

1 Testing the social, cognitive, and motor foundations of
2 Paleolithic skill reproduction

3 Justin Pargeter* Cheng Liu† Megan Beney Kilgore‡ Aditi Majoe§
4 Dietrich Stout¶

5 **Abstract**

6 Stone tools provide key evidence of human cognitive evolution but remain challenging
7 to interpret. Tool-making skill-learning has been understudied even though: 1) the most
8 salient cognitive demands of tool-making should occur during learning, and 2) variation
9 in learning aptitude would have provided the raw material for any past selection acting on
10 tool-making ability. However, we know very little about the cognitive prerequisites of learning
11 under different information transmission conditions that may have prevailed during the
12 Paleolithic. This paper presents results from an exploratory study to trial new experimental
13 methods for studying the effect of learning conditions and individual differences on Oldowan
14 flake-tool-making skill acquisition. We trained 23 participants for 2 hours to make stone
15 flakes under two different instructional conditions (observation only vs. direct active teaching).
16 We employed appropriate raw materials, a moderate practice time duration, and in-person,
17 fully interactive instruction. Participant performance was evaluated through an analysis of
18 the stone artifacts produced. We compared performance across experimental groups with
19 respect to individual participant differences in grip strength, motor accuracy, and cognitive
20 function measured for the study. Our results show that 2 hours are insufficient to document
21 learning-related performance change. However, teaching reduces variability in knapping
22 rate, methods, and outcomes during early-stage learning, thus increasing the reliability of
23 skill reproduction. Instruction also altered knapping quality vs. quantity trade-offs in the two
24 groups and dramatically changed the effects of individual differences in strength, visuospatial
25 working memory, and social learning tendencies on knapping outcomes. Our results provide
26 further support for the hypothetical co-evolution of teaching, language, and tool-making,
27 suggest that the presence/absence of instruction can fundamentally alter learning-related
28 selection pressures on individuals, and offer lessons for the design of future experiments.

29 **Keywords:** Oldowan; Stone tool-making; Social learning; Individual variation; Cognitive
30 aptitudes; Motor skills

31 **Contents**

32 1 Introduction	2
33 1.1 Individual Differences	4

*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; raylc1996@outlook.com

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

§Independent Researcher; am2752@cantab.ac.uk

¶Department of Anthropology, Emory University, Atlanta, GA, USA; dwstout@emory.edu

34	1.2 Teaching, Language, and Tool Making	7
35	1.3 Raw materials and knapping skill	9
36	2 Materials and Methods	10
37	2.1 Participants	10
38	2.2 Study Visit	11
39	2.3 Individual Difference Measures	11
40	2.4 Stone Tool Making	13
41	2.5 Lithic Analysis	19
42	2.6 Statistical Analyses	21
43	3 Results	21
44	3.1 Principal Component analyses	22
45	3.2 Relationships between Performance Measures	24
46	3.3 Do trained, untrained, and expert knappers perform differently?	25
47	3.4 Does performance change over time?	29
48	3.5 Do individual differences in motor skill and psychometric measures predict flaking performance?	32
49	3.6 Behavioral observations	40
51	4 Discussion	41
52	4.1 Variance Reduction	42
53	4.2 Knapping Behaviors	42
54	4.3 Learning Strategies	44
55	4.4 Limitations and Prospects	46
56	5 Conclusions	47
57	References	48

58 **1 Introduction**

59 Stone tools have long been seen as a key source of evidence for understanding human behavioral
 60 and cognitive evolution (Darwin, 1871; Oakley, 1949; Washburn, 1960). Pathbreaking attempts to
 61 infer specific cognitive capacities from this evidence largely focused on the basic requirements
 62 of tool production (Gowlett, 1984; Isaac, 1976; Wynn, 1979; Wynn & Coolidge, 2004). More
 63 recently, researchers have directed increasing attention to the processes and demands of stone
 64 tool-making skill acquisition (Cataldo et al., 2018; Duke & Pargeter, 2015; Geribàs et al., 2010;
 65 Hecht, Gutman, Khreisheh, et al., 2015; Lombao et al., 2017; Morgan et al., 2015; Nonaka et al.,
 66 2010; Pargeter et al., 2020; Pargeter et al., 2019; Putt et al., 2017, 2019; Putt et al., 2014; Roux
 67 et al., 1995; Shipton, 2018; Stout et al., 2005; Stout et al., 2011; Stout, 2002; Stout & Khreisheh,
 68 2015). This increased attention is motivated by the expectation that tool-making's most salient

69 cognitive demands should occur during learning rather than routine expert performance (Stout
70 & Khreisheh, 2015). Researchers are also increasingly focused on the relevance of different social
71 learning mechanisms such as imitation (Rein et al., 2014; Stout et al., 2019), emulation (Tehrani &
72 Riede, 2008; Wilkins, 2018), and language (Cataldo et al., 2018; Lombao et al., 2017; Morgan et al.,
73 2015; Ohnuma et al., 1997; Putt et al., 2017; Putt et al., 2014) to the reproduction of Paleolithic
74 technologies.

75 Studies investigating these questions have used a range of different experimental designs (e.g.,
76 varying technological goals/instructions, training times, raw materials, live vs. recorded instruc-
77 tion, lithic/skill assessment metrics, pseudo-knapping tasks etc.) and reached disparate conclu-
78 sions regarding skill acquisition's neurocognitive and social foundations. It is plausible that these
79 discordant results reflect actual diversity in how humans acquire and master stone tool-making
80 skills. However, this failure of results to generalize across artificial experimental manipulations (cf.
81 Yarkoni, 2020) also raises doubts regarding the external validity (Eren et al., 2016) of conclusions
82 concerning real-world Paleolithic learning contexts. To address this, we conducted an exploratory
83 study that draws on lessons from previous research to balance the pragmatic and theoretical
84 trade-offs inherent in experimental studies of stone knapping skill acquisition (Pargeter et al.,
85 2019; Stout & Khreisheh, 2015).

86 Compared to artificial experimental tasks, learning real-world skills like stone knapping is highly
87 demanding of time and materials and difficult to control experimentally without sacrificing
88 generalizability to real world conditions. Prior efforts have attempted to navigate these challenges
89 by using various combinations of 1) inauthentic raw materials that are less expensive, easier to
90 standardize, and/or easier to knap, 2) video-recorded instruction that is uniform across partici-
91 pants and less demanding of experimenter time, 3) short learning periods, 4) small sample sizes,
92 and 5) single learning conditions. The difficulty of interpreting results from this growing literature
93 led Stout and Khreisheh (2015: 870, emphasis original) to call for "studies with sufficient sample
94 sizes to manipulate learning conditions (e.g. instruction, motivation) and assess individual varia-
95 tion (e.g. performance, psychometrics, neuroanatomy) that *also* have realistic learning periods."
96 The current study attempts to strike a viable balance between these demands by investigating
97 early-stage learning of a relatively simple technology (least effort, "Oldowan," flake production
98 (Reti, 2016; Shea, 2016)) under two instructional conditions while collecting data on individual
99 differences in strength, coordination, cognition, and social learning. Unlike any previous study,

¹⁰⁰ this allows us to address the likelihood that interactions with individual participant differences in
¹⁰¹ aptitude or learning style might impact group effects of training conditions.

¹⁰² We focus on early-stage learning because it is relatively rapid, variable across individuals, and
¹⁰³ predictive of later outcomes (Pargeter et al., 2019; Putt et al., 2019; Stout & Khreich, 2015). The
¹⁰⁴ study thus provides a reasonable expectation of generating meaningful data on skill and learning
¹⁰⁵ variation while minimizing training costs. Moreover, understanding the minimum training times
¹⁰⁶ necessary to detect changes in tool making skill will help archaeologists design more realistic and
¹⁰⁷ cost-effective experiments. We limited our study to only two learning conditions (observation
¹⁰⁸ only vs. active teaching) to further manage costs. This study design targets a key controversy in
¹⁰⁹ human evolution, namely the origins of teaching and language (Gärdenfors & Högberg, 2017;
¹¹⁰ Morgan et al., 2015) while avoiding artificial manipulations of dubious relevance to real-world
¹¹¹ Paleolithic learning (e.g., Cataldo et al., 2018). These choices allowed us to invest more in other
¹¹² aspects of research design that we identified as theoretically necessary, including measuring
¹¹³ individual differences in cognition and behavior, including an in-person, fully interactive teaching
¹¹⁴ condition, and using naturalistic raw materials. The sample size remained small in this internally
¹¹⁵ funded exploratory study, but we could easily scale it up at funding levels typical of pre- and
¹¹⁶ post-doctoral research grants in archaeology.

¹¹⁷ 1.1 Individual Differences

¹¹⁸ “*The many slight differences... being observed in the individuals of the same species inhabiting*
¹¹⁹ *the same confined locality, may be called individual differences... These individual differences are*
¹²⁰ *of the highest importance to us, for they are often inherited... and they thus afford materials for*
¹²¹ *natural selection to act on and accumulate...*” (Darwin, 1859, Chapter 2)

¹²² Individuals vary in aptitude and learning style for skills (Jonassen & Grabowski, 1993), but research
¹²³ on knapping skill acquisition has largely ignored this. These studies have instead focused on
¹²⁴ group effects of different experimental conditions. There are good pragmatic reasons for this,
¹²⁵ as individual difference studies typically require larger sample sizes and other data collection.
¹²⁶ However, overlooking these distinctions is not ideal since individual differences can provide
¹²⁷ valuable insight into cognition and behavior's mechanisms, development, and evolution (Boogert
¹²⁸ et al., 2018). Researchers can use patterns of association between cognitive traits and behavioral
¹²⁹ performance to test hypotheses about the cognitive demands of learning particular skills and

130 the likely targets of natural selection acting on aptitude. More prosaically, individual differences
131 can introduce an unexamined and uncontrolled source of variation in group-level results. This
132 effect is especially true in the relatively small samples of convenience typical of experimental
133 archaeology.

134 While testing hypotheses in evolutionary cognitive archaeology remains a considerable challenge
135 ([Wynn, 2017](#)), investigating individual variation in modern research participants represents one
136 promising direction. For any behavior of archaeological interest, we could expect that standing
137 variation in modern populations should remain relevant to normal variation in learning aptitude.
138 The presence of trait variation without impact on learning aptitude would provide strong evidence
139 against the plausibility of the proposed evolutionary relationship. We would not expect an an
140 absence of variation (i.e., past fixation and rigorous developmental canalization) given the known
141 variability of human brains and cognition ([Barrett, 2020; Sherwood & Gómez-Robles, 2017](#)). Any
142 confirmatory findings of trait-aptitude correspondence would have the testable implication that
143 humans should be evolutionarily derived along the same dimension ([Hecht, Gutman, Bradley, et
144 al., 2015](#)).

145 To date, a small number of “neuroarchaeological” studies have reported associations between
146 individual knapping performance and brain structure or physiological responses. Hecht et al.
147 ([2015](#)) reported training-related changes in white matter structure correlated with individual
148 differences in practice time and striking accuracy change. The regional patterning of white matter
149 changes also varied across individuals. Only those who displayed early increases under the right
150 ventral precentral gyrus (premotor cortex involved in movement planning and guidance) showed
151 striking accuracy improvement over the training period. Putt et al. ([2019](#)) similarly found that the
152 proportion of flakes-to-shatter produced by individuals during handaxe-making correlated with
153 dorsal precentral gyrus (motor cortex) activation. Pargeter et al. ([2020](#)) used a flake prediction
154 paradigm (modeled after [Nonaka et al., 2010](#)) to confirm that striking force and accuracy are
155 important determinants of handaxe-making success. These findings point to the central role of
156 perceptual-motor systems ([Stout & Chaminade, 2007](#)) and coordination ([Roux et al., 1995](#)) in
157 knapping skill acquisition. In addition, Putt et al. ([2019](#)) found successful flake production to
158 be associated with prefrontal (working memory/cognitive control) activation, and Stout et al.
159 ([2015](#)) found that prefrontal activation correlated with success at a strategic judgement (platform
160 selection) task which in turn was predictive of success at out-of-scanner handaxe production.

161 Such investigations are thus starting to chart the contributions of different neural systems to
162 aspects of knapping skill acquisition, including motor and cognitive control aspects. To date,
163 however, the cognitive/functional interpretation of systems identified in this manner has relied
164 mainly on informal reverse inference (reasoning backward from observed activations to inferred
165 mental processes) from published studies of other tasks that activated the same regions. Re-
166 searchers widely regard this approach as problematic (Poldrack, 2011). Here we take a more direct,
167 psychometric approach to measuring individual differences in perceptual-motor coordination
168 and cognition.

169 Studies implement psychometric instruments (e.g., tasks, questionnaires) to assess variation in
170 cognitive traits and states, such as fluid intelligence, working memory, attention, motivation, and
171 personality. These measures have been of theoretical interest to the broader field of cognitive
172 archaeology (e.g., Wynn & Coolidge, 2016). It is thus surprising that experimental studies have
173 almost entirely neglected them. In the only published example we are aware of, Pargeter et
174 al. (2019) reported significant effects of variation in planning and problem solving (Tower of
175 London test (Shallice et al., 1982)) and cognitive set shifting (Wisconsin Card Sort test (Grant
176 & Berg, 1948)) on early stage handaxe learning. Of course, cognition is not the only thing that
177 can affect knapping performance. Flake prediction experiments highlight the importance of
178 regulating movement speed/accuracy trade-offs (Nonaka et al., 2010; Pargeter et al., 2020) and
179 studies of muscle recruitment (Marzke et al., 1998) and manual pressure (Key & Dunmore, 2018;
180 Williams-Hatala et al., 2018) during knapping highlight basic strength requirements. Along these
181 lines, Key and Lycett (2019) found that individual differences in hand size, shape, and especially
182 grip strength were better predictors of force loading during stone tool use than were attributes of
183 the tools themselves. However, we are unaware of any such studies of biometric influences on
184 variation in knapping success. Finally, individual, and contextual differences in adopting social
185 vs. individual learning strategies (Miu et al., 2020) might also affect learning outcomes. We are
186 again unaware of any previous studies that have assessed such effects. This study evaluated all
187 participants with a battery of tests, including grip strength, movement speed/accuracy, spatial
188 working memory, fluid intelligence, and tendency to use social information. We were particularly
189 interested in the possibility that these variables might not only impact learning generally but
190 might also have different effects under different learning conditions.

191 **1.2 Teaching, Language, and Tool Making**

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

195 Possible links between tool making and language have been a subject of speculation for nearly
196 150 years ([Engles, 2003, p. \[1873\]](#)), if not longer ([Hewes, 1993](#)), although compelling empirical
197 tests have remained elusive. Over 25 years ago, Toth and Schick ([1993](#)) suggested that experiments
198 teaching modern participants to make stone tools in verbal and non-verbal conditions could
199 test the importance of language in the social reproduction of Paleolithic technologies. Ohnuma
200 et al. ([1997](#)) were the first to implement this suggestion in a study of Levallois flake production,
201 followed by more recent studies of handaxe making ([Putt et al., 2017; Putt et al., 2014](#)) and simple
202 flake production ([Cataldo et al., 2018; Lombao et al., 2017; Morgan et al., 2015](#)). This study
203 design trajectory reflects a recent interest in the hypothesis that language might be an adaptation
204 for teaching (e.g., [Laland, 2017; Stout & Chaminade, 2012](#)). Teaching and learning demands
205 of Paleolithic tool-making would thus provide evidence of selective contexts favoring language
206 evolution ([Montagu, 1976; Morgan et al., 2015; Stout, 2010](#)).

207 However, Toth and Schick ([1993](#)) pointed out that extinct hominid learning strategies and capaci-
208 ties might differ from modern experimental participants. Even leaving aside potential species
209 differences in social learning (cf. [Morgan et al., 2015; Stout et al., 2019](#)), reliance on explicit verbal
210 instruction varies widely across modern human societies (e.g., [Boyette & Hewlett, 2017](#)). The
211 WEIRD (Western, educated, industrialized, rich, democratic ([Henrich et al., 2010](#))) teachers and
212 learners typical of knapping experiments arguably represent an extreme bias toward such instruc-
213 tion. Simply instructing such participants not to speak during an experiment (or to demonstrate
214 but not gesture, etc. ([Morgan et al., 2015](#))) is likely to underestimate the efficacy of non-verbal
215 teaching and learning in cultural contexts where it is more common, let alone in a hypothetical
216 pre-linguistic hominid species (cf. [Fuentes, 2015](#)).

217 Such concerns are exacerbated in experiments using pre-recorded instructional videos or brief
218 training periods. Video does not allow the interactive teaching favored even in formal academic
219 knapping classes (e.g., [Shea, 2015](#)) and is almost undoubtedly typical of traditional learning
220 contexts (e.g., [Stout, 2002](#)). Researchers do not know how video presentation generally affects the

efficacy of teaching stone knapping or the relative effectiveness of different forms of instruction. Moreover, 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 multimodal (Levinson & Holler, 2014) communication strategies as they would naturally do, for example, through pointing and pantomime. Simply removing a communication channel without allowing 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 & Khriesheh, 2015; Whiten, 2015). Even the most extended training times (Pargeter et al., 2019; Stout & Khriesheh, 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, Khriesheh, 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 providing instruction or interacting with learners in any way during this activity. This absence of teaching is again a familiar social context for our participants and did not require any novel behaviors from the instructor. Although we make no assumptions regarding learning mechanisms, it matches the “imitation/emulation” condition of Morgan et al. (2015). We did not include a “reverse engineering” or “end-state emulation” condition where only finished products were visible. Researchers have advocated these states as an important baseline or control condition (Whiten, 2015) to distinguish observational from individual learning. However, they are 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

253 observation/imitation capacities (Hecht, Gutman, et al., 2013; Hecht, Murphy, et al., 2013; Stout
254 et al., 2019).

