–274. https://doi.org/10.1111/j.1749-6632.1976.tb25493.x

Morgan, T. J. H., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M., Cross, C. P., Evans, C., Kearney, R., de la Torre, I., Whiten, A., & Laland, K. N. (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, T., Bril, B., & Rein, R. (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), 155–167. https://doi.org/10.1016/j.jhevol.2010.04.006

Oakley, K. P. (1949). *Man the toolmaker*. Trustees of the British Museum.

Ohnuma, K., Aoki, K., & Akazawa, A. T. (1997). Transmission of tool-making through verbal and non-verbal commu-nication: Preliminary experiments in levallois flake production. *Anthropological Science*, *105*(3), 159–168. https://doi.org/10.1537/ase.105.159

Pargeter, J., Khreisheh, N., Shea, J. J., & Stout, D. (2020). Knowledge vs. know-how? Dissecting the foundations of stone knapping skill. *Journal of Human Evolution*, *145*, 102807. https://doi.org/10.1016/j.jhevol.2020.102807

Pargeter, J., Khreisheh, N., & Stout, D. (2019). Understanding stone tool-making skill acquisition: Experimental methods and evolutionary implications. *Journal of Human Evolution*, *133*, 146–166. https://doi.org/10.1016/j.jhevol.2019.05.010

Pelegrin, J. (1990). Prehistoric Lithic Technology : Some Aspects of Research. *Archaeological Review from Cambridge*, *9*(1), 116–125. /paper/Prehistoric-Lithic-Technology-%3A-Some-Aspects-of-Pelegrin/5e02fc2a5280ac128727275ab6b833756e6a6056

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697. https://doi.org/10.1016/j.neuron.2011.11.001

Prasciunas, M. M. (2007). Bifacial Cores and Flake Production Efficiency: An Experimental Test of Technological Assumptions. *American Antiquity*, *72*(2), 334–348. https://doi.org/10.2307/40035817

Putt, S. S. (2015). The origins of stone tool reduction and the transition to knapping: An experimental approach. *Journal of Archaeological Science: Reports*, *2*, 51–60. https://doi.org/10.101

6/j.jasrep.2015.01.004

Putt, S. S., Wijeakumar, S., Franciscus, R. G., & Spencer, J. P. (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, S. S., Wijeakumar, S., & Spencer, J. P. (2019). Prefrontal cortex activation supports the emergence of early stone age toolmaking skill. *NeuroImage*, *199*, 57–69. https://doi.org/10.1016/j.neuroimage.2019.05.056

Putt, S. S., Woods, A. D., & Franciscus, R. G. (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, R., Nonaka, T., & Bril, B. (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

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., Fogarty, L., Ghirlanda, S., Lillicrap, T., & Laland, K. N. (2010). Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science*, *328*(5975), 208–213. https://doi.org/10.1126/science.1184719

Reti, J. 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, V., Bril, B., & Dietrich, G. (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, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W. (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, H., Kasagi, F., Yamada, M., & Fujita, S. (2007). Grip Strength Predicts Cause-Specific Mortality in Middle-Aged and Elderly Persons. *The American Journal of Medicine*, *120*(4), 337–342. https://doi.org/10.1016/j.amjmed.2006.04.018

Schillinger, K., Mesoudi, A., & Lycett, S. J. (2014). Copying Error and the Cultural Evolution of “Additive” vs. “Reductive” Material Traditions: An Experimental Assessment. *American Antiquity*, *79*(1), 128–143. https://doi.org/10.7183/0002-7316.79.1.128

Shallice, T., Broadbent, D. E., & Weiskrantz, L. (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, J. J. (2015). Making and using stone tools: Advice for learners and teachers and insights for archaeologists. *Lithic Technology*, *40*(3), 231–248. https://doi.org/10.1179/2051618515Y.0000000011

Shea, J. J. (2016). *Stone tools in human evolution: Behavioral differences among technological primates*. Cambridge University Press. https://doi.org/10.1017/9781316389355

Sherwood, C. C., & Gómez-Robles, A. (2017). Brain plasticity and human evolution. *Annual Review of Anthropology*, *46*(1), 399–419. https://doi.org/10.1146/annurev-anthro-102215-100009

Stout, D. (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

Stout, D. (2010). Possible relations between language and technology in human evolution. In A. Nowell & I. Davidson (Eds.), *Stone tools and the evolution of human cognition* (pp. 159–184). University Press of Colorado.

Stout, D. (2013). Neuroscience of technology. In P. J. Richerson & M. H. Christiansen (Eds.), *Cultural evolution: Society, technology, language, and religion* (pp. 157–173). The MIT Press.

