

1 Testing the motor and cognitive foundations of Paleolithic  
2 social transmission

3 Justin Pargeter\* Megan Beney Kilgore† Cheng Liu‡ Dietrich Stout§

4 **Abstract**

5 Stone tools provide key evidence of human cognitive evolution but remain difficult to  
6 interpret. Toolmaking skill-learning in particular has been understudied even though: 1) the  
7 most salient cognitive demands of toolmaking should occur during learning, and 2) variation  
8 in learning aptitude would have provided the raw material for any past selection acting on  
9 tool making ability. However, we actually know very little about the cognitive prerequisites  
10 of learning under different information transmission conditions that may have prevailed  
11 during the Paleolithic. This paper presents results from a pilot experimental study to trial new  
12 experimental methods for studying the effect of learning conditions and individual differences  
13 on Oldowan flake-tool making skill acquisition. We trained 23 participants for 2 hours to make  
14 stone flakes under two different instructional conditions (observation only vs. direct active  
15 teaching) employing appropriate raw materials, practice time, and real human interaction.  
16 Participant performance was evaluated through analysis of the stone artifacts produced.  
17 Performance was compared both across experimental groups and with respect to individual  
18 participant differences in grip strength, motor accuracy, and cognitive function measured  
19 for the study. Our results show aptitude to be associated with fluid intelligence in a verbally  
20 instructed group and with a tendency to use social information in an observation-only group.  
21 These results have implications for debates surrounding the cumulative nature of human  
22 culture, the relative contributions of knowledge and know-how for stone tool making, and  
23 the role of evolved psychological mechanisms in “high fidelity” transmission of information,  
24 particularly through imitation and teaching.

25 **Keywords:** Oldowan; Stone toolmaking; Social learning; Individual variation; Cognitive  
26 aptitudes; Motor skills

27 **Contents**

28 <b>1 Introduction</b>	<b>2</b>
29   1.1 Individual Differences . . . . .	4
30   1.2 Teaching, Language, and Tool Making . . . . .	7
31   1.3 Raw materials and knapping skill . . . . .	9
32 <b>2 Materials and Methods</b>	<b>10</b>
33   2.1 Participants . . . . .	10
34   2.2 Study Visit . . . . .	11

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

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

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

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

35	2.3 Individual Difference Measures . . . . .	11
36	2.4 Stone Tool Making . . . . .	13
37	2.5 Lithic Analysis . . . . .	19
38	2.6 Statistical Analyses . . . . .	20
39	<b>3 Results</b>	<b>20</b>
40	3.1 Principal Component analyses . . . . .	21
41	3.2 Relationships between Performance Measures . . . . .	22
42	3.3 Do trained, untrained, and expert knappers perform differently? . . . . .	24
43	3.4 Does performance change over time? . . . . .	26
44	3.5 Do individual differences in motor skill and psychometric measures predict flaking performance? . . . . .	28
45	3.6 Behavioral observations . . . . .	35
46		
47	<b>4 Discussion</b>	<b>36</b>
48	4.1 Variance Reduction . . . . .	37
49	4.2 Knapping Behaviors . . . . .	37
50	4.3 Learning Strategies . . . . .	39
51	4.4 Limitations and Prospects . . . . .	41
52	<b>5 Conclusions</b>	<b>42</b>
53	<b>6 Acknowledgments</b>	<b>42</b>
54	<b>7 Figures</b>	<b>42</b>
55	<b>References</b>	<b>42</b>

## 56 **1 Introduction**

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

68 social learning mechanisms such as imitation (Rein et al., 2014; Stout et al., 2019), emulation  
69 (Tehrani & Riede, 2008; Wilkins, 2018), and language (Cataldo et al., 2018; Lombao et al., 2017;  
70 Morgan et al., 2015; Ohnuma et al., 1997; Putt et al., 2017; Putt et al., 2014) to the reproduction of  
71 Paleolithic technologies.