255 We selected a two-hour learning period for both pragmatic and theoretical reasons. Pargeter et al.
256 (Pargeter et al., 2019) found that even ~90 hours of fully interactive instruction and practice was
257 insufficient to achieve handaxe-making skills comparable to the later Acheulean site of Boxgrove
258 (García-Medrano et al., 2019; Stout et al., 2014), and estimated actual time to mastery as ranging
259 from 121 to 441 hours for different participants. However, they observed the greatest, fastest, and
260 most individually variable skill increases during the first 20 hours of practice. In addition, initial
261 performance moderately correlated with later achievement. This result suggests that studying
262 early-stage learning may be a practical alternative, especially for research investigating individual
263 differences in aptitude. Studies of simple flake production similarly document considerable initial
264 variation (Stout & Khriesheh, 2015) and rapid early progress (Putt et al., 2019; Stout & Khriesheh,
265 2015; Stout & Semaw, 2006). We designed the current study to test the utility of studying learning
266 and variation during the first two hours of simple flaking instruction/practice in hopes of finding
267 a viable compromise between experimental realism and cost.

268 **1.3 Raw materials and knapping skill**

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

273 Lithic raw materials vary in size, shape, and fracture mechanical properties that affect the difficulty
274 of achieving different knapping goals (Eren et al., 2014). Unfortunately, it can be difficult and/or
275 expensive to procure authentic raw materials. Experimental studies of knapping skill have often
276 used proxy materials such as flint (Cataldo et al., 2018; Morgan et al., 2015; Nonaka et al., 2010),
277 limestone (Stout & Semaw, 2006), porcelain (Khriesheh et al., 2013), or heat-treated chert (Putt et
278 al., 2017, 2019; Putt et al., 2014) to model Oldowan and early Acheulean technologies executed in
279 other materials. As well as being more readily available, these proxies are generally easier to knap.
280 This study design has the benefit of reducing required practice time. Still, it is unclear how it might
281 affect learning demands more generally or the efficacy of different learning conditions/strategies
282 precisely.

283 Some designs have attempted to closely match experimental and archaeological raw material
284 types (Duke & Pargeter, 2015; Pargeter et al., 2019; Stout et al., 2011). However, raw materials
285 vary across individual clasts within as well as between types. This choice has led to an interest
286 in standardizing experimental core morphology (Nonaka et al., 2010) and composition, even if
287 this means using artificial materials such as porcelain (Khreicheh et al., 2013), brick (Geribàs
288 et al., 2010; Lombao et al., 2017), or foam blocks (Schillinger et al., 2014). Such manipulations
289 enhance experimental control and internal validity (Eren et al., 2016) at the expense of external
290 generalizability to actual archaeological conditions. Specifically, they allow more robust results
291 from smaller samples but eliminate a core element of real-world knapping skill: the ability to
292 produce consistent results from variable materials (Pelegrin, 1990; Stout, 2013). For example,
293 Pargeter et al. (2020) found that predicting specific flaking outcomes on actual handaxe preforms
294 was both more complex and less technologically important than expected from previous work
295 with standardized, frustum-shaped cores (Nonaka et al., 2010). The alternative to control is to
296 incorporate raw material size, shape, and composition as experimental variables (e.g., Stout et al.,
297 2019). This choice allows consideration of raw material selection and response to variation as
298 aspects of skill but correspondingly increases the sample sizes required to identify patterning.
299 In considering these issues, we again chose to explore a middle path between pragmatism and
300 realism by employing commercially purchased basalt like that known from East African Oldowan
301 sites. We allowed clast size and shape to vary within set limits and selected the particular clasts
302 provided to each participant to approximate the same distribution.

303 2 Materials and Methods

304 The Emory Institutional Review Board approved this research (IRB00113024). All participants
305 provided written informed consent and completed a video release form (<https://databrary.org/support/irb/release-template.html>).

307 2.1 Participants

308 Twenty-four adult participants with no prior stone knapping experience were recruited from the
309 Emory community using paper fliers and e-mail listserv advertisements. Eleven participants
310 (6 female, five male) completed the uninstructed condition, and 12 (8 female, four male) com-
311 pleted the instructed condition. We could not replace one participant who failed to attend their

312 scheduled session, resulting in a total sample of 23.

313 **2.2 Study Visit**

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

322 We compensated participants for their time with a \$30 gift card. They also had the opportunity
323 to earn a performance bonus of \$5, \$10, \$15 or \$20 on the gift card. They were told that this
324 bonus would depend on “how well they did” on the last core of their practice session. The actual
325 performance measure was not specified, but to allow on the spot payment we used a simple
326 measure of the percentage of starting weight removed from the final core such that: > 30% earned
327 \$5, > 40% earned \$10, > 50% earned \$15, > 75% earned \$20.

328 **2.3 Individual Difference Measures**

329 We used five individual difference measures for this study:

- 330 1) *Grip strength* was measured in kilograms using an electronic hand dynamometer (Camry
331 EH101). We measured strength twice and kept the higher value. Grip strength is a simple
332 measure well correlated with overall muscular strength (Wind et al., 2010) and a range of
333 other health and fitness measures (Sasaki et al., 2007). Researchers have hypothesized that
334 grip strength is relevant for generating kinetic energy for fracture initiation (Nonaka et al.,
335 2010) as well as control and support of the hammerstone (Williams-Hatala et al., 2018) and
336 core (Faisal et al., 2010; Key & Dunmore, 2015).
- 337 2) *Motor accuracy* was assessed using a “Fitts Law” reciprocal tapping task. Fitts Law describes
338 the trade-off between speed and accuracy in human movement, classically measured
339 by tapping back and forth between two targets of varying size and spacing (Fitts, 1954).

340 Archaeologists have proposed (Pargeter et al., 2020; Stout, 2002) that managing this trade-
341 off is critical to the accurate application of appropriate force seen in skilled knapping
342 (Nonaka et al., 2010; Roux et al., 1995). We implemented this test on a Surface Pro tablet
343 running free software (FittsStudy Version 4.2.8, default settings) developed by the Accessible
344 Computing Experiences lab (Jacob O. Wobbrock, director) at the University of Washington
345 (depts.washington.edu/acelab/proj/fittsstudy/index.html). Participants use a touchscreen
346 pen to tap between ribbons on the screen, with average movement time as the performance
347 metric.

348 3) *Visuospatial working memory* is the capacity to “hold in mind,” which researchers
349 have hypothesized to be important in stone tool-making performance (Coolidge &
350 Wynn, 2005). It also might support a learning process known as ‘chunking,’ in which
351 subjects combine multiple items or operations into summary chunks stored in long-term
352 memory, which researchers think is essential in acquiring knapping and other skills
353 (Pargeter et al., 2019). We measured visuospatial working memory using a free n-back
354 task (wmp.education.uci.edu/software/) developed by the Working Memory and Plasticity
355 Laboratory at the University of California, Irvine (Susanne Jaeggi, PI) and implemented
356 in E-Prime software on a desktop computer. In this task, instructors tell participants to
357 remember the position of blue squares presented sequentially on the screen and touch
358 a key when the current position matches that 1, 2, 3...n iterations back. Progression to
359 blocks with increasing values of n is contingent on exceeding a threshold success rate. We
360 measured n-back task performance as the highest n achieved.

361 4) *Fluid intelligence* (Cattell, 1963) refers to the capacity to engage in abstract reasoning and
362 problem solving in a way that is minimally dependent on prior experience. It complements
363 “crystallized intelligence” (the ability to apply learned procedures and knowledge) as one of
364 the two factors (g_f, g_c) comprising so-called “general intelligence” (g). Fluid intelligence is
365 closely related to the executive control of attention and manipulation of the information
366 held in working memory (Engle, 2018). Researchers hypothesize that fluid intelligence
367 supports technological innovation (Coolidge & Wynn, 2005) and/or the intentional learning
368 of new skills (Stout & Khreich, 2015; Unsworth & Engle, 2005). We measured fluid
369 intelligence using the short version (Bilker et al., 2012) of the classic Raven Progressive Ma-
370 trices task, which requires participants to complete increasingly difficult pattern matching

371 questions.

372 5) The *use of social information* for learning and decision-making varies across individuals
373 and societies (Molleman et al., 2019). Such variation is a key topic for understanding
374 social learning and cultural evolutionary processes (Heyes, 2018; Kendal et al., 2018; Miu
375 et al., 2020) and represents a potential confound for assessing experimental effects of
376 different social learning conditions. We measured participants' tendency to rely on social
377 information vs. their insights using the Berlin Estimate AdjuStment Task (BEAST) developed
378 by Molleman et al. (2019). In this task, participants are presented with large arrays of items
379 on a screen and asked to estimate the number present. They are then provided with
380 another person's estimate and allowed to give a second estimate. The participants' average
381 adjustment between the first and second estimates measures their propensity to rely on
382 social information.

383 **2.4 Stone Tool Making**

384 After individual difference testing, participants engaged in a 2-hour stone tool making session,
385 with a 15-minute break after 1 hour. During these breaks, participants were instructed not to seek
386 additional training or information on stone tool-making (i.e., via the internet). For each group of
387 participants, we randomly assigned subjects to one of two experimental conditions: no teaching
388 or teaching. In both conditions, participants were first allowed to inspect and handle examples
389 made by the expert model/instructor (Figure 1) of the kind of stone tools (flakes) they were asked
390 to produce. They were told that their objective was to detach as many flakes as possible from
391 the materials provided. These instructions meant that even the uninstructed condition included
392 some minimal instruction (being told the objective). However, this was unavoidable without
393 creating a more elaborate and naturalistic context where participants would develop their own
394 technological goals. We expected that such a design would also increase behavioral variability,
395 demanding correspondingly larger samples of participants to identify patterns and making direct
396 comparisons with the instructed condition problematic.



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

397 **2.4.1 Raw Materials**

398 Each participant was provided with nine nodules for use over the 2-hour experiment. We made
399 these nodules from larger chunks of a fine-grained basalt purchased from neolithics.com by
400 fracturing them with a sledgehammer. This basalt has not been mechanically compared (e.g.,
401 rebound hardness, [Braun et al., 2009](#)) to East African basalts. Still, the basalt appears qualitatively
402 like finer-grained examples at Gona (D. Stout, pers. obs.). Spalling produced irregular, angular
403 chunks (**Figure 2**) for the experiment, weighing between 459g - 1876g (mean = 975g). All cores
404 were weighed, measured (Length, Width, Thickness), and painted white so that we could dis-
405 criminate new fracture surfaces from those created during production (see supplementary Table
406 1). Cores were sorted by shape and weight and then distributed evenly to each participant. As a

407 result, there were no significant difference across participants in the mean weight (ANOVA, df
408 = 22, F=0.3, p = 0.9; Levene test of homogeneity of variance = 1.04, df1=22, df2 = 184, p = 0.4) or
409 shape (Length × Width/Thickness: ANOVA, df = 22, F=0.4, p = 0.9; Levene statistic = 0.6, df1=22,
410 df2 = 184, p = 0.9) of cores provided. This was also true for cores provided to participants across
411 the two experimental conditions (Instructed vs. Uninstructed mean weight = 1001g vs. 956g, t =
412 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 =
413 205, p = 0.9, Levene's Test F = 3.8, p= 0.05). Participants were, however, allowed to choose which
414 cores to work on so that differences in the weight and shape of cores used across participants and
415 conditions could still emerge because of selection bias.

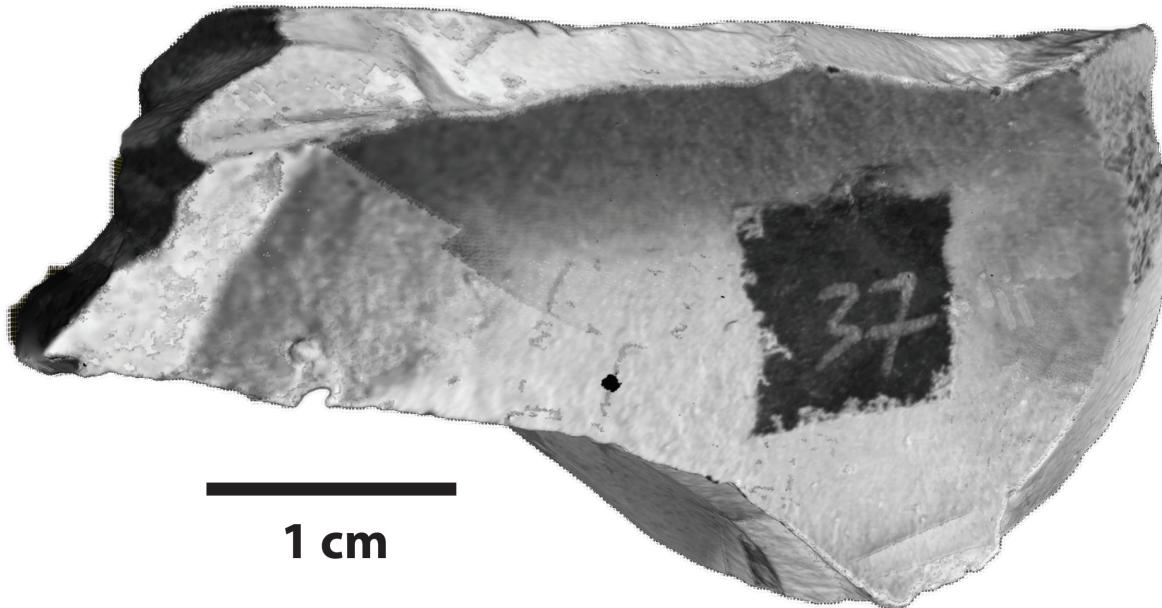


Figure 2: 3D scan of a basalt core prior to knapping. Cores were spray painted white to facilitate subsequent technological analyses of the flake scars and flaking intensity. Scale in the background is in 10mm blocks.

416 Sixty pounds of 3-to-5 inch basalt “Mexican Beach Pebbles” were purchased from a landscaping
417 supply company as hammerstones in the experiment. Of these, we selected 90 as suitable for use.
418 These weighed between 213g-1360g (mean = 425) and varied in elongation (L/W = 1.01 to 2.65)
419 and relative thickness ($L\times W/T$ = 90.48 to 283.67). We placed roughly two-thirds of the stones
420 in the middle of the knapping area ([Figure 3](#)) for participants to freely choose from during the
421 experiment. Broken hammerstones were replaced from the reserve to maintain a consistent
422 number and range of choices. We numbered each hammerstone and recorded each participants'

⁴²³ choices along with the number of the core(s) they worked on with a particular hammerstone.

⁴²⁴ **2.4.2 Experimental Conditions**

⁴²⁵ In both conditions, three researchers were present to record activities and collect materials.
⁴²⁶ Participants were seated in a circle ([Shea, 2015](#)), and experiments were video recorded using
⁴²⁷ two cameras. Participants were free to select hammerstones from the common pile and work
⁴²⁸ on any of their nine assigned cores in any order they preferred ([Figure 3](#)). However, each core
⁴²⁹ and all associateddebitage were collected before participants were allowed to start working on a
⁴³⁰ new core, so it was not possible to partially work and then return to a particular core later. The
⁴³¹ order of cores used and associated hammerstones were recorded for each participant during the
⁴³² experiment.



Figure 3: Subjects selecting hammerstones.

⁴³³ In the uninstructed condition, a researcher (DS) sat with the participants and made stone tools
⁴³⁴ but remained silent and made no effort to facilitate learning (e.g., through gesture, modified
⁴³⁵ performance, facial expression, attention direction, or verbal instruction). Over the 2 hours, the
⁴³⁶ researcher completely reduced four cores (one every ~30 minutes). We did not restrict participants

437 from talking to each other, creating an unnatural and potentially stressful social context that
438 might affect learning. Participants were asked to avoid any form of communication about the
439 tool-making task specifically, and they complied with this request. Participants in this condition
440 thus had the opportunity to observe tool-making (**Figure 4**) by an expert and/or by other learners,
441 should they choose to do so, but received no intentional instruction.



Figure 4: Subjects observing expert knapper in the uninstructed condition.

442 In the instructed condition, there were no restrictions on participant interaction, and the re-
443 searcher engaged in direct active teaching ([Kline, 2015](#)) of tool-making techniques through verbal
444 instruction, demonstration, gesture, and shaping of behavior. The instructor has a moderate level
445 of experience teaching basic knapping skills to students in undergraduate archaeology classes
446 and participants in previous knapping research (e.g., [Stout et al., 2011](#)). The pedagogical strategy
447 employed was based on the instructor's own learning experiences and theoretical interpreta-
448 tions (e.g., [Pargeter et al., 2020](#)) and focused on coaching participants in effective body postures,
449 movement patterns, and grips and the assessment of viable core morphology.

450 **2.5 Lithic Analysis**

451 **Table 1** lists details for the study's various lithic attributes. We weighed and measured all the fin-
452 ished cores (L, W, T). Delta weight was calculated as (nodule start weight-core end weight)/nodule
453 start weight. All detached pieces (DPs) were collected and weighed. We did not sort DPs into
454 types (e.g., whole flakes, fragments), which would have greatly increased processing time. It is not
455 clear that such distinctions add relevant information regarding utility/desirability beyond that
456 supplied by metrics ([Stout et al., 2019](#)). All DPs larger than 40mm in maximum dimension were
457 photographed and measured. It is conventional in Early Stone Age lithic analysis to employ a 20
458 mm cut-off. We selected a higher threshold for both pragmatic (analysis time) and theoretical
459 reasons. Flake use experiments have shown that flakes weighing less than 5–10 g or with a surface
460 area below 7–10 cm² ([Prasciunas, 2007](#)) or with a maximum dimension <50–60 mm ([Key & Lycett,
461 2014](#)) become markedly inefficient for basic cutting tasks. Similarly, data from Oldowan replica-
462 tion experiments ([Stout et al., 2019](#)) show that the utility index (flake cutting edge/flake mass^{1/3}) *
463 (1 - exp[-0.31 * (flake maximum dimension - 1.81)]) developed by Morgan et al. (2015) falls off
464 rapidly below 40mm maximum dimension (Mean Utility < 40mm = 0.508; >=40mm = 0.946; t=
465 11.99, df = 707, p < 0.000). By including weight in our cut-off criteria, we also avoid skewing
466 the flake shape distribution by selectively retaining long, thin pieces (i.e., MD > 40, weight < 5g)
467 while discarding rounder pieces of similar (or greater) weight and area. We recorded the absolute
468 number of above-threshold DPs produced per core and calculated a mass-based proportion of
469 larger DPs by dividing the combined weight of DPs >40mm in maximum dimension and 5g in
470 weight by the weight of all DPs. Higher values show cores from which proportionally more of the
471 detached mass comprises large pieces.