Stout, D., Apel, J., Commander, J., & Roberts, M. (2014). Late Acheulean technology and cognition at Boxgrove, UK. *Journal of Archaeological Science*, *41*, 576–590. https://doi.org/10.1016/j.jas.2013.10.001

Stout, D., & Chaminade, T. (2007). The evolutionary neuroscience of tool making. *Neuropsychologia*, *45*(5), 1091–1100. https://doi.org/10.1016/j.neuropsychologia.2006.09.014

Stout, D., & Chaminade, T. (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, D., & Hecht, E. E. (2017). Evolutionary neuroscience of cumulative culture. *Proceedings of* *the National Academy of Sciences*, *114*(30), 7861–7868. https://doi.org/10.1073/pnas.1620738114

Stout, D., Hecht, E., Khreisheh, N., Bradley, B., & Chaminade, T. (2015). Cognitive Demands of Lower Paleolithic Toolmaking. *PLOS ONE*, *10*(4), e0121804. https://doi.org/10.1371/journal.pone.0121804

Stout, D., & Khreisheh, N. (2015). Skill Learning and Human Brain Evolution: An Experimental Approach. *Cambridge Archaeological Journal*, *25*(4), 867–875. https://doi.org/10.1017/S0959774315000359

Stout, D., Passingham, R., Frith, C., Apel, J., & Chaminade, T. (2011). Technology, expertise and social cognition in human evolution. *The European Journal of Neuroscience*, *33*(7), 1328–1338. https://doi.org/10.1111/j.1460-9568.2011.07619.x

Stout, D., Quade, J., Semaw, S., Rogers, M. J., & Levin, N. E. (2005). Raw material selectivity of the earliest stone toolmakers at Gona, Afar, Ethiopia. *Journal of Human Evolution*, *48*(4), 365–380. https://doi.org/10.1016/j.jhevol.2004.10.006

Stout, D., Rogers, M. J., Jaeggi, A. V., & Semaw, S. (2019). Archaeology and the origins of human cumulative culture: A case study from the earliest oldowan at gona, ethiopia. *Current Anthropology*, *60*(3), 309–340. https://doi.org/10.1086/703173

Stout, D., & Semaw, S. (2006). Knapping skill of the earliest stone toolmakers: Insights from the study of modern human novices. In N. Toth & K. Schick (Eds.), *The Oldowan: Case studies into the earliest Stone Age* (pp. 307–320). Stone Age Institute Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.

Tehrani, J. J., & Riede, F. (2008). Towards an archaeology of pedagogy: Learning, teaching and the generation of material culture traditions. *World Archaeology*, *40*(3), 316–331. https://doi.org/10.1080/00438240802261267

Tennie, C., Premo, L. S., Braun, D. R., & McPherron, S. P. (2017). Early stone tools and cultural transmission: Resetting the null hypothesis. *Current Anthropology*, *58*(5), 652–672. https://doi.org/10.1086/693846

Toelch, U., Bruce, M. J., Newson, L., Richerson, P. J., & Reader, S. M. (2014). Individual consistency and flexibility in human social information use. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1776), 20132864. https://doi.org/10.1098/rspb.2013.2864

Toth, N., & Schick, K. (1993). Early stone industries and inferences regarding language and cognition. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution* (pp. 346–362). Cambridge University Press.

Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & Cognition*, *33*(2), 213–220. https://doi.org/10.3758/BF03195310

Vostroknutov, A., Polonio, L., & Coricelli, G. (2018). The Role of Intelligence in Social Learning. *Scientific Reports*, *8*(1), 6896. https://doi.org/10.1038/s41598-018-25289-9

Washburn, S. L. (1960). Tools and human evolution. *Scientific American*, *203*(3), 62–75. https://doi.org/10.1038/scientificamerican0960-62

Whiten, A. (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, J. C. (1994). *Flintknapping: Making and Understanding Stone Tools*. University of Texas Press.

Wilkins, J. (2018). The Point is the Point: Emulative social learning and weapon manufacture in the Middle Stone Age of South Africa. In M. J. O'Brien, B. Buchanan, & M. I. Eren (Eds.), *Convergent Evolution in Stone-Tool Technology* (pp. 153–174). The MIT Press.

Williams-Hatala, E. M., Hatala, K. G., Gordon, M., Key, A., Kasper, M., & Kivell, T. L. (2018). The manual pressures of stone tool behaviors and their implications for the evolution of the human hand. *Journal of Human Evolution*, *119*, 14–26. https://doi.org/10.1016/j.jhevol.2018.02.008

Williams-Hatala, E. M., Hatala, K. G., Key, A., Dunmore, C. J., Kasper, M., Gordon, M., & Kivell, T. L. (2021). Kinetics of stone tool production among novice and expert tool makers. *American Journal of Physical Anthropology*, *174*(4), 714–727. https://doi.org/10.1002/ajpa.24159

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

Wind, A. E., Takken, T., Helders, P. J. M., & Engelbert, R. H. H. (2010). Is grip strength a predictor for total muscle strength in healthy children, adolescents, and young adults? *European Journal of Pediatrics*, *169*(3), 281–287. https://doi.org/10.1007/s00431-009-1010-4

Wynn, T. (1979). The intelligence of later acheulean hominids. *Man*, *14*(3), 371–391. https://doi.org/10.2307/2801865

Wynn, T. (2017). Evolutionary cognitive archaeology. In T. Wynn & F. Coolidge (Eds.), *Cognitive models in Palaeolithic archaeology* (pp. 1–20). Oxford University Press.

Wynn, T., & Coolidge, F. L. (2004). The expert Neandertal mind. *Journal of Human Evolution*, *46*(4), 467–487. https://doi.org/10.1016/j.jhevol.2004.01.005

Wynn, T., & Coolidge, F. L. (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, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685