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

83 Learning real-world skills like stone knapping is highly demanding of time and materials and  
84 difficult to control experimentally without sacrificing generalizability to real world conditions.  
85 Prior efforts have attempted to navigate these challenges by using various combinations of 1)  
86 inauthentic raw materials that are less expensive, easier to standardize, and/or easier to knap,  
87 2) video-recorded instruction that is uniform across participants and less demanding of experi-  
88 menter time, 3) short learning periods, 4) small sample sizes, and 5) single learning conditions.  
89 The difficulty of interpreting results from this growing literature led Stout and Khreisheh (2015:  
90 870, emphasis original) to call for “studies with sufficient sample sizes to manipulate learning  
91 conditions (e.g. instruction, motivation) and assess individual variation (e.g. performance, psy-  
92 chometrics, neuroanatomy) that *also* have realistic learning periods.” The current study attempts  
93 to strike a viable balance between these demands by investigating early-stage learning of a rela-  
94 tively simple technology (least effort, “Oldowan,” flake production (Reti, 2016; Shea, 2016) under  
95 two instructional conditions while collecting data on individual differences in strength, coordina-  
96 tion, cognition, social learning, self-control, and task engagement. Unlike any previous study, this  
97 allows us to address the likelihood that group effects of training conditions might be impacted by  
98 interactions with individual participant differences in aptitude, motivation, or learning style.

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

## 114 1.1 Individual Differences

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

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

<sup>129</sup> group level results. This is especially true in the relatively small “samples of convenience” typical  
<sup>130</sup> of experimental archaeology.

<sup>131</sup> While testing hypotheses in evolutionary cognitive archaeology remains a considerable challenge  
<sup>132</sup> ([Wynn, 2017](#)), investigation of individual variation in modern research participants represents  
<sup>133</sup> one promising direction. For any particular behavior of archaeological interest, it is expected that  
<sup>134</sup> standing variation in modern populations should remain relevant to normal variation in learning  
<sup>135</sup> aptitude. The presence of trait variation without impact on learning aptitude would provide  
<sup>136</sup> strong evidence against the plausibility of the proposed evolutionary relationship. An absence  
<sup>137</sup> of variation (i.e., past fixation and rigorous developmental canalization) is not expected given  
<sup>138</sup> the known variability of human brains and cognition ([Barrett, 2020](#); [Sherwood & Gómez-Robles,](#)  
<sup>139</sup> [2017](#)). Any confirmatory findings of trait-aptitude correspondence would then have the testable  
<sup>140</sup> implication that humans should be evolutionarily derived along the same dimension (e.g. [Hecht,](#)  
<sup>141</sup> [Gutman, Bradley, et al., 2015](#)).

<sup>142</sup> To date, a small number of “neuroarchaeological” studies have reported associations between  
<sup>143</sup> individual knapping performance and brain structure or physiological responses. Hecht et al.  
<sup>144</sup> ([2015](#)) reported training-related changes in white matter integrity (fractional anisotropy [FA])  
<sup>145</sup> that correlated with individual differences in practice time and striking accuracy change. The  
<sup>146</sup> regional patterning of FA changes also varied across individuals, with only those individuals  
<sup>147</sup> who displayed early increases in FA under the right ventral precentral gyrus (premotor cortex)  
<sup>148</sup> involved in movement planning and guidance) showing striking accuracy improvement over the  
<sup>149</sup> training period. Putt et al. ([2019](#)) similarly found that the proportion of flakes to shatter produced  
<sup>150</sup> by individuals during handaxe making correlated with dorsal precentral gyrus (motor cortex)  
<sup>151</sup> activation. Pargeter et al. ([2020](#)) used a flake prediction paradigm (modeled after [Nonaka et](#)  
<sup>152</sup> [al., 2010](#)) to confirm that striking force and accuracy are important determinants of handaxe-  
<sup>153</sup> making success. These findings all point to the central role of perceptual-motor systems ([Stout &](#)  
<sup>154</sup> [Chaminade, 2007](#)) and coordination ([Roux et al., 1995](#)) in knapping skill acquisition. In addition,  
<sup>155</sup> Putt et al. ([2019](#)) also found successful flake production to be associated with prefrontal (working  
<sup>156</sup> memory/cognitive control) activation and Stout et al. ([2015](#)) found that prefrontal activation  
<sup>157</sup> correlated with success at a strategic judgement (platform selection) task which in turn was  
<sup>158</sup> predictive of success at out-of-scanner handaxe production. Such investigations are thus starting  
<sup>159</sup> to chart out the more specific contributions of different neural systems to particular aspects of

<sup>160</sup> knapping skill acquisition. To date, however, the cognitive/functional interpretation of systems  
<sup>161</sup> identified in this manner has largely relied on informal reverse inference (reasoning backward  
<sup>162</sup> from observed activations to inferred mental processes) from published studies of other tasks  
<sup>163</sup> that activated the same regions, an approach which is widely regarded as problematic (Poldrack,  
<sup>164</sup> 2011).