Table 1: Overview of lithic variables and recording methods used in the study.

Variable	Variable.class	Definition	Recording.method
Delta mass	Ratio	Final core mass as a percentage of original nodule mass. Higher values show more completely flaked nodules.	Digital scale
Mass of large detached pieces / total debitage mass	Ratio	Combined mass of all detached pieces >40mm in maximum dimension and 5g in mass divided by the mass of all material detached from a specific core. Higher values show cores with more detached pieces per knapped material.	Calipers, counts, digital scale
Length	Linear measurement	Measurement along each core and flake's longest axis in mm.	2D photogrammetry
Width	Linear measurement	Measurement taken at 10% increments along a detached piece's longest axis starting at the detached piece's tip (in mm). Same measurement taken orthogonal to a core's length.	2D photogrammetry
Thickness	Linear measurement	Measurement taken at the maximum point in a detached piece's profile and orthogonal to width on the cores.	Calipers

472 We defined DP length as the longest axis and width as the maximum dimension orthogonal to
 473 length for measurement (see supplementary Table 2). Our thickness measurement follows the
 474 maximum dimension orthogonal to the plane formed by L and W, which we measured using
 475 calipers. L, W, and plan-view area measurements were taken from photographs captured using
 476 a Canon Rebel T3i fitted with a 60 mm macro lens and attached to a photographic stand with
 477 adjustable upper and lower light fittings. The camera was positioned directly above the flakes
 478 and kept at a constant height. We arranged DPs irrespective of any technological features so that
 479 the longest axis was vertical and placed the broader end toward the bottom of the photograph.

- 480 • Photographs were post-processed using Equalight software to adjust for lens and lighting
 481 falloff resulting from bending light through a lens and its aperture, affecting measurements
 482 taken from photos. We shot each image with a scale then used this scale to rectify the
 483 photograph's pixel scale to a real-world measurement scale in Adobe Photoshop. Images
 484 were converted to binary black and white format, and silhouettes of the tools were extracted
 485 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

487 length starting at the tip of each DP (i.e., 10% width = tip, 90% width = base). We used the
488 built-in ImageJ tool to measure DP area.

489 **2.6 Statistical Analyses**

490 We adopted an information-theoretic approach to evaluate the association between psychometric-
491 ric, motor-skill, training measures, and technological outcomes ([Burnham & Anderson, 2002](#)).
492 Information-theoretic techniques provide methods for model selection using all possible com-
493 binations of variables while avoiding problems associated with significance-threshold stepwise
494 selection. We used the small sample corrected Akaike information criterion (AICc) to rate each
495 possible combination of predictors on the balance between the goodness of fit (likelihood of
496 the data given the model) and parsimony (number of parameters). The AICc consists of the
497 log-likelihood (i.e., how well does the model fit the data?) and a penalty term for the number of
498 parameters that must be estimated in the model (i.e., how parsimonious is the model?), with a
499 correction for small sample sizes (AICc converges to the standard AIC at large samples). A lower
500 AICc indicates a more generalizable model, and we use it to compare and rank various possible
501 models. Each analysis begins with a complete model that includes all predictors of interest. All
502 possible combinations of predictors are then fitted, and the resulting models are ranked and
503 weighted based on their AICc. We chose the “best” model because it had the lowest AICc score.
504 Continuous predictors were centered such that zero represents the sample average, and units are
505 standard deviations. The full model was fitted with the *lm* function in R 4.0.3, and the *glmulti*
506 package ([Bartoń 2015](#)) was used for multi-modal selection and model comparison.

507 **3 Results**

508 Following a recent protocol to enhance the reproducibility and data transparency of archaeo-
509 logical research ([Marwick, 2017](#)), detailed results of all analyses and assessments of the data
510 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
511 flaking performance and individual differences. We were particularly interested in 1) group-level
512 effects of experimental conditions, 2) individual differences in aptitude and learning, and 3)
513 potential interactions between learning conditions and individual differences. To address these
514

515 questions, we employed data reduction (Principal Component Analysis) to derive two summary
516 metrics of flaking performance, compared these factors across the two experimental condi-
517 tions, and built multivariate models examining the relations between our various psychometric
518 measures, subject's motor skill scores, and our two lithic performance factors.

519 **3.1 Principal Component analyses**

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

522 **3.1.1 Detached Piece size and shape**

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

528 The analysis yielded three factors explaining a total of 90% of the variance for the entire 11
529 measurement variable set (Table 2). Factor 1 tracks size with higher scores indicating larger
530 pieces since all 11 measures positively load this factor. Factor 2's loadings track increasing
531 thickness and length but not width. As factor 2 scores increase, pieces thus get thicker, longer, and
532 relatively narrower, resembling irregular splinters. Factor 3 tracks increasing proximal breadth
533 and decreasing thickness. As factor 3 scores go up, pieces get thinner overall and broader at the
534 base. Factor 3, therefore, tracks pieces with a more typical flat, flake-like morphology having a
535 thin cross-section and tapering shape. We used these three shape coordinates to approximate DP
536 size and shape in the project's flaking performance factor analysis.

Table 2: Lithic size/shape PCA factor loadings.

Variable	Factor.1	Factor.2	Factor.3
Variance %	78	6.2	5.4
Interpretation	Size	Relative thickness and elongation	Relative thinness and proximal breadth
Flake thickness	0.68	0.58	-0.35
Flake length	0.81	0.31	-0.11
Width-10% (tip)	0.82	-0.31	-0.16
Width-20%	0.91	-0.27	-0.19
Width-30%	0.94	-0.2	-0.16
Width-40%	0.95	-0.07	-0.05
Width-50%	0.96	-0.06	-0.03
Width-60%	0.94	-0.06	0.01
Width-70%	0.95	0.05	0.18
Width-80%	0.92	0.11	0.34
Width-90% (base)	0.8	0.12	0.47

537 3.1.2 Lithic flaking performance measures

538 We conducted a second principal component analysis better to understand the relationship
 539 between our various lithic performance measurements and reduce data dimensionality. This
 540 analysis examined the study's six lithic performance measures (count of large pieces [$>40\text{mm}$
 541 and 5g], mass of large pieces relative to total detached mass, core delta mass, and the three flake
 542 shape factors) (see supplementary Table 3 for details). These measures were summarized for each
 543 core, and individual factor scores were calculated from these core-specific measures. Bartlett's
 544 Test of Sphericity was significant ($\chi^2(6) = 3185$, $p < .01$) indicating that the set of variables are at
 545 least adequately related for factor analysis.

546 The analysis yielded two factors explaining 56% of the variance for the entire set of variables
 547 (**Table 3**). Factor 1 (hereafter "Quantity") explains 28.7% of the variance and tracks flaking quantity
 548 due to high positive loadings on large DP count and mass ratio and core delta mass. Performance
 549 factor 2 (hereafter "Quality") covers 27% of the sample variance and measures flaking quality
 550 as reflected in high positive loadings on Shape Factors 1 (size) and 3 (thin, "flake-like" shape)
 551 and a negative loading on Shape Factor 2 ("splinter-like" thickness and elongation). High scores
 552 on Quality thus reflect the making of larger, relatively thinner, and more typically flake-shaped
 553 vs. splinter-shaped DPs.

Table 3: Overview of principal components results on the six flake performance measures. DP = detached piece.

Variable	Factor.1	Factor.2
Variance %	28.7	27
Interpretation	Flaking quantity (higher values = greater quantity detached)	Flaking quality (higher values = larger, more "flake-shaped" pieces)
Mass of DPs > 40mm and 5g : total mass of all DPs	0.61	-0.48
Number of DPs > 40mm and 5g	0.84	0.4
Core delta mass	0.78	-0.23
Shape PC factor 1 (size)	0.11	0.63
Shape PC factor 2 (thickness and elongation)	0.01	-0.71
Shape PC factor 3 (thinness and base breadth)	0.13	0.55

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

561 3.2 Relationships between Performance Measures

562 This approach also allowed us to compare the relationship between Total Productivity, Quantity,
 563 and Quality across our two experimental groups (**Figures 5 and 6**). As might be expected, we
 564 found that per-core Quantity and Total Productivity are positively correlated in both groups
 565 (**Figure 5a**), although this relationship is twice as strong in the instructed ($F[1, 9] = 33, p < 0.01$,
 566 Adj. $R^2 = 0.8$) compared to uninstructed ($F[1, 8] = 8, p = 0.02$, Adj. $R^2 = 0.4$) group. Interestingly,
 567 we also found evidence of a negative correlation between Total Productivity and Quality in the
 568 uninstructed group ($F[1, 8] = 28, p = <0.01$, Adj. $R^2 = 0.7$), but no relation in the instructed group
 569 (**Figure 5b**). A qualitatively similar trend concerning Quality vs. Quantity (**Figure 5c**) did not
 570 achieve significance ($F[1, 17] = 0.6, p = 0.2$).

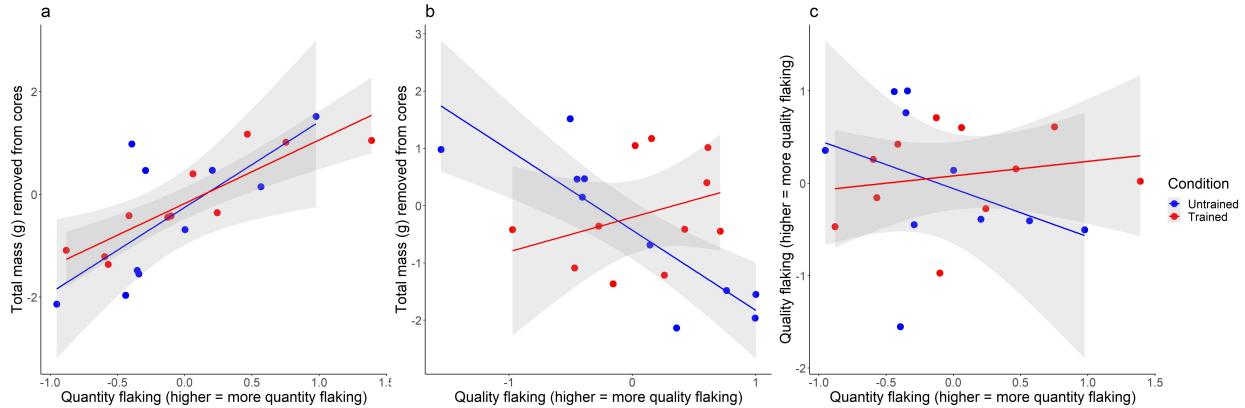


Figure 5: Relationships between flaking quantity, quality, productivity, and the two training conditions. Each dot represents a participant, colors represent training conditions.

571 Thus, it appears that instructed participants achieved higher Total Productivity by increasing
 572 average flaking Quantity across cores and without sacrificing Quality. In contrast, uninstructed
 573 participants found other ways to vary Total Productivity (e.g., number of cores knapped rather
 574 than Quantity per core, see **Table 5**, **Figure 8 & 9**) and generally increased productivity at the
 575 expense of Quality. Experimental artifacts illustrating these trade-offs are presented in **Figure 6**.

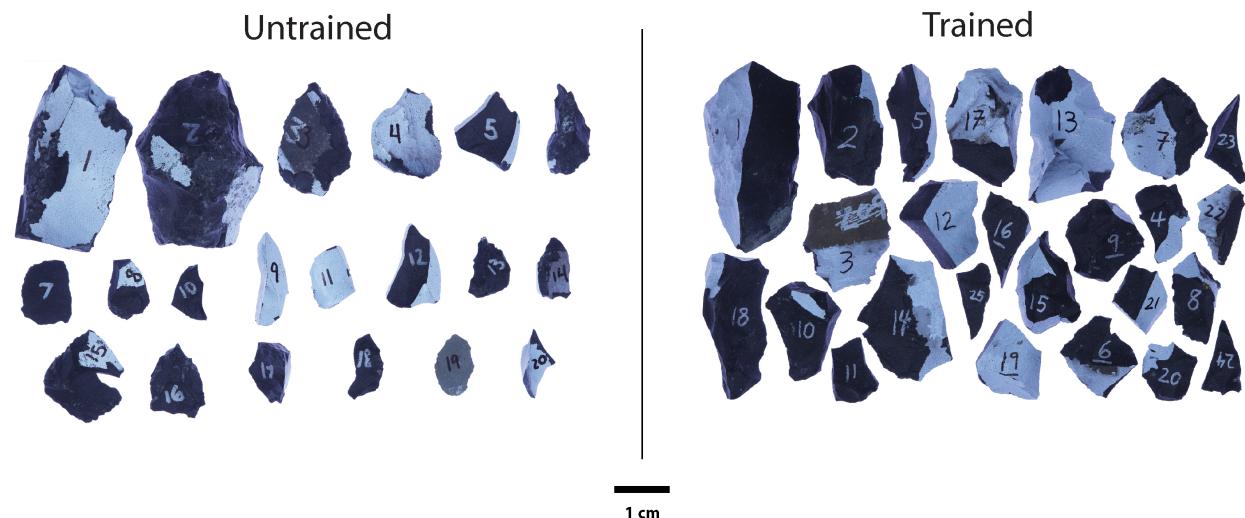


Figure 6: Comparison of untrained and trained detached pieces illustrating the ability of Trained participants to maximize Quality and Productivity at the same time.

576 3.3 Do trained, untrained, and expert knappers perform differently?

577 Here we compare our flaking outcomes (DP size/shape and flaking performance factors) between
 578 the instructed and uninstructed groups. We include the expert demonstrator/instructor as a
 579 performance benchmark.

580 **Table 4** summarizes the results of ANOVA tests group-level difference in central tendency on
581 various performance measures. We found no significant differences between the instructed and
582 uninstructed groups on our flaking Quantity and Quality factors. In contrast, three-way flake
583 size and shape comparisons between our expert knapper and the two novice groups showed
584 that the expert knapper made significantly more large flakes (effect size = 0.14). The expert also
585 had a significantly higher core delta mass signal than either of the novice groups (effect size =
586 0.26) and left significantly smaller finished cores (effect size = 0.27) (**Figure 7**). All three of these
587 results show either medium or large effect sizes. The instructed group's data distributions tended
588 towards the expert sample in all three comparisons, although they did not differ significantly
589 from the uninstructed group (**Figure 7**). We also observed a significant difference in shape factor
590 2 (splinters) driven by the expert's lower values but with a very small (< 0.01) effect size. These
591 results show that mean core reduction intensity and large flake production rates distinguish
592 expert and novice performance. In contrast, novices in experimental groups produced similar
593 mean detached piece sizes and shapes as the expert trainer.

Table 4: Summary group-wise comparison statistics contrasting flake performance metrics by experiment condition and with the expert instructor. All comparisons are three-way except for the two skill PC factors and the total mass removed, which are two-way comparisons between the subject population as these factors were not applied to the expert data sample. Eta2 = ANOVA effect size (>0.1 = medium effect size, >0.2 = large effect size). * = non-parametric permuted p-values to account for unequal sample variances. SS = sum of squared differences, df = degrees of freedom, MS = mean squared difference.

Variable	Parameter	SS	df	MS	F	p	Eta2
Skill PC factor 1 (quantity flaking)*	Group	0.03	1	0.03	0.01	0.8	na
	Residuals	40.49	21	1.9	NA		
Skill PC factor 2 (quality flaking)*	Group	2.03	1	na	1.20	0.2	na
	Residuals	33.8	21	na	NA		
Flake factor 1 (size)	Group	22.3	2	11.1	1.30	0.2	na
	Residuals	36741.35	4328	8.5	NA		
Flake factor 2 (relative thickness)	Group	11	2	5.46	8.30	<0.01	<0.01
	Residuals	2842.2	4328	0.66	NA		
Flake factor 3 (basal relative to tip width)*	Group	0.04	2	na	0.90	0.9	na
	Residuals	2404	4328	na	NA		
Flakes > 40mm and 5g*	Group	4967	2	na	23.50	<0.01	0.14
	Residuals	19496	185	na	NA		
Mass of flakes : flaked mass	Group	0.06	2	0.03	2.30	0.1	na
	Residuals	2.4	183	0.01	NA		
Core delta mass*	Group	0.6	2	na	13.70	<0.01	0.26
	Residuals	4.4	184	na	NA		
Total cores used	Group	8.5	1	8.5	2.20	0.1	na
	Residuals	79.4	21	3.7	NA		
Total flaked mass	Group	22623898.1	1	2262389	0.50	0.4	na
	Residuals	39348298797	21	4451571	NA		

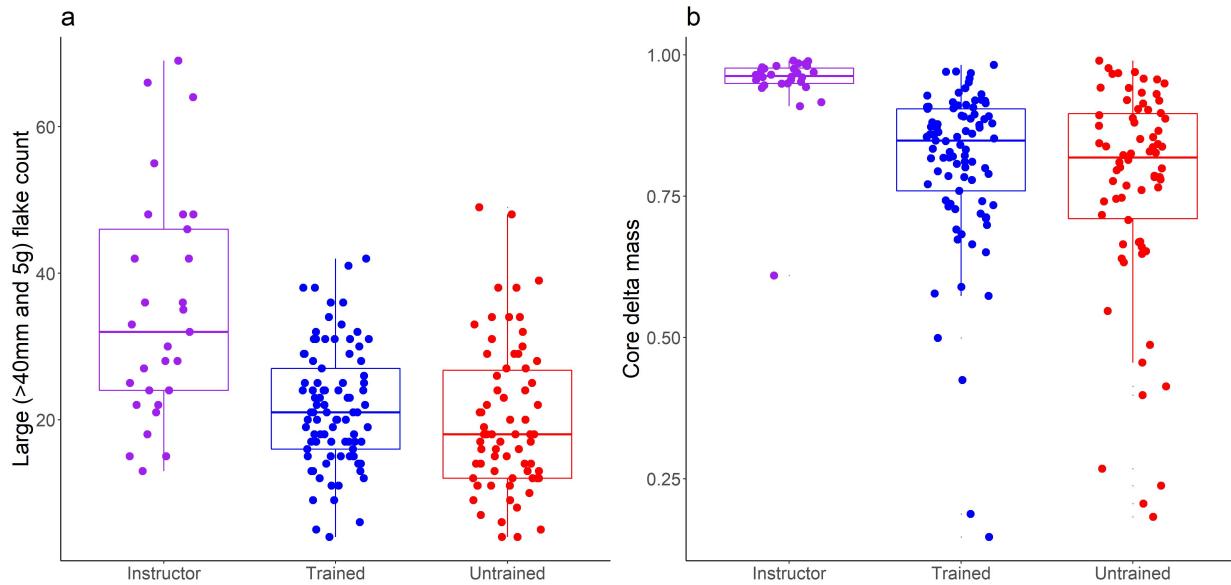


Figure 7: Results showing significant differences between the instructor (expert) and novice flaking performance.

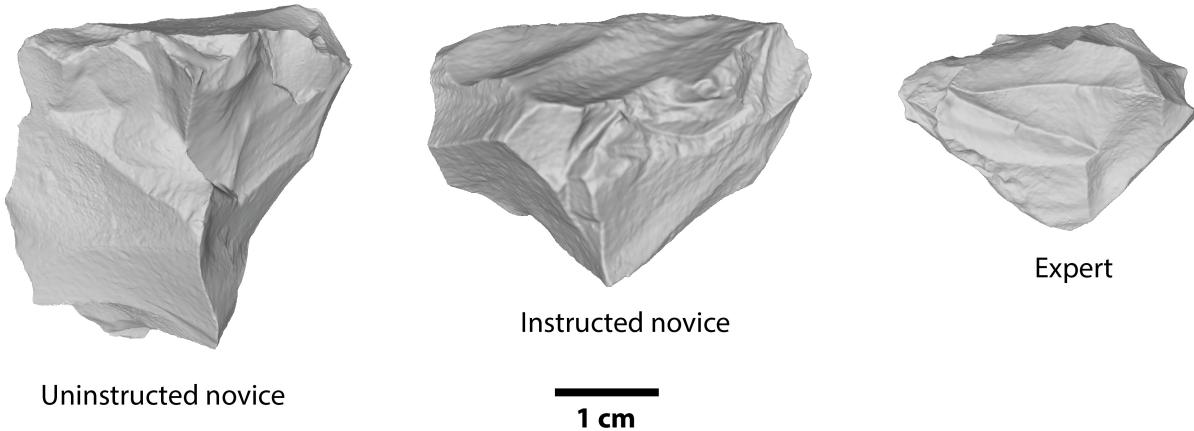


Figure 8: Comparison of uninstructed, instructed, and expert cores.

594 While we did not find significant differences in central tendency between our two experimental
 595 groups, results (**Figure 7**) did indicate lower variance in the instructed group. To test whether
 596 training reduced variability in performance outcomes between subjects, we compared vari-
 597 ance metrics between the instructed and uninstructed individuals using the F-test on either
 598 core-averaged or flake-specific variances. **Table 5** and **Figure 9** present the results from these
 599 comparisons showing significant variance differences predominantly in flaking Quality, number
 600 of large DPs, core delta mass, and the total amount of flaked mass). In most instances, variance
 601 in the uninstructed group exceeds that of instructed individuals by 1.5 to 4.7 times. The most
 602 salient effect of instruction was thus not to shift mean performance but to reduce variability by
 603 eliminating the skew (generally toward poorer outcomes) seen in the uninstructed group (**Figure**
 604 **9**), rather than to shift the mean.

Table 5: F-test results comparing variances between trained and untrained subjects across our various flaking performance metrics.

Variable	df	F	95% CI	p	Var.ratio
Skill PC factor 1 (quantity flaking)	21	1.1	0.3 - 4.2	0.8	1.1
Skill PC factor 2 (quality flaking)	21	2.1	1.3 - 3.3	<0.01	2.1
Flake factor 1 (size)	1379	1.1	0.9 - 1.2	0.05	1.1
Flake factor 2 (relative thickness)	1379	0.9	0.8 - 1.0	0.6	0.9
Flake factor 3 (basal relative to tip width)	1379	1.0	0.9 - 1.2	0.06	1.0
Detached pieces > 40mm and 5g	69	1.5	0.9 - 2.4	0.05	1.5
Mass of detached pieces : flaked mass	69	1.3	0.8 - 2.1	0.1	1.3
Core delta mass	69	1.7	1.1 - 2.7	0.01	1.7

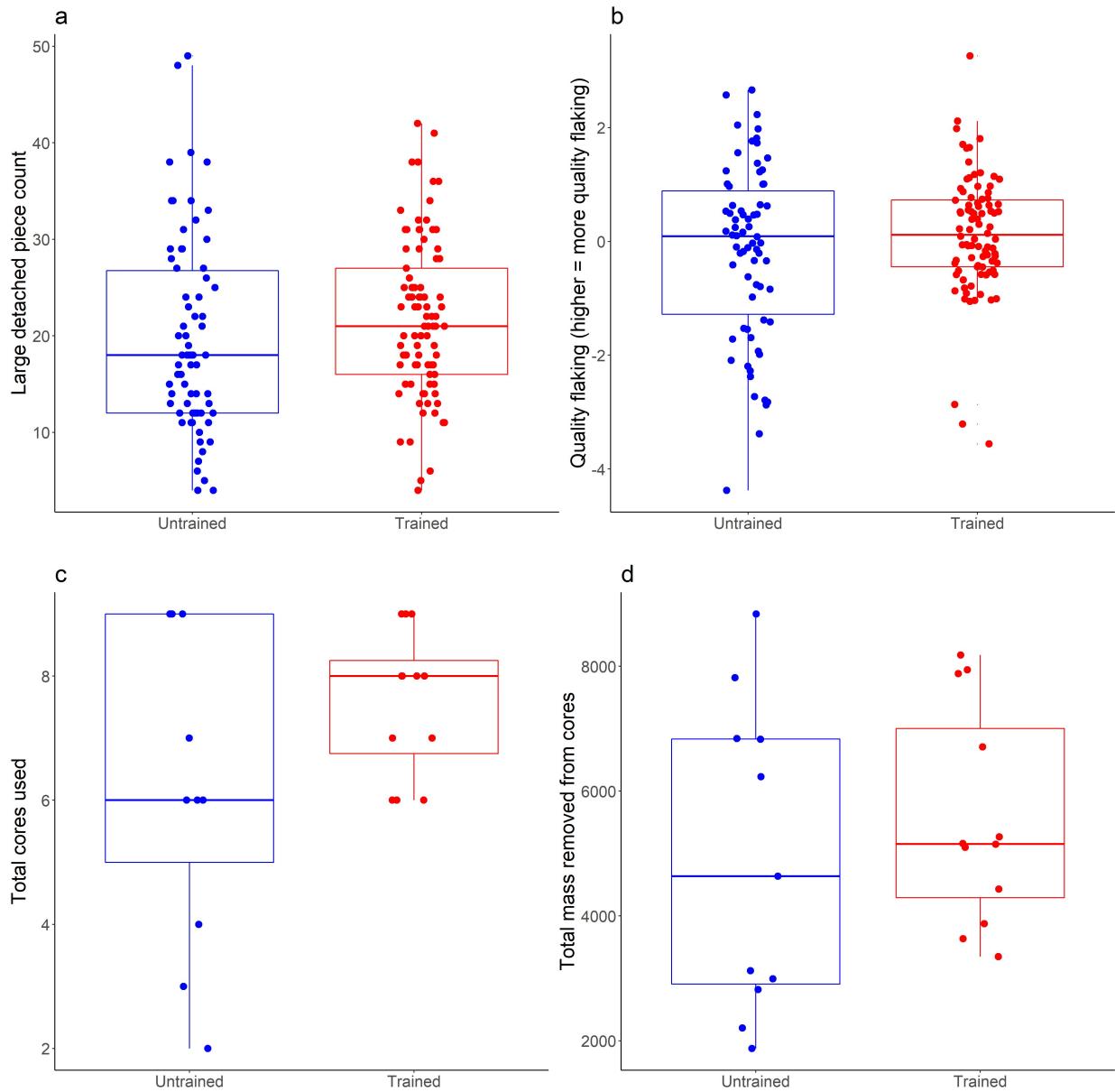


Figure 9: Variance comparisons between instructed and uninstructed individuals across various flaking performance metrics. Note the lower variance of the trained group in all the plots.

605 3.4 Does performance change over time?

606 In addition to comparing overall performance during the two-hour experiment, we also wanted
 607 to determine if groups or individuals differed in learning (i.e., performance change) over the
 608 period. We calculated the learning stage as the ordinal number of each core out of the total
 609 number knapped by each subject (i.e., core 2 of 4 or 4 of 8, both equal 50% complete) for these
 610 analyses. These relative core use-order percentages were then binned into 20 percent brackets

for core-order and group-level comparisons. We tracked flaking outcomes 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 6 shows no significant training effects across the two performance measures either as grouped data or between individuals (**Figure 10** and **Figure 11**). This result demonstrated that flaking outcomes did not change dramatically across the study interval. An inspection of individual learning curves confirms this lack of significant learning effects (**Figure 11**). The main training effect was core starting mass (with a substantial effect size = 0.25). On average, core starting masses start low and increase, showing that participants selected smaller nodules first. As participants depleted the smaller nodules in their allotment, they knapped larger, less preferred nodules. This preference for smaller cores is somewhat less pronounced in the uninstructed group, as indicated by a small main effect of learning condition and generally higher starting nodule masses for the uninstructed group (**Figure 11**).

Table 6: Summary group-wise comparison statistics contrasting flake performance metrics by experiment condition and relative core order. All comparisons are three-way. Eta2 = ANOVA effect size (>0.1 = medium effect size, >0.2 = large effect size). SS = sum of squared differences, df = degrees of freedom, MS = mean squared difference.

Variable	Parameter	SS	df	MS	F	p	Eta2
Skill factor 1 (quantity flaking)	Condition	1.80	1	1.80	1.00	0.3	na
	Relative core order	4.90	4	1.20	0.70	0.6	na
	Condition*Relative core order	1.00	4	0.20	0.10	1	na
	Residuals	263.00	148	1.80	NA		
Skill factor 2 (quality flaking)	Condition	3.40	1	3.40	2.20	0.1	na
	Relative core order	9.00	4	2.20	1.40	0.2	na
	Condition*Relative core order	14.80	4	3.70	2.40	0.1	na
	Residuals	231.70	148	1.60	NA		
Core starting mass	Condition	0.30	1	0.30	4.72	0.03	0.03
	Relative core order	3.27	4	0.82	12.70	<0.01	0.25
	Condition*Relative core order	0.30	4	0.07	1.17	0.32	0.03
	Residuals	9.58	149	0.06	NA		

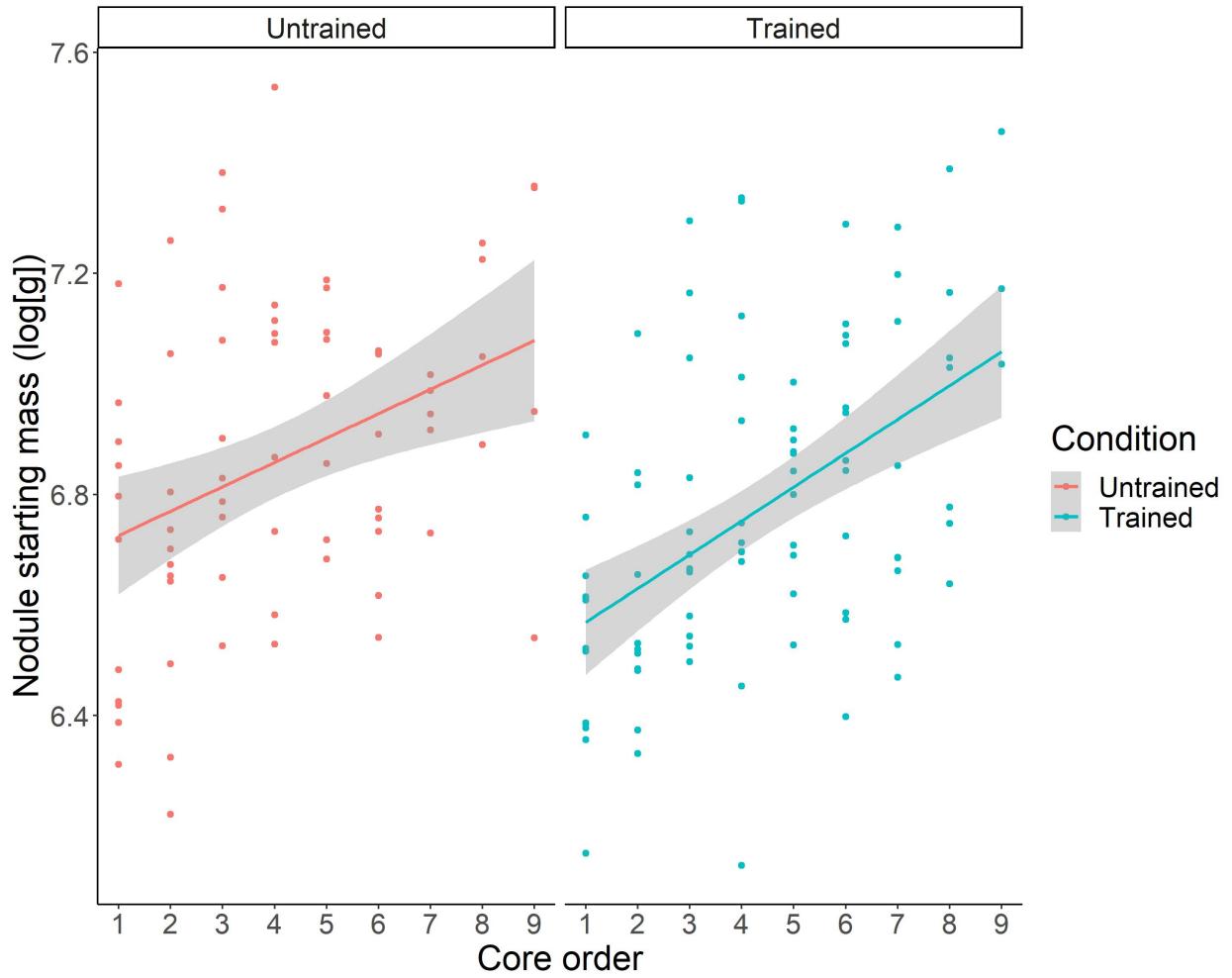


Figure 10: Comparisons of nodule starting mass across the study period by training condition. Results show a significant relationship between nodule starting mass and length of time in the study.

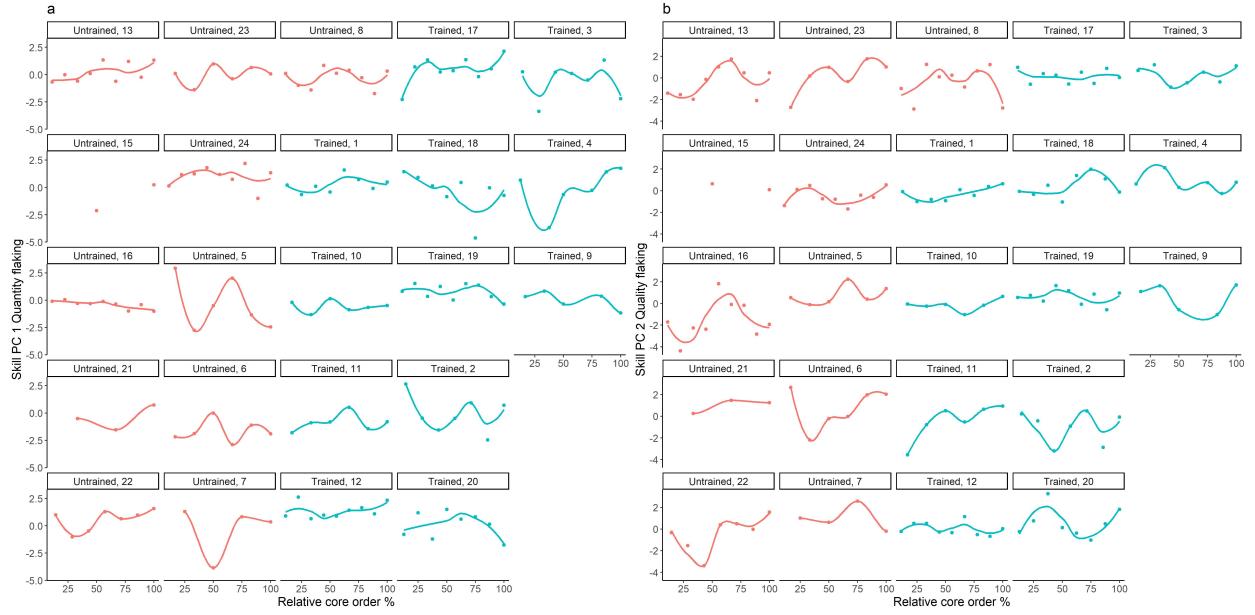


Figure 11: Individual learning curves derived from each subject's relative core use order and the two flake performance factors.