<sup>165</sup> Here we take a more direct, psychometric approach to measuring individual differences in  
<sup>166</sup> perceptual-motor coordination and cognition. Psychometric instruments (e.g., tasks, question-  
<sup>167</sup> naires) are designed to assess variation in cognitive traits and states, such as fluid intelligence,  
<sup>168</sup> working memory, attention, motivation, and personality, that have been of theoretical interest to  
<sup>169</sup> cognitive archaeologists (e.g., Wynn & Coolidge, 2016). It is thus surprising that they have been  
<sup>170</sup> almost entirely neglected in experimental studies of knapping skill. In the only published example  
<sup>171</sup> we are aware of, Pargeter et al. (2019) reported significant effects of variation in planning and  
<sup>172</sup> problem solving (Tower of London test (Shallice et al., 1982)) and cognitive set shifting (Wisconsin  
<sup>173</sup> Card Sort test (Grant & Berg, 1948)) on early stage handaxe learning. Of course, cognition is not  
<sup>174</sup> the only thing that can affect knapping performance. Flake prediction experiments highlight the  
<sup>175</sup> importance of regulating movement speed/accuracy trade-offs (Nonaka et al., 2010; Pargeter et  
<sup>176</sup> al., 2020) and studies of muscle recruitment (Marzke et al., 1998) and manual pressure (Key & Dun-  
<sup>177</sup> more, 2018; Williams-Hatala et al., 2018) during knapping highlight basic strength requirements.  
<sup>178</sup> Along these lines, Key and Lycett (2019) found that individual differences in hand size, shape, and  
<sup>179</sup> especially grip strength were better predictors of force loading during stone tool use than were  
<sup>180</sup> attributes of the tools themselves. However, we are unaware of any such studies of biometric  
<sup>181</sup> influences on variation in knapping success. Finally, the time and effort demands of knapping  
<sup>182</sup> skill acquisition suggest that differences in personality (e.g., self-control and “grit” (Pargeter et  
<sup>183</sup> al., 2019), motivation (Stout, 2002), and social vs. individual learning strategies (Miu et al., 2020)  
<sup>184</sup> might also affect learning outcomes. We are again unaware of any previous studies that have  
<sup>185</sup> assessed such effects. In this study, we assessed all participants with a battery of tests including  
<sup>186</sup> grip strength, movement speed/accuracy, spatial working memory, fluid intelligence, self-control,  
<sup>187</sup> tendency to use social information, and motivation/engagement with the tool making task. We  
<sup>188</sup> were particularly interested in the possibility that these variables might not only impact learning  
<sup>189</sup> generally, but might also have different effects under different learning conditions.

190 **1.2 Teaching, Language, and Tool Making**

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

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

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

216 Such concerns are exacerbated in experiments using pre-recorded instructional videos or ex-  
217 tremely short training periods. Video does not allow the interactive teaching that is favored even  
218 in formal academic knapping classes (e.g., [Shea, 2015](#)) and is almost certainly typical of traditional  
219 learning contexts (e.g., [Stout, 2002](#)). It is not known how video presentation affects the efficacy of

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

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

252 et al., 2019).