624 3.5 Do individual differences in motor skill and psychometric measures predict 625 flaking performance?

626 One of the experiment's primary goals was to test if individual perceptual-motor and cognitive
627 variation measures predict success in stone flaking across different training conditions. We
628 built three multivariate models to address this goal examining the relations between training
629 conditions, individual difference measures, and our three lithic performance measures (overall
630 productivity and average per-core Quantity and Quality). These models enabled us to deter-
631 mine which psychometric and motor skill factors are better predictors of a participant's flaking
632 performance in the study.

633 We considered all possible interactions between five individual difference measures, core size,
634 training condition, and the three performance measures, with each subject providing one data
635 point. Each model's continuous predictors (highest n-back level, Raven's Progressive Matrix
636 score, BEAST score, starting nodule mass, Fitts score, and grip strength) were centered such that
637 zero represents the sample average, and units are standard deviations. Our two motor skill and
638 strength measures (grip strength and Fitt's performance scores) are also strongly correlated (F
639 $[1,19] = 15, p < 0.01, R^2 = 0.41$). However, these two measures track complementary components
640 of athleticism (strength vs. speed/accuracy trade-offs), so we decided to include both in the model

641 selection process.

642 All three models follow the same complete model statement as follows:

643 *Flaking performance variable ~ Training condition + Highest n-back level + Raven's Progressive*

644 *Matrix score + BEAST score + Fitt's score + Grip strength*

645 For our two per-core performance factors (Quantity and Quality), it is also relevant to consider how

646 individual core features may have affected performance. We found no evidence of individual or

647 group level practice effects over the two hours, so we did not include core order in the models. We

648 did, however, find that subjects selected progressively larger nodules throughout the experiment.

649 It is thus essential to understand whether nodule variability had any impact on our flaking results.

650 Because starting nodule size (mass) and shape were strongly correlated ($F [1,157] = 186$, $p <$

651 0.01 , $R^2 = 0.54$), we included nodule mass as a covariate to control for any variance in flaking

652 performance that nodule differences may drive.

653 **3.5.1 Model 1: Individual differences and overall productivity**

654 Our first model examined variance in overall flaking productivity measured by each subject's

655 combined flaked mass (nodule starting mass-core final mass). This productivity metric provides

656 a basic measure of variation in individuals' success detaching pieces and reducing cores from a

657 standardized (see Methods) raw material supply. From the same candidate pool size of 55893

658 possible multivariate models, the best performing model returned an AICc value of -18 (Average

659 AIC = -13). This model comprised the following statement with two main and three interaction

660 effects:

661 *Total flaked mass ~ Training condition + Grip strength + RPM × Highest n-back level + Fitts*

662 *score × BEAST score + Grip strength × RPM*

663 This model explains a statistically significant and substantial proportion of variance in flaking

664 productivity ($R^2 = 0.84$, $F (6, 14) = 12.7$, $p < 0.01$, adj. $R^2 = 0.77$). A model residuals normality

665 test shows no significant differences with the normal distribution ($p = 0.72$) indicating that this

666 relationship is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity ($BP = 2$, df

667 = 6, $p = 0.8$).

Table 7: Summary and estimates for the overall flaking productivity model.

Covariate	Estimate	95% CI	t(12)	p
(Intercept)	-0.7	[-1.06, -0.33]	-4.11	0
Condition[Trained]*BEAST	-1.3	[-1.86, -0.73]	-4.91	<0.01
Condition[Trained]*Highest nBack	-1.3	[-1.82, -0.75]	-5.15	<0.01
Grip strength	0.9	[0.59, 1.21]	6.24	<0.01
Training condition	0.7	[0.21, 1.22]	3.02	0.01
Visuospatial nBack	0.7	[0.29, 1.17]	3.57	<0.01
BEAST	0.6	[0.17, 0.96]	3.08	0.01

668 **Table 7** presents this model's coefficients and summary outputs, wherein baseline refers to the
 669 uninstructed condition with all continuous predictors at the sample average. The parameter
 670 estimates for the continuous predictors reflect the expected change in utility for one standard
 671 deviation change in the predictor variable. We found significant ($p < 0.05$) and substantial (Stan-
 672 dardized Estimate ≥ 0.50 , i.e. a 50% change in variable) main effects of Grip Strength, Visuospatial
 673 nBack, and BEAST. The main effect of Grip Strength (**Figure 12**), irrespective of learning condition,
 674 indicates the basic importance of strength in generating higher production rates among naive
 675 knappers, at least when we do not consider efficiency and quality.

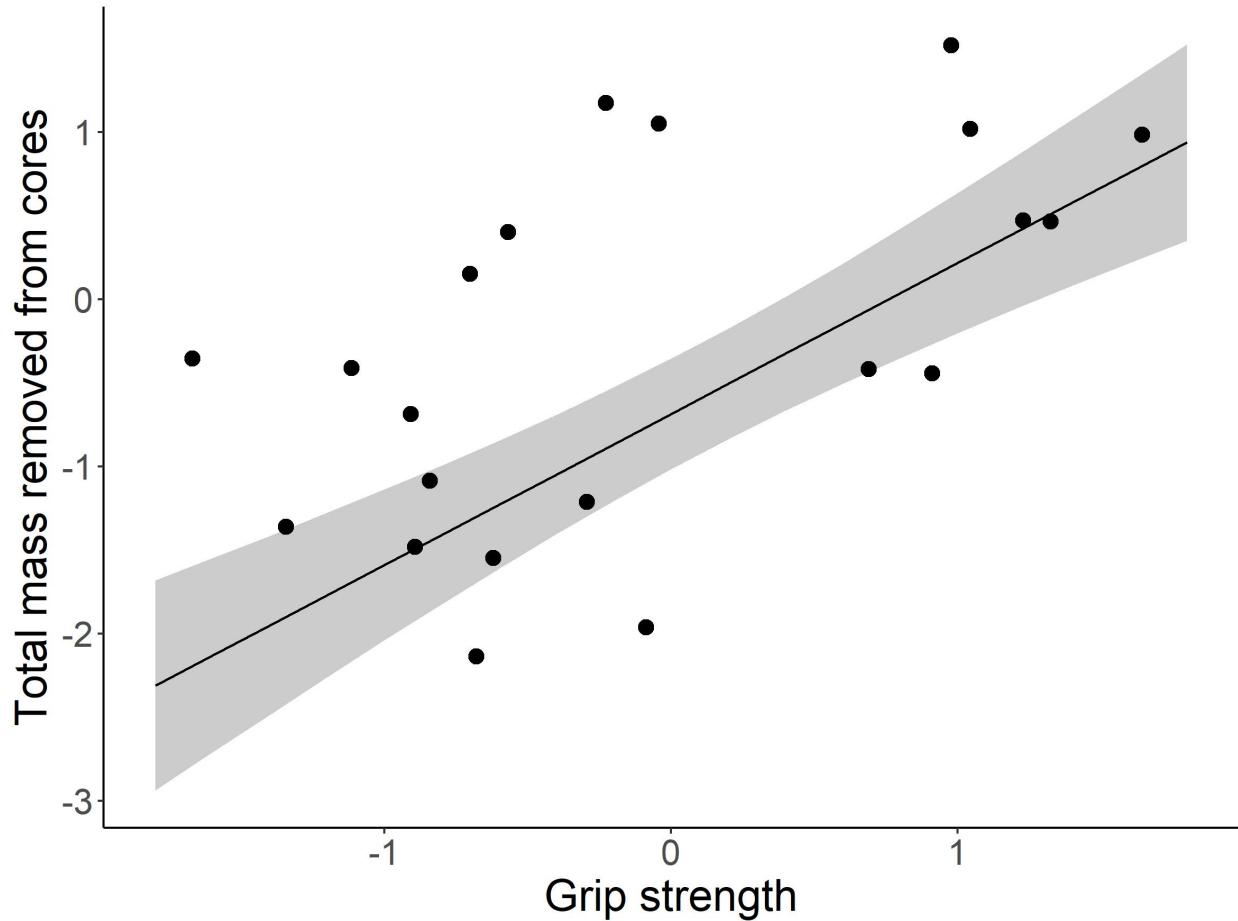


Figure 12: Significant main effect of grip strength on overall flaking productivity.

676 Effects of visuospatial working memory capacity and social information use are more complicated,
 677 as indicated by strong interactions with learning condition ([Figure 13](#)). In each case, higher scores
 678 were associated with better performance in the uninstructed group but *worse* performance in
 679 the instructed group. Positive effects in the uninstructed group were as expected, given the
 680 hypothesized importance of spatial cognition ([Coolidge & Wynn, 2005](#)) and social learning
 681 ([Morgan et al., 2015](#)) in acquiring knapping skills. Negative effects in the instructed group are
 682 unexpected but presumably reflect differences in learning strategies adopted under instruction.

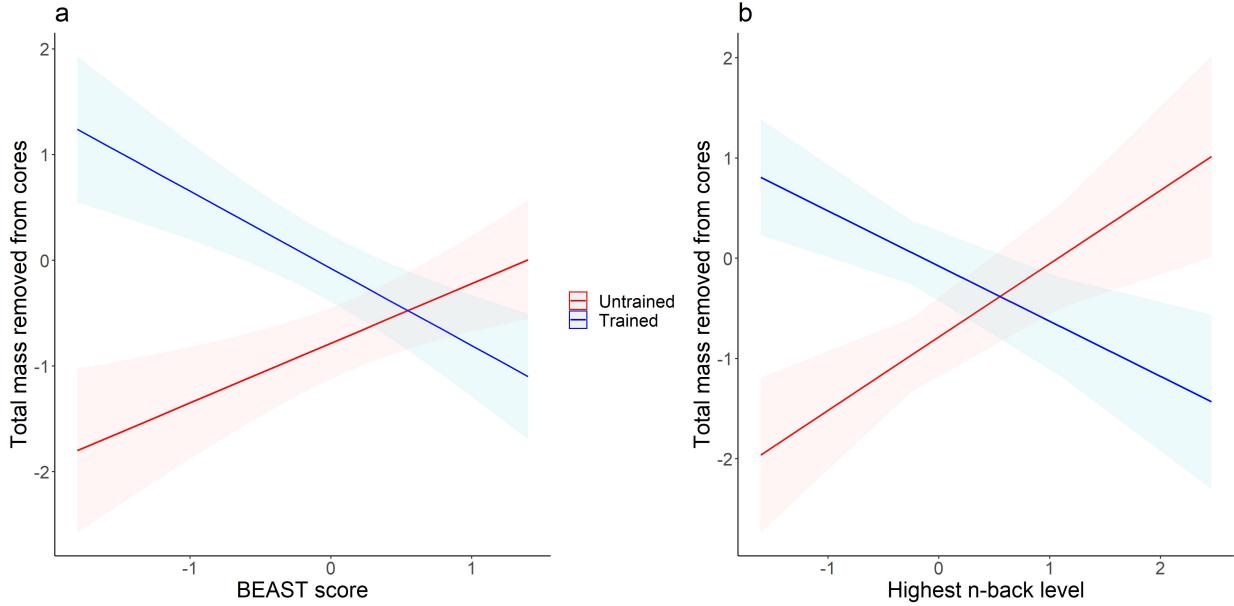


Figure 13: Two significant interaction effects of training condition, social information, and visuo-spatial working memory on overall flaking productivity.

683 3.5.2 Model 2: Individual differences and quality flaking

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

689 *Quantity ~ Highest n-back level + BEAST score + Fitt's score + Grip strength + Training
 690 condition × Highest n-back level + Training condition × BEAST score + Training condition × Grip
 691 strength + Nodule mass (as control)*

692 This model explains a statistically significant and substantial proportion of variance in quantity
 693 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
 694 significant differences with the normal distribution ($p = 0.38$) indicating that this relationship
 695 (as required) is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (whether
 696 variance for all observations in our data set are the same) ($BP = 4.4$, $df = 8$, $p = 0.8$).

697 **Table 8** presents this model's coefficients and summary outputs, following the same format as
 698 **Table 6** and **Table 7**.

Table 8: Summary and estimates for Quantity flaking model.

Covariate	Estimate	95% CI	t(12)	p
(Intercept)	0.0	[-0.20, 0.24]	0.2	0.85
Training Condition[Trained]*BEAST score	-0.9	[-1.37, -0.42]	-4.1	<0.01
Training Condition[Trained]*Highest n-back level	-0.7	[-1.23, -0.26]	-3.3	<0.01
Training Condition[Trained]*Grip strength	0.7	[0.08, 1.27]	2.5	0.03
BEAST score	0.3	[-0.02, 0.66]	2.1	0.06
Fitts score	-0.3	[-0.59, 0.02]	-2.0	0.06
Highest n-back level	0.2	[-0.19, 0.58]	1.1	0.29
Grip Strength	-0.1	[-0.48, 0.29]	-0.6	0.6
Nodule mass	-0.1	[-0.35, 0.08]	-1.3	0.21

699 The Quantity model roughly paralleled results for Total Production, yielding substantial and
 700 significant interactions between training condition, n-back level, BEAST scores, and grip strength.
 701 As with Total Production, higher visuospatial n-back levels and BEAST scores were associated
 702 with lower Quantity scores in the instructed group but higher or unchanged DP Quantity in the
 703 uninstructed group (**Figure 14**).

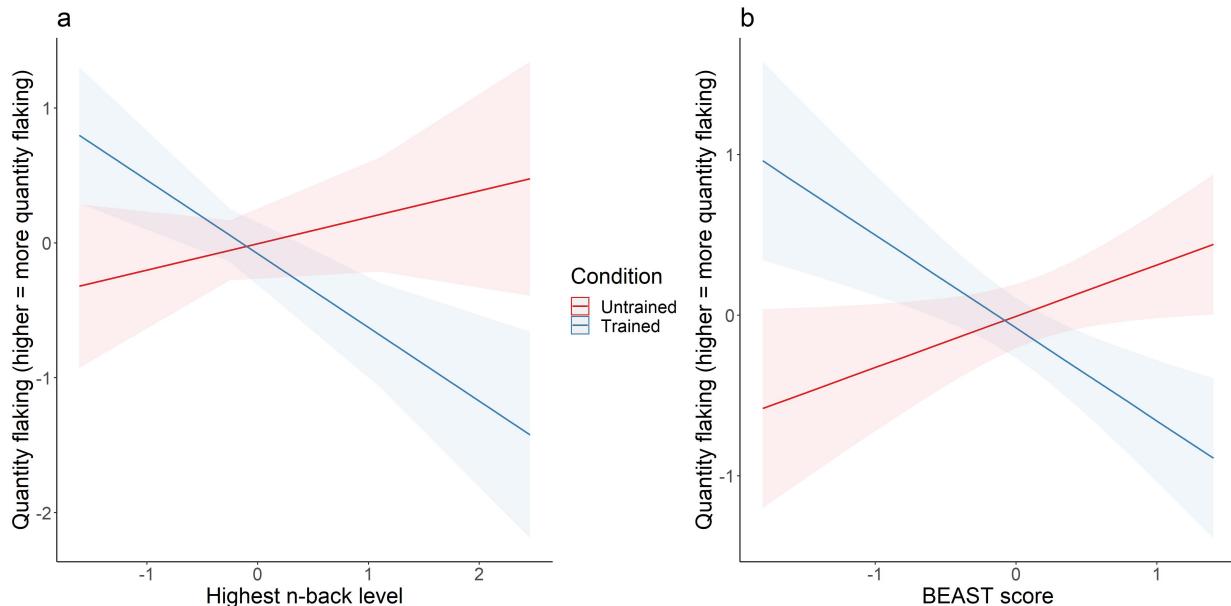


Figure 14: Two significant interaction effects of training condition, visuo-spatial working memory, and social information, on overall flaking quality.

704 Unlike Total Productivity, an interaction with learning condition mediated the effect of Grip
 705 Strength on per-core Quantity (**Figure 15**). Only instructed individuals translated Grip Strength
 706 into a more intense reduction of individual cores, including delta mass and the number and

⁷⁰⁷ proportion of larger pieces. Thus, high Grip Strength enabled individuals in both groups to
⁷⁰⁸ produce more debitage.

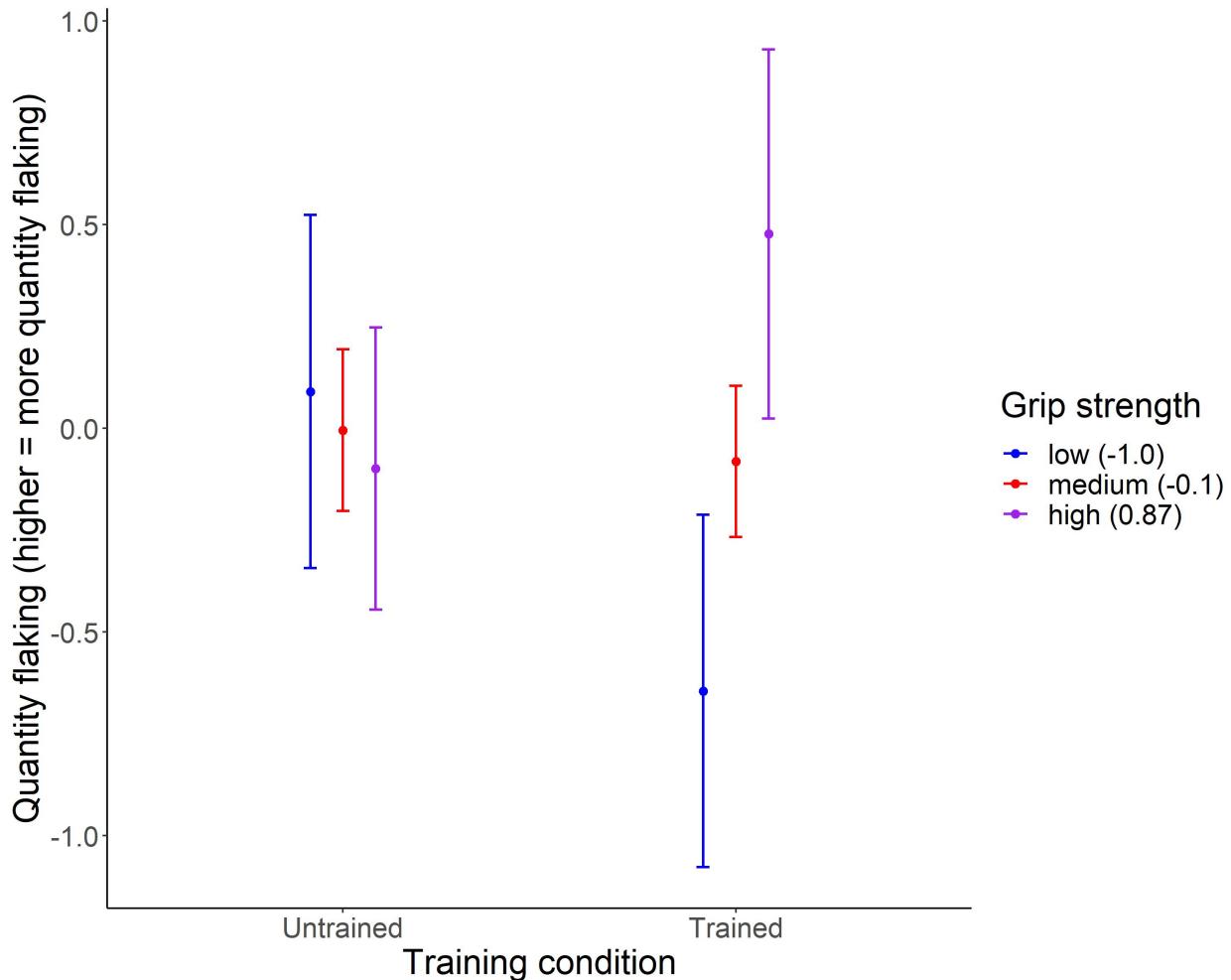


Figure 15: Three significant interaction effects of training condition, visuo-spatial working memory and social information, and training and grip strength use on quantity flaking.

⁷⁰⁹ 3.5.3 Model 3: Individual differences and quality flaking

⁷¹⁰ Our third model examining variance in Quality follows the same complete model statement we
⁷¹¹ used for Quantity. From the same candidate pool size of 55893 possible multivariate models, the
⁷¹² best performing model returned an AICc value of 32 (Average AIC = 39). This model comprised
⁷¹³ the following statement with three main and four interaction effects:

⁷¹⁴ *Quality flaking ~ Highest n-back level + Fitt's score + Grip strength + Fitt's score × BEAST score + Grip*
⁷¹⁵ *strength × BEAST score + Grip strength × Fitt's score + Training condition × Grip strength + Nodule*
⁷¹⁶ *mass (as control)*

717 This model explains a statistically significant and substantial proportion of variance in Quality
 718 ($R^2 = 0.75$, $F(8, 12) = 4.6$, $p < 0.01$, adj. $R^2 = 0.6$) in the absence of any main training effects. A
 719 model residuals normality test shows no significant differences with the normal distribution (p
 720 = 0.41) indicating that this relationship is linear. A Breusch-Pagan test showed no evidence for
 721 heteroskedasticity (BP = 7, df = 8, p = 0.5).

Table 9: Summary and estimates for the Quality flaking model.

Covariate	Estimate	95% CI	t(12)	p
(Intercept)	-0.1	[-0.31, 0.17]	-0.62	0.55
Training condition[Trained]*Fitts score	-0.5	[-1.15, 0.06]	-1.97	0.07
Grip strength	-0.4	[-0.86, 0.07]	-1.84	0.09
BEAST score*RPM	0.3	[0.06, 0.55]	2.70	0.02
Highest n-back level*Fitts score	-0.3	[-0.57, -0.09]	-2.98	0.01
Highest n-back level	-0.3	[-0.54, 0.01]	-2.10	0.06
Fitts score	0.2	[-0.30, 0.75]	0.93	0.37
Nodule starting mass	0.1	[-0.06, 0.33]	1.48	0.17
Grip strength*Training condition[Trained]	0.2	[-0.44, 0.85]	0.68	0.51

722 **Table 9** presents this model's coefficients and summary outputs following the same data format as
 723 **Table 7 & Table 8**. The results show no effects that were both significant and substantial (Estimate
 724 ≥ 0.5).

725 The Quality model did produce two statistically ($p < 0.05$) significant interaction effects (RPM *
 726 BEAST & Fitts * n-back). However, these interactions had relatively small effects on Quality (<0.5),
 727 and we believe that interpreting these results from our small, exploratory study would be inap-
 728 propriate. **Figure 16** shows the uneven distribution of data points for these interactions, which
 729 suggests vulnerability to leveraging effects of a small number of extreme value combinations.

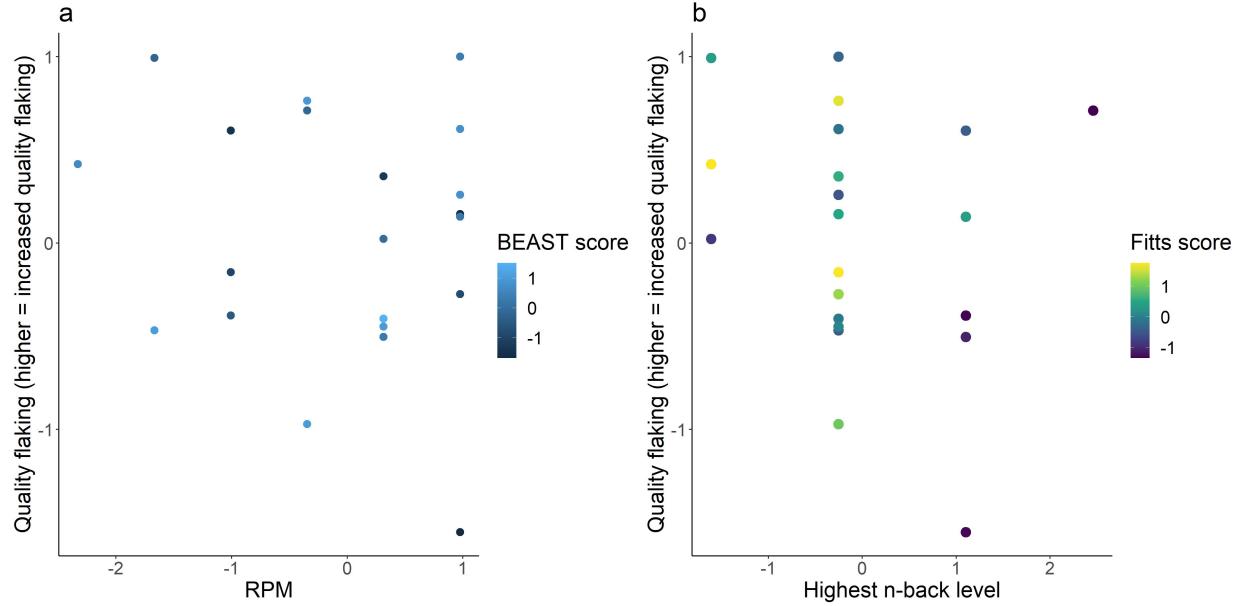


Figure 16: Significant interactions with low estimated effect sizes from the quality flaking model.

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

732 3.6 Behavioral observations

733 We designed this exploratory study primarily to trial experimental design elements such as train-
734 ing time, conditions, and raw materials and to collect preliminary data on individual differences
735 and training effects on knapping outcomes. We thus focused on collecting quantitative psychome-
736 tric and lithic data. However, we also considered that quantifying participant knapping behaviors
737 and products could be important for future studies. To support methods development in this
738 regard, we made ad hoc notes on observed behaviors during the experiments and video-recorded
739 all experiments to enable later, more systematic analyses yet to be completed. However, even
740 casual behavior observation was sufficient to reveal an unexpected effect. Whereas all instructed
741 participants copied the general posture and technique of the expert (freehand knapping seated
742 in a chair), fully half (6) of the uninstructed participants experimented with or even knapped all
743 their cores using the floor as support (**Figure 17**). Three of these participants were in the same
744 session, the only group composed of just three individuals. In this group, knapping on the ground
745 appears to have been transmitted from one participant to the other two based on appearance
746 order and the point of gaze of participants.

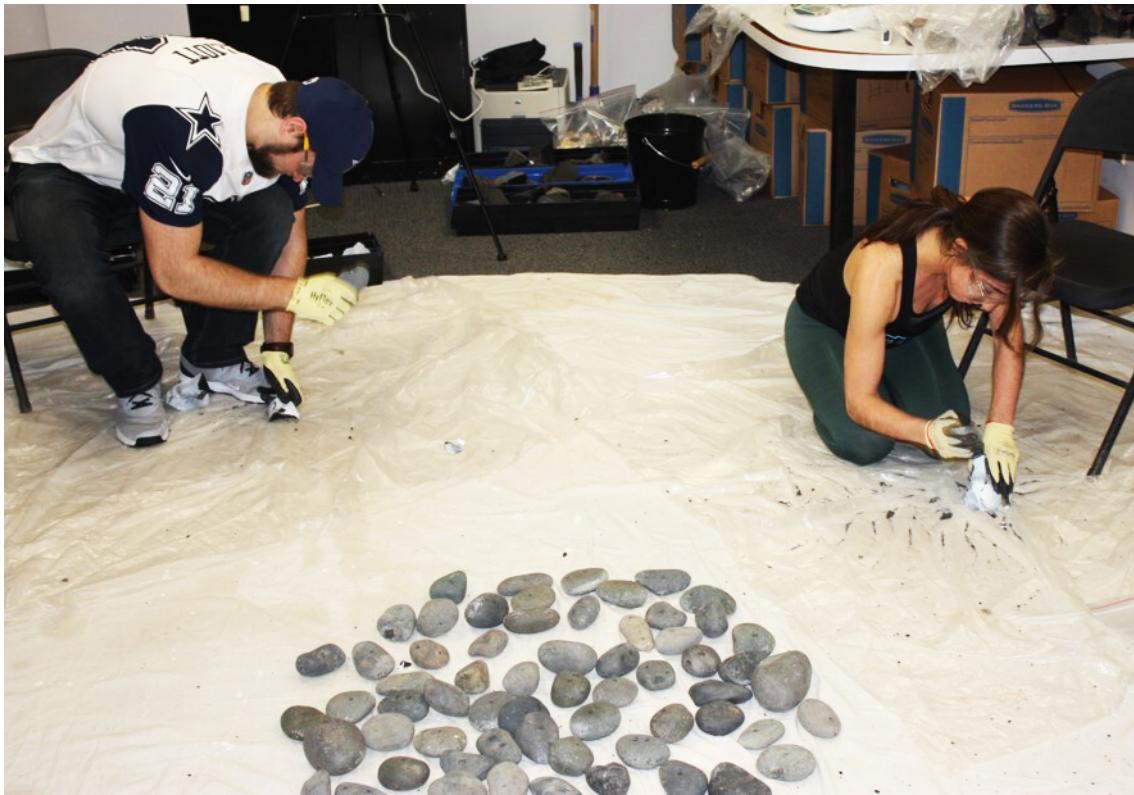


Figure 17: Novices in the uninstructed condition developing and transmitting a form of bipolar technology involving core reduction on the floor.

747 4 Discussion

748 The most salient finding of our exploratory study is that the presence/absence of teaching clearly
749 impacted knapping performance but did so in nuanced and individually variable ways that
750 previous research has not explored. We did not observe any significant differences in mean
751 performance between our experimental conditions. Some non-significant tendencies toward
752 enhanced instructed group performance suggest that a larger participant sample might detect
753 significant effects, but also that the size of any such effects would likely remain small. This result
754 could reasonably lead one to conclude that teaching does not substantially facilitate early-stage
755 knapping skill acquisition ([Ohnuma et al., 1997](#); [Putt et al., 2014](#)). Looking closer, however, it is
756 evident that teaching has several significant effects.

757 **4.1 Variance Reduction**

758 In our experiment, the strongest effects of teaching were to reduce variance (**Figure 9, Table**
759 **5**) rather than shift mean values. Instruction acted as a “safety net” that homogenized perfor-
760 mance by reducing the frequency of inferior outcomes (i.e., learning failures). This finding
761 provides additional support for the hypothesis that teaching would have increased the reliability
762 of Oldowan skill reproduction ([Morgan et al., 2015](#)) while simultaneously corroborating the view
763 that naïve knappers can achieve basic flaking competence in its absence ([Tennie et al., 2017](#)).
764 Our results thus do not imply that novices required teaching or that teaching was even present
765 during Oldowan times ([Gärdenfors & Högberg \(2017\)](#)). Instead, the results serve to reinforce
766 the plausibility of co-evolutionary scenarios positing the cost/reliability of technological skill
767 acquisition as a selection pressure favoring the evolution of teaching and language ([Morgan et al.,](#)
768 [2015; Stout, 2010; Stout & Hecht, 2017](#)).

769 **4.2 Knapping Behaviors**

770 The current study complements the transmission chain design of [Morgan et al. \(2015\) \(2015\)](#) by
771 finding similar effects in more naturalistic learning contexts. Our design further allowed us to
772 examine individual variation to understand better how teaching produces its effects. Whereas
773 transmission chains are optimized to investigate iterative learning effects (see [Caldwell et al.,](#)
774 [2020](#)), they necessarily involve each participant’s different instructor/model. We sacrificed this
775 iterative component to consider how the presence/absence of teaching affected the behavior of
776 individuals under otherwise standardized learning conditions.

777 We found that a critical impact of teaching was to alter basic flake production strategies, as
778 reflected in the relationship between Total Productivity and detached piece Quality (**Figure 5b**).
779 Whereas uninstructed participants achieved greater Productivity at the expense of Quality and
780 vice versa, these dimensions were unrelated across instructed participants. Thus, even though
781 uninstructed participants gained the highest values on each metric, only instructed individuals
782 managed to maximize both simultaneously. Indeed, a core function of teaching is to reduce the
783 search space that learners must explore and increase the likelihood of discovering globally instead
784 of locally optimum solutions (cf. [Hinton & Nowlan, 1996; Stout, 2013](#)). In our study, uninstructed
785 individuals explored a greater range of basic behavioral variations not seen in the instructed
786 group, including knapping on the floor, concentrating on working just a few cores (2-4, **Figure**

787 9c) over the practice period. Uninstructed individuals also showed less constrained nodule size
788 preferences (**Figure 8**). Notably, this variation occurred even in the presence of an observable
789 expert example, suggesting it may be interesting for future experiments to address the impact of
790 social context, expectations, and relationships on observational learning strategies ([Kendal et al.,
791 2018](#)).

792 We also found a strong positive effect of Grip Strength on Total Productivity independent of
793 instruction (**Figure 12**). While it is tempting to interpret this with respect to the demands for hand
794 strength specifically, it is important to remember that grip strength correlates with total muscle
795 strength ([Wind et al., 2010](#)) and overall fitness ([Sasaki et al., 2007](#)). Thus, it is best taken to indicate
796 some importance of fitness generally in increasing the rate and intensity of core reduction by
797 naïve knappers, potentially affecting the rate of work and the kinetic energy of the swing and the
798 handling of core and hammerstone. It thus provides further support for hypotheses positing stone
799 tool-making as a selection pressure on the functional anatomy of hand, arm, and shoulder (e.g.,
800 [Williams-Hatala et al., 2018](#)), but initially appears orthogonal to variations in learning condition
801 and knapping behaviors in our study. However, we also found that the effect of Grip Strength on
802 per-core knapping Quantity is dependent on teaching (**Figure 15**). The absence of this effect in
803 the uninstructed group reflects the weaker association between Total Productivity and per-core
804 Quality across these participants (**Figure 5a**). Grip Strength increased uninstructed Productivity
805 specifically by allowing them to knap more cores rather than to reduce individual cores more
806 heavily. In keeping with this, uninstructed Grip Strength positively correlates with Total Cores
807 knapped ($R^2 = 0.54$, $p = 0.01$). Conversely, strength allowed instructed participants to increase
808 their average Quantity per core without affecting the total number of cores knapped ($R^2 = 0.18$, $p =$
809 0.165). Thus, strength appears to have affected core reduction rate and intensity in different ways,
810 depending on teaching. This difference is likely related to the homogenizing effect of teaching on
811 knapping rate (all instructed participants knapped six or more cores) and methods. Subjectively,
812 uninstructed participants' knapping behaviors often appeared more physically demanding (e.g.,
813 a greater number of non-productive blows, rapid and unregulated battering). This result would
814 imply different demands on strength and aerobic fitness ([Mateos et al., 2019; Williams-Hatala et
815 al., 2021](#)). However, future research will need to test this observation systematically.

816 In this respect, it is also important to note that we do not know how well the knapping objectives
817 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). Still, 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). Nevertheless, the effect of instruction to constrain behavioral exploration and homogenize outcomes is clear. We expect this effect to generalize to teaching alternative knapping goals and behaviors, although future work should test this prediction.