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

266 **1.3 Raw materials and knapping skill**

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

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

281 To address this, some studies have attempted to more closely match experimental and archaeo-

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

## 2 Materials and Methods

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

### 2.1 Participants

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

<sup>311</sup> **2.2 Study Visit**

<sup>312</sup> Participants were asked to visit the Paleolithic Technology Lab at Emory University to complete  
<sup>313</sup> one three-hour session. Participants were scheduled to attend in six groups of four, however one  
<sup>314</sup> of these groups had only three participants due to a no-show on the day of the experiment. Each  
<sup>315</sup> visit began with the collection of individual differences measures, which took approximately one  
<sup>316</sup> hour. After that, participants undertook 105 minutes (two hours minus a 15-minute break after 1  
<sup>317</sup> hour) of stone tool making practice. This session was video-recorded, and all lithic products were  
<sup>318</sup> collected. After the tool making task, participants completed an “exit questionnaire” comprising  
<sup>319</sup> the Intrinsic Motivation Inventory (see below).

<sup>320</sup> Participants were compensated for their time with a \$30 gift card. They also had the opportunity  
<sup>321</sup> to earn a performance bonus of \$5, \$10, \$15 or \$20 on the gift card. They were told that this  
<sup>322</sup> bonus would depend on “how well they did” on the last core of their practice session. The actual  
<sup>323</sup> performance measure was not specified, but in order to allow on the spot payment a simple  
<sup>324</sup> measure of the percentage of starting weight removed from the final core was used such that: >  
<sup>325</sup> 30% earned \$5, > 40% earned \$10, > 50% earned \$15, > 75% earned \$20.

<sup>326</sup> **2.3 Individual Difference Measures**

<sup>327</sup> We used five individual difference measures for this study:

<sup>328</sup> 1) Grip strength was measured in kilograms using an electronic hand dynamometer (Camry  
<sup>329</sup> EH101). Strength was measured twice and the higher value recorded. Grip strength is a  
<sup>330</sup> simple measure that is well correlated with overall muscular strength (Wind et al., 2010) and  
<sup>331</sup> a range of other health and fitness measures (Sasaki et al., 2007). It is hypothesized to be  
<sup>332</sup> relevant to generating kinetic energy for fracture initiation (Nonaka et al., 2010) as well as  
<sup>333</sup> control and support of the hammerstone (Williams-Hatala et al., 2018) and core (Faisal et  
<sup>334</sup> al., 2010; Key & Dunmore, 2015).

<sup>335</sup> 2) Motor accuracy was assessed using a “Fitts Law” reciprocal tapping task. Fitts Law describes  
<sup>336</sup> the trade-off between speed and accuracy in human movement, classically measured  
<sup>337</sup> by tapping back and forth between two targets of varying size and spacing (Fitts, 1954).  
<sup>338</sup> Archaeologists have proposed (Pargeter et al., 2020; Stout, 2002) that management of this  
<sup>339</sup> trade-off is critical to the accurate application of appropriate force seen in skilled knapping

340 (Nonaka et al., 2010; Roux et al., 1995). We implemented this test on a Surface Pro tablet  
341 running free software (FittsStudy Version 4.2.8, default settings) developed by the Accessible  
342 Computing Experiences lab (Jacob O. Wobbrock, director) at the University of Washington  
343 ([depts.washington.edu/acelab/proj/fittsstudy/index.html](http://depts.washington.edu/acelab/proj/fittsstudy/index.html)). Participants use a touchscreen  
344 pen to tap between ribbons on the screen, with average movement time as the performance  
345 metric.

- 346 3) Visuospatial working memory is the capacity to “hold in mind,” which researchers have  
347 hypothesized to be important in stone toolmaking performance (Coolidge & Wynn, 2005). It  
348 also might support a learning process known as ‘chunking,’ in which multiple items or operations  
349 are combined into summary chunks stored in long term memory, that is thought to be  
350 important in the acquisition of knapping and other skills (Pargeter et al., 2019). We measured  
351 visuospatial working memory using a free n-back task ([wmp.education.uci.edu/software/](http://wmp.education.uci.edu/software/))  
352 developed by the Working Memory and Plasticity Laboratory at the University of California,  
353 Irvine (Susanne Jaeggi, PI) and implemented in E-Prime software on a desktop computer.  
354 In this task, participants are asked to remember the position of blue squares presented  
355 sequentially on the screen and touch a key when the current position matches that 1, 2,  
356 3...n iterations back. Progression to blocks with increasing values of n is contingent on  
357 exceeding a threshold success rate. Performance was measured as the highest n achieved.
- 358 4) Fluid intelligence