4.3 Learning Strategies

One primary goal of this experiment was to test the viability of a moderate, two-hour learning period for skill acquisition studies. Unfortunately, we found that this duration was insufficient to capture learning effects for Oldowan-like flake production. We cannot attribute the lack of performance change over the period (Figure 11) to a ceiling effect (i.e., rapid task mastery at the outset of the practice period). 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 & Khriesheh, 2015). Future investigations of learning variation across individuals and/or experimental conditions may thus need to incorporate more extended practice periods to capture skill acquisition processes. In theory, future experiments might use shorter knapping trials to assess the variation in initial performance across individuals and under different conditions 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

848 expectation (Wilson et al., 2021). Teaching and other social learning resources have the potential
849 to provide low-cost information about task structure and payoffs (Kendal et al., 2018; Rendell
850 et al., 2010), which, if adopted, would be expected to affect exploration/exploitation decisions.
851 Such adoption is itself known to be influenced by individual cognitive differences, for example, if
852 higher fluid intelligence allows observers to better understand observed tasks (Vostroknutov et
853 al., 2018) or if individuals vary in their tendency to use and value social information (Molleman et
854 al., 2019; Toelch et al., 2014).

855 Our study did not observe any effect of fluid intelligence (RPM) on knapping outcomes. Still,
856 we found strong learning condition interactions with participant visuospatial working memory
857 and social information use tendency (Figure 13). As expected, uninstructed individuals with
858 higher scores on these dimensions displayed higher Total Productivity and average per-core
859 flaking Quantity (although the effect on n-Back on Quantity did not achieve significance). We
860 attribute these effects to an increased ability to hold relevant morphological/spatial information
861 in mind and a tendency to benefit from observing successful strategies of others, including
862 the expert model. In contrast, instructed individuals with higher scores tended to have lower
863 Productivity and Quantity. We interpret this unexpected effect as an increased tendency to
864 privilege exploratory learning behavior overexploitation. We suggest that instructed participants
865 might knap more slowly and less productively if higher working memory capacity inclined them
866 to experiment more with morphological/spatial variables highlighted by the instructor. Instructed
867 subjects' predisposition to use social information might have caused them to invest greater time
868 and effort attending to and trying out observed actions and/or instructions. These suggestions
869 remain to be tested by further work.

870 Putt et al. (2014) report a similar negative effect of instruction on knapping outcomes during
871 early-stage learning. They interpret this result to reflect learners experimenting with advanced
872 techniques before they have the perceptual-motor skill to execute them (Stout & Khreich, 2015;
873 Whiten, 2015). Unfortunately, the training period in our current experiment was insufficient to
874 capture learning effects, so we have no evidence of the impact of these individual differences
875 and putative exploration/exploitation trade-offs to the ultimate achievement of expertise. Future
876 research might explore such effects with more detailed behavioral data instead of purely lithic
877 data and with more extended learning periods.

878 **4.4 Limitations and Prospects**

879 Our exploratory study produced several robust results concerning the effects of instruction and
880 individual differences on lithic products. However, we had a small sample size, short training
881 duration, and a lack of detailed quantification of observed behaviors. We can hopefully address
882 these limitations in future studies building on the methods and evidence presented here. For
883 example, it is notable that our study failed to document any reliable effects on knapping Quality.
884 This lack of a result might reflect an actual lack of such effects. Still, it may also indicate a need for
885 more sensitive measures and/or increased sample size and training duration to identify subtle or
886 delayed effects. One aspect of our attempt to balance pragmatic costs and benefits in our study
887 was to test the efficacy of relatively limited lithic analysis. More detailed ongoing investigations of
888 core morphology and debitage features (e.g., typology, cutting edge length, platform dimensions)
889 may yet reveal a more reliable signal of knapping quality. Results of the Quality model also
890 seem to suffer from the uneven distribution and discrete rather than continuous nature of
891 scores on our RPM and n-Back tests. We could address concerns about sampling variation on
892 these dimensions with larger samples or by pre-screening participants to ensure more even
893 representation. Alternative psychometric tests (e.g., complete rather than the short version of the
894 RPM) might also provide more sensitive and continuous measures.

895 Another major limitation our study shares with all other published experiments on knapping skill
896 acquisition is that we do not address variation in social and cultural context or teaching style
897 ([Kline, 2015](#)) . Currently, we have little basis other than personal experience/tradition ([Callahan,](#)
898 [1979; Shea, 2015; Whittaker, 1994](#)) and theoretical speculation ([Stout, 2013; Whiten, 2015](#)) from
899 which to assess which pedagogical techniques are most effective even in WEIRD contexts. No
900 study to date has considered how variation in teacher skill ([Shea, 2015](#)) or social relationship to
901 participants might impact learning under different conditions. Addressing these questions would
902 require a significant research program, including cross-cultural comparative studies ([Barrett,](#)
903 [2020; Fuentes, 2016](#)) and more naturalistic study designs. While costly, such research would
904 produce results of broad relevance to anthropologists, biologists, psychologists, and sociologists
905 interested in teaching and learning.

906 5 Conclusions

907 Our exploratory study produced findings with both theoretical and methodological implications.
908 Perhaps the clearest result is that the effects of different experimental learning conditions and
909 their interaction with individual cognitive and motor differences are complicated, and researchers
910 must interpret them with care. For example, we did not find robust effects of instruction on mean
911 group performance. Still, we did find that teaching reduced variation knapping rate, methods, and
912 outcomes, thus providing a “safety net” against failed transmission. This increase in reliability of
913 skill reproduction offers further support for the hypothetical co-evolution of teaching, language,
914 and tool-making ([Morgan et al., 2015](#); [Stout, 2010](#)). At the same time, we found that two hours of
915 practice was insufficient to document learning-related performance change. This result implies
916 that we may need longer experiment durations to study learning, even for basic Oldowan-like
917 flaking. Researchers should take care when interpreting results from shorter studies. It is thus
918 important to have a quantitative standard of proficiency, ideally archaeological ([Pargeter et](#)
919 [al., 2019](#); [Stout et al., 2019](#)) but at least experimental (this study), against which to compare
920 performance. Here, it was clear that participants remained in an early learning stage far from
921 expert performance and that teaching had important effects on how learners explored the novel
922 task space. Instruction allowed learners to maximize both quality and quantity rather than trading
923 off between them and dramatically changed the effects of individual differences in strength,
924 visuospatial working memory, and social learning tendencies on knapping outcomes. These
925 results highlight the value of including individual difference measures in knapping experiments
926 and suggest that the presence/absence of teaching in the Paleolithic would have fundamentally
927 altered associated selective pressures ([Fuentes, 2015](#); [Stout et al., 2019](#)) in addition to affecting the
928 reliability of skill reproduction. To unravel these complicated and nuanced relationships, future
929 experiments might benefit from more extended instruction periods, video-based quantitative
930 analyses of knapping behaviors ([Geribàs et al., 2010](#); [Stout et al., 2021](#)), more refined psychometric
931 and lithic measures, a greater diversity of social and cultural contexts ([Barrett, 2020](#)), and explicit
932 attention to teacher skill, style, and interpersonal relationships with participants ([Bevilacqua et](#)
933 [al., 2019](#)).

934 **References**

- 935 Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation
936 supports flexible tool use and physical reasoning. *Proceedings of the National Academy of
937 Sciences*, 117(47), 29302–29310. <https://doi.org/10.1073/pnas.1912341117>
- 938 Barrett, H. C. (2020). Towards a Cognitive Science of the Human: Cross-Cultural Approaches and
939 Their Urgency. *Trends in Cognitive Sciences*, 24(8), 620–638. [https://doi.org/10.1016/j.tics.2020.05.007](https://doi.org/10.1016/j.tics.202
940 0.05.007)
- 941 Bevilacqua, D., Davidesco, I., Wan, L., Chaloner, K., Rowland, J., Ding, M., Poeppel, D., & Dikker, S.
942 (2019). Brain-to-Brain Synchrony and Learning Outcomes Vary by Student–Teacher Dynamics:
943 Evidence from a Real-world Classroom Electroencephalography Study. *Journal of Cognitive
944 Neuroscience*, 31(3), 401–411. https://doi.org/10.1162/jocn_a_01274
- 945 Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Develop-
946 opment of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test.
947 *Assessment*, 19(3), 354–369. <https://doi.org/10.1177/1073191112446655>
- 948 Boogert, N. J., Madden, J. R., Morand-Ferron, J., & Thornton, A. (2018). Measuring and under-
949 standing individual differences in cognition. *Philosophical Transactions of the Royal Society B:
950 Biological Sciences*, 373(1756), 20170280. <https://doi.org/10.1098/rstb.2017.0280>
- 951 Boyette, A. H., & Hewlett, B. S. (2017). Autonomy, Equality, and Teaching among Aka Foragers and
952 Ngandu Farmers of the Congo Basin. *Human Nature*, 28(3), 289–322. [https://doi.org/10.1007/s12110-017-9294-y](https://doi.org/10.1007/
953 s12110-017-9294-y)
- 954 Braun, D. R., Aldeias, V., Archer, W., Arrowsmith, J. R., Baraki, N., Campisano, C. J., Deino, A.
955 L., DiMaggio, E. N., Dupont-Nivet, G., Engda, B., Feary, D. A., Garello, D. I., Kerfelew, Z.,
956 McPherron, S. P., Patterson, D. B., Reeves, J. S., Thompson, J. C., & Reed, K. E. (2019). Earliest
957 known Oldowan artifacts at >2.58 Ma from Ledi-Geraru, Ethiopia, highlight early technological
958 diversity. *Proceedings of the National Academy of Sciences*, 116(24), 11712–11717. <https://doi.org/10.1073/pnas.1820177116>
- 960 Braun, D. R., Plummer, T., Ferraro, J. V., Ditchfield, P., & Bishop, L. C. (2009). Raw material quality
961 and Oldowan hominin toolstone preferences: Evidence from Kanjera South, Kenya. *Journal of
962 Archaeological Science*, 36(7), 1605–1614. <https://doi.org/10.1016/j.jas.2009.03.025>

- 963 Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical*
964 *Information-Theoretic Approach* (2nd ed.). Springer-Verlag. <https://doi.org/10.1007/b97636>
- 965 Caldwell, C. A., Atkinson, M., Blakey, K. H., Dunstone, J., Kean, D., Mackintosh, G., Renner, E., &
966 Wilks, C. E. H. (2020). Experimental assessment of capacities for cumulative culture: Review
967 and evaluation of methods. *WIREs Cognitive Science*, 11(1), e1516. <https://doi.org/10.1002/wcs.1516>
- 969 Callahan, E. (1979). The basics of biface knapping in the eastern fluted point tradition: A manual
970 for flintknappers and lithic analysts. *Archaeology of Eastern North America*, 7(1), 1–180.
971 <https://www.jstor.org/stable/40914177>
- 972 Cataldo, D. M., Migliano, A. B., & Vinicius, L. (2018). Speech, stone tool-making and the evolution
973 of language. *PLOS ONE*, 13(1), e0191071. <https://doi.org/10.1371/journal.pone.0191071>
- 974 Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of*
975 *Educational Psychology*, 54(1), 1–22. <https://doi.org/10.1037/h0046743>
- 976 Coolidge, F. L., & Wynn, T. (2005). Working Memory, its Executive Functions, and the Emergence
977 of Modern Thinking. *Cambridge Archaeological Journal*, 15(1), 5–26. <https://doi.org/10.1017/S0959774305000016>
- 979 Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or, The Preservation of*
980 *Favoured Races in the Struggle for Life* (1st ed.). John Murray.
- 981 Darwin, C. (1871). *The descent of man, and selection in relation to sex* (1st ed.). John Murray.
- 982 Duke, H., & Pargeter, J. (2015). Weaving simple solutions to complex problems: An experimental
983 study of skill in bipolar cobble-splitting. *Lithic Technology*, 40(4), 349–365. <https://doi.org/10.1179/2051618515Y.0000000016>
- 985 Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psycho-*
986 *logical Science*, 13(2), 190–193. <https://doi.org/10.1177/1745691617720478>
- 987 Engles, F. (2003). The part played by labour in the transition from ape to man. In R. C. Scharff &
988 V. Dusek (Eds.), *Philosophy of Technology – The Technological Condition: An Anthology* (pp.
989 71–77). Blackwell.

- 990 Eren, M. I., Bradley, B. A., & Sampson, C. G. (2011). Middle Paleolithic Skill Level and the Individual
991 Knapper: An Experiment. *American Antiquity*, 76(2), 229–251. <https://doi.org/10.7183/0002-7316.76.2.229>
- 993 Eren, M. I., Lycett, S. J., Patten, R. J., Buchanan, B., Pargeter, J., & O'Brien, M. J. (2016). Test, model,
994 and method validation: The role of experimental stone artifact replication in hypothesis-
995 driven archaeology. *Ethnoarchaeology: Journal of Archaeological, Ethnographic and Experi-
996 mental Studies*, 8(2), 103–136. <https://doi.org/10.1080/19442890.2016.1213972>
- 997 Eren, M. I., Roos, C. I., Story, B. A., von Cramon-Taubadel, N., & Lycett, S. J. (2014). The role of raw
998 material differences in stone tool shape variation: an experimental assessment. *Journal of
999 Archaeological Science*, 49, 472–487. <https://doi.org/10.1016/j.jas.2014.05.034>
- 1000 Faisal, A., Stout, D., Apel, J., & Bradley, B. (2010). The Manipulative Complexity of Lower Paleolithic
1001 Stone Toolmaking. *PLOS ONE*, 5(11), e13718. <https://doi.org/10.1371/journal.pone.0013718>
- 1002 Fitts, P. M. (1954). The information capacity of the human motor system in controlling the
1003 amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–391. <https://doi.org/10.1037/h0055392>
- 1005 Fuentes, A. (2016). The Extended Evolutionary Synthesis, Ethnography, and the Human Niche:
1006 Toward an Integrated Anthropology. *Current Anthropology*, 57(S13), S13–S26. <https://doi.org/10.1086/685684>
- 1008 Fuentes, A. (2015). Integrative Anthropology and the Human Niche: Toward a Contemporary
1009 Approach to Human Evolution. *American Anthropologist*, 117(2), 302–315. <https://doi.org/10.1111/aman.12248>
- 1011 García-Medrano, P., Ollé, A., Ashton, N., & Roberts, M. B. (2019). The Mental Template in Handaxe
1012 Manufacture: New Insights into Acheulean Lithic Technological Behavior at Boxgrove, Sussex,
1013 UK. *Journal of Archaeological Method and Theory*, 26(1), 396–422. <https://doi.org/10.1007/s10816-018-9376-0>
- 1015 Gärdenfors, P., & Höglberg, A. (2017). The archaeology of teaching and the evolution of homo
1016 docens. *Current Anthropology*, 58(2), 188–208. <https://doi.org/10.1086/691178>
- 1017 Geribàs, N., Mosquera, M., & Vergès, J. M. (2010). What novice knappers have to learn to become
1018 expert stone toolmakers. *Journal of Archaeological Science*, 37(11), 2857–2870. <https://doi.or>

- 1019 g/10.1016/j.jas.2010.06.026
- 1020 Gowlett, J. A. J. (1984). Mental abilities of early man: A look at some hard evidence. *Higher
1021 Education Quarterly*, 38(3), 199–220. <https://doi.org/10.1111/j.1468-2273.1984.tb01387.x>
- 1022 Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting
1023 to new responses in a weigl-type card-sorting problem. *Journal of Experimental Psychology*,
1024 38(4), 404–411. <https://doi.org/10.1037/h0059831>
- 1025 Hecht, E. E., Gutman, D. A., Bradley, B. A., Preuss, T. M., & Stout, D. (2015). Virtual dissection and
1026 comparative connectivity of the superior longitudinal fasciculus in chimpanzees and humans.
1027 *NeuroImage*, 108, 124–137. <https://doi.org/10.1016/j.neuroimage.2014.12.039>
- 1028 Hecht, E. E., Gutman, D. A., Khreicheh, N., Taylor, S. V., Kilner, J. M., Faisal, A. A., Bradley, B. A.,
1029 Chaminade, T., & Stout, D. (2015). Acquisition of Paleolithic toolmaking abilities involves
1030 structural remodeling to inferior frontoparietal regions. *Brain Structure & Function*, 220(4),
1031 2315–2331. <https://doi.org/10.1007/s00429-014-0789-6>
- 1032 Hecht, E. E., Gutman, D. A., Preuss, T. M., Sanchez, M. M., Parr, L. A., & Rilling, J. K. (2013). Process
1033 versus product in social learning: Comparative diffusion tensor imaging of neural systems
1034 for action executionobservation matching in macaques, chimpanzees, and humans. *Cerebral
1035 Cortex*, 23(5), 1014–1024. <https://doi.org/10.1093/cercor/bhs097>
- 1036 Hecht, E. E., Murphy, L. E., Gutman, D. A., Votaw, J. R., Schuster, D. M., Preuss, T. M., Orban,
1037 G. A., Stout, D., & Parr, L. A. (2013). Differences in neural activation for object-directed
1038 grasping in chimpanzees and humans. *The Journal of Neuroscience*, 33(35), 14117–14134.
1039 <https://doi.org/10.1523/JNEUROSCI.2172-13.2013>
- 1040 Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302),
1041 29–29. <https://doi.org/10.1038/466029a>
- 1042 Hewes, G. W. (1993). A history of speculation on the relation between tools and language. In K.
1043 R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution* (pp. 20–31).
1044 Cambridge University Press.
- 1045 Heyes, C. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transac-
1046 tions of the Royal Society B: Biological Sciences*, 373(1743), 20170051. <https://doi.org/10.1098/rstb.2017.0051>

- 1048 Hinton, G. E., & Nowlan, S. J. (1996). How learning can guide evolution. In R. K. Belew & M.
1049 Mitchell (Eds.), *Adaptive individuals in evolving populations: Models and algorithms* (pp.
1050 447–454). Addison-Wesley Publishing Company.
- 1051 Isaac, G. L. (1976). Stages of Cultural Elaboration in the Pleistocene: Possible Archaeological
1052 Indicators of the Development of Language Capabilities. *Annals of the New York Academy of
1053 Sciences*, 280(1), 275–288. <https://doi.org/10.1111/j.1749-6632.1976.tb25494.x>
- 1054 Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and
1055 instruction*. Lawrence Erlbaum.
- 1056 Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social
1057 Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, 22(7),
1058 651–665. <https://doi.org/10.1016/j.tics.2018.04.003>
- 1059 Key, A. J. M., & Dunmore, C. J. (2015). The evolution of the hominin thumb and the influence
1060 exerted by the non-dominant hand during stone tool production. *Journal of Human Evolution*,
1061 78, 60–69. <https://doi.org/10.1016/j.jhevol.2014.08.006>
- 1062 Key, A. J. M., & Dunmore, C. J. (2018). Manual restrictions on Palaeolithic technological behaviours.
1063 *PeerJ*, 6, e5399. <https://doi.org/10.7717/peerj.5399>
- 1064 Key, A. J. M., & Lycett, S. J. (2014). Are bigger flakes always better? An experimental assessment of
1065 flake size variation on cutting efficiency and loading. *Journal of Archaeological Science*, 41,
1066 140–146. <https://doi.org/10.1016/j.jas.2013.07.033>
- 1067 Key, A. J. M., & Lycett, S. J. (2019). Biometric variables predict stone tool functional performance
1068 more effectively than tool-form attributes: a case study in handaxe loading capabilities.
1069 *Archaeometry*, 61(3), 539–555. <https://doi.org/10.1111/arcm.12439>
- 1070 Khreisheh, N. N., Davies, D., & Bradley, B. A. (2013). Extending Experimental Control: The Use of
1071 Porcelain in Flaked Stone Experimentation. *Advances in Archaeological Practice*, 1(1), 38–46.
1072 <https://doi.org/10.7183/2326-3768.1.1.37>
- 1073 Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of
1074 teaching behavior in humans and other animals. *The Behavioral and Brain Sciences*, 38, e31.
1075 <https://doi.org/10.1017/S0140525X14000090>

- 1076 Laland, K. N. (2017). The origins of language in teaching. *Psychonomic Bulletin & Review*, 24(1),
1077 225–231. <https://doi.org/10.3758/s13423-016-1077-7>
- 1078 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>
- 1081 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>
- 1084 Marwick, B. (2017). Computational Reproducibility in Archaeological Research: Basic Principles
1085 and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*,
1086 24(2), 424–450. <https://doi.org/10.1007/s10816-015-9272-9>
- 1087 Marzke, M. W., Toth, N., Schick, K., Reece, S., Steinberg, B., Hunt, K., Linscheid, R. L., & An, K.-N.
1088 (1998). EMG study of hand muscle recruitment during hard hammer percussion manufacture
1089 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](https://doi.org/10.1002/(SICI)1096-8644(199803)105:3%3C315::AID-AJPA3%3E3.0.CO;2-Q)
- 1091 Mateos, A., Terradillos-Bernal, M., & Rodríguez, J. (2019). Energy Cost of Stone Knapping. *Journal
1092 of Archaeological Method and Theory*, 26(2), 561–580. [https://doi.org/10.1007/s10816-018-9382-2](https://doi.org/10.1007/s10816-018-
1093 9382-2)
- 1094 Miu, E., Gulley, N., Laland, K. N., & Rendell, L. (2020). Flexible learning, rather than inveterate
1095 innovation or copying, drives cumulative knowledge gain. *Science Advances*, 6(23), eaaz0286.
1096 <https://doi.org/10.1126/sciadv.aaz0286>
- 1097 Molleman, L., Kurvers, R. H. J. M., & van den Bos, W. (2019). Unleashing the BEAST: a brief
1098 measure of human social information use. *Evolution and Human Behavior*, 40(5), 492–499.
1099 <https://doi.org/10.1016/j.evolhumbehav.2019.06.005>
- 1100 Montagu, A. (1976). Toolmaking, Hunting, and the Origin of Language. *Annals of the New York
1101 Academy of Sciences*, 280(1), 266–274. <https://doi.org/10.1111/j.1749-6632.1976.tb25493.x>
- 1102 Morgan, T. J. H., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M.,
1103 Cross, C. P., Evans, C., Kearney, R., de la Torre, I., Whiten, A., & Laland, K. N. (2015). Experimental
1104 evidence for the co-evolution of hominin tool-making teaching and language. *Nature*

- 1105 *Communications*, 6(1), 6029. <https://doi.org/10.1038/ncomms7029>
- 1106 Nonaka, T., Bril, B., & Rein, R. (2010). How do stone knappers predict and control the outcome
1107 of flaking? Implications for understanding early stone tool technology. *Journal of Human
1108 Evolution*, 59(2), 155–167. <https://doi.org/10.1016/j.jhevol.2010.04.006>
- 1109 Oakley, K. P. (1949). *Man the toolmaker*. Trustees of the British Museum.
- 1110 Ohnuma, K., Aoki, K., & Akazawa, A. T. (1997). Transmission of tool-making through verbal
1111 and non-verbal communication: Preliminary experiments in levallois flake production.
1112 *Anthropological Science*, 105(3), 159–168. <https://doi.org/10.1537/ase.105.159>
- 1113 Pargeter, J., Khreisheh, N., Shea, J. J., & Stout, D. (2020). Knowledge vs. know-how? Dissecting
1114 the foundations of stone knapping skill. *Journal of Human Evolution*, 145, 102807. <https://doi.org/10.1016/j.jhevol.2020.102807>
- 1116 Pargeter, J., Khreisheh, N., & Stout, D. (2019). Understanding stone tool-making skill acquisition:
1117 Experimental methods and evolutionary implications. *Journal of Human Evolution*, 133,
1118 146–166. <https://doi.org/10.1016/j.jhevol.2019.05.010>
- 1119 Pelegrin, J. (1990). Prehistoric Lithic Technology : Some Aspects of Research. *Archaeological
1120 Review from Cambridge*, 9(1), 116–125. [https://paperkit.net/paper/Prehistoric-Lithic-Technology-%3A-Some-
Aspects-of-Pelegrin/5e02fc2a5280ac128727275ab6b833756e6a6056](https://paperkit.net/paper/Prehistoric-Lithic-Technology-%3A-Some-
1121 Aspects-of-Pelegrin/5e02fc2a5280ac128727275ab6b833756e6a6056)
- 1122 Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to
1123 large-scale decoding. *Neuron*, 72(5), 692–697. <https://doi.org/10.1016/j.neuron.2011.11.001>
- 1124 Prascunas, M. M. (2007). Bifacial Cores and Flake Production Efficiency: An Experimental Test of
1125 Technological Assumptions. *American Antiquity*, 72(2), 334–348. [https://doi.org/10.2307/40035817](https://doi.org/10.2307/40
1126 035817)
- 1127 Putt, S. S. (2015). The origins of stone tool reduction and the transition to knapping: An experi-
1128 mental approach. *Journal of Archaeological Science: Reports*, 2, 51–60. [https://doi.org/10.1016/j.jasrep.2015.01.004](https://doi.org/10.101
1129 6/j.jasrep.2015.01.004)
- 1130 Putt, S. S., Wijekumar, S., Franciscus, R. G., & Spencer, J. P. (2017). The functional brain networks
1131 that underlie Early Stone Age tool manufacture. *Nature Human Behaviour*, 1(6), 1–8. <https://doi.org/10.1038/s41562-017-0102>

- 1133 Putt, S. S., Wijekumar, S., & Spencer, J. P. (2019). Prefrontal cortex activation supports the
1134 emergence of early stone age toolmaking skill. *NeuroImage*, 199, 57–69. <https://doi.org/10.1016/j.neuroimage.2019.05.056>
- 1136 Putt, S. S., Woods, A. D., & Franciscus, R. G. (2014). The role of verbal interaction during
1137 experimental bifacial stone tool manufacture. *Lithic Technology*, 39(2), 96–112. <https://doi.org/10.1179/0197726114Z.00000000036>
- 1139 Rein, R., Nonaka, T., & Bril, B. (2014). Movement Pattern Variability in Stone Knapping: Im-
1140 plications for the Development of Percussive Traditions. *PLOS ONE*, 9(11), e113567. <https://doi.org/10.1371/journal.pone.0113567>
- 1142 Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., Fogarty, L.,
1143 Ghirlanda, S., Lillicrap, T., & Laland, K. N. (2010). Why Copy Others? Insights from the
1144 Social Learning Strategies Tournament. *Science*, 328(5975), 208–213. <https://doi.org/10.1126/science.1184719>
- 1146 Reti, J. S. (2016). Quantifying Oldowan Stone Tool Production at Olduvai Gorge, Tanzania. *PLOS
1147 ONE*, 11(1), e0147352. <https://doi.org/10.1371/journal.pone.0147352>
- 1148 Roux, V., Bril, B., & Dietrich, G. (1995). Skills and learning difficulties involved in stone knapping:
1149 The case of stone-bead knapping in khambhat, india. *World Archaeology*, 27(1), 63–87. <https://doi.org/10.1080/00438243.1995.9980293>
- 1151 Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W.
1152 (2017). ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*,
1153 18(1), 529. <https://doi.org/10.1186/s12859-017-1934-z>
- 1154 Sasaki, H., Kasagi, F., Yamada, M., & Fujita, S. (2007). Grip Strength Predicts Cause-Specific
1155 Mortality in Middle-Aged and Elderly Persons. *The American Journal of Medicine*, 120(4),
1156 337–342. <https://doi.org/10.1016/j.amjmed.2006.04.018>
- 1157 Schillinger, K., Mesoudi, A., & Lycett, S. J. (2014). Copying Error and the Cultural Evolution
1158 of “Additive” vs. “Reductive” Material Traditions: An Experimental Assessment. *American
1159 Antiquity*, 79(1), 128–143. <https://doi.org/10.7183/0002-7316.79.1.128>
- 1160 Shallice, T., Broadbent, D. E., & Weiskrantz, L. (1982). Specific impairments of planning. *Philoso-
1161 sophical Transactions of the Royal Society of London. B, Biological Sciences*, 298(1089), 199–209.

- 1162 <https://doi.org/10.1098/rstb.1982.0082>
- 1163 Shea, J. J. (2015). Making and using stone tools: Advice for learners and teachers and insights for
1164 archaeologists. *Lithic Technology*, 40(3), 231–248. <https://doi.org/10.1179/2051618515Y.0000>
1165 000011
- 1166 Shea, J. J. (2016). *Stone tools in human evolution: Behavioral differences among technological*
1167 *primates*. Cambridge University Press. <https://doi.org/10.1017/9781316389355>
- 1168 Sherwood, C. C., & Gómez-Robles, A. (2017). Brain plasticity and human evolution. *Annual Review*
1169 *of Anthropology*, 46(1), 399–419. <https://doi.org/10.1146/annurev-anthro-102215-100009>
- 1170 Shipton, C. (2018). Biface Knapping Skill in the East African Acheulean: Progressive Trends and
1171 Random Walks. *African Archaeological Review*, 35(1), 107–131. <https://doi.org/10.1007/s104>
1172 37-018-9287-1
- 1173 Stout, D. (2002). Skill and cognition in stone tool production: An ethnographic case study from
1174 irian jaya. *Current Anthropology*, 43(5), 693–722. <https://doi.org/10.1086/342638>
- 1175 Stout, D. (2010). Possible relations between language and technology in human evolution. In A.
1176 Nowell & I. Davidson (Eds.), *Stone tools and the evolution of human cognition* (pp. 159–184).
1177 University Press of Colorado.
- 1178 Stout, D. (2013). Neuroscience of technology. In P. J. Richerson & M. H. Christiansen (Eds.),
1179 *Cultural evolution: Society, technology, language, and religion* (pp. 157–173). The MIT Press.
- 1180 Stout, D., Apel, J., Commander, J., & Roberts, M. (2014). Late Acheulean technology and cognition
1181 at Boxgrove, UK. *Journal of Archaeological Science*, 41, 576–590. <https://doi.org/10.1016/j.jas.>
1182 2013.10.001
- 1183 Stout, D., & Chaminade, T. (2007). The evolutionary neuroscience of tool making. *Neuropsycholo-*
1184 *gia*, 45(5), 1091–1100. <https://doi.org/10.1016/j.neuropsychologia.2006.09.014>
- 1185 Stout, D., & Chaminade, T. (2012). Stone tools, language and the brain in human evolution.
1186 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585), 75–87. <https://doi.org/10.1098/rstb.2011.0099>
- 1188 Stout, D., Chaminade, T., Apel, J., Shafti, A., & Faisal, A. A. (2021). The measurement, evolution,
1189 and neural representation of action grammars of human behavior. *Scientific Reports*, 11(1),

- 1190 13720. <https://doi.org/10.1038/s41598-021-92992-5>
- 1191 1192 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.1620738>
- 1193 114
- 1194 1195 1196 Stout, D., Hecht, E., Khriesheh, 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>
- 1197 1198 1199 Stout, D., & Khriesheh, N. (2015). Skill Learning and Human Brain Evolution: An Experimental Approach. *Cambridge Archaeological Journal*, 25(4), 867–875. <https://doi.org/10.1017/S0959774315000359>
- 1200 1201 1202 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>
- 1203 1204 1205 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>
- 1206 1207 1208 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>
- 1209 1210 1211 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.
- 1212 1213 Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.
- 1214 1215 1216 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>
- 1217 1218 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/10.1086/692100>

- 1219 [//doi.org/10.1086/693846](https://doi.org/10.1086/693846)
- 1220 Toelch, U., Bruce, M. J., Newson, L., Richerson, P. J., & Reader, S. M. (2014). Individual consistency
1221 and flexibility in human social information use. *Proceedings of the Royal Society B: Biological
1222 Sciences*, 281(1776), 20132864. <https://doi.org/10.1098/rspb.2013.2864>
- 1223 Toth, N., & Schick, K. (1993). Early stone industries and inferences regarding language and
1224 cognition. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution*
1225 (pp. 346–362). Cambridge University Press.
- 1226 Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and
1227 learning: Evidence from the serial reaction time task. *Memory & Cognition*, 33(2), 213–220.
1228 <https://doi.org/10.3758/BF03195310>
- 1229 Vostroknutov, A., Polonio, L., & Coricelli, G. (2018). The Role of Intelligence in Social Learning.
1230 *Scientific Reports*, 8(1), 6896. <https://doi.org/10.1038/s41598-018-25289-9>
- 1231 Washburn, S. L. (1960). Tools and human evolution. *Scientific American*, 203(3), 62–75. <https://doi.org/10.1038/scientificamerican0960-62>
- 1232 Whiten, A. (2015). Experimental studies illuminate the cultural transmission of percussive tech-
1233 nologies in homo and pan. *Philosophical Transactions of the Royal Society B: Biological
1234 Sciences*, 370(1682), 20140359. <https://doi.org/10.1098/rstb.2014.0359>
- 1235 Whittaker, J. C. (1994). *Flintknapping: Making and Understanding Stone Tools*. University of Texas
1236 Press.
- 1237 Wilkins, J. (2018). The Point is the Point: Emulative social learning and weapon manufacture
1238 in the Middle Stone Age of South Africa. In M. J. O'Brien, B. Buchanan, & M. I. Eren (Eds.),
1239 *Convergent Evolution in Stone-Tool Technology* (pp. 153–174). The MIT Press.
- 1240 Williams-Hatala, E. M., Hatala, K. G., Gordon, M., Key, A., Kasper, M., & Kivell, T. L. (2018). The
1241 manual pressures of stone tool behaviors and their implications for the evolution of the human
1242 hand. *Journal of Human Evolution*, 119, 14–26. <https://doi.org/10.1016/j.jhevol.2018.02.008>
- 1243 Williams-Hatala, E. M., Hatala, K. G., Key, A., Dunmore, C. J., Kasper, M., Gordon, M., & Kivell, T.
1244 L. (2021). Kinetics of stone tool production among novice and expert tool makers. *American
1245 Journal of Physical Anthropology*, 174(4), 714–727. <https://doi.org/10.1002/ajpa.24159>

- 1247 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.
- 1248
- 1249 <https://doi.org/10.1016/j.cobeha.2020.10.001>
- 1250 Wind, A. E., Takken, T., Helders, P. J. M., & Engelbert, R. H. H. (2010). Is grip strength a predictor for
- 1251 total muscle strength in healthy children, adolescents, and young adults? *European Journal of*
- 1252 *Pediatrics*, 169(3), 281–287. <https://doi.org/10.1007/s00431-009-1010-4>
- 1253 Wynn, T. (1979). The intelligence of later acheulean hominids. *Man*, 14(3), 371–391. <https://doi.org/10.2307/2801865>
- 1254
- 1255 Wynn, T. (2017). Evolutionary cognitive archaeology. In T. Wynn & F. Coolidge (Eds.), *Cognitive*
- 1256 *models in Palaeolithic archaeology* (pp. 1–20). Oxford University Press.
- 1257 Wynn, T., & Coolidge, F. L. (2004). The expert Neandertal mind. *Journal of Human Evolution*,
- 1258 46(4), 467–487. <https://doi.org/10.1016/j.jhevol.2004.01.005>
- 1259 Wynn, T., & Coolidge, F. L. (2016). Archeological insights into hominin cognitive evolution.
- 1260 *Evolutionary Anthropology: Issues, News, and Reviews*, 25(4), 200–213. <https://doi.org/10.1002/evan.21496>
- 1261
- 1262 Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. <https://doi.org/10.1017/S0140525X20001685>
- 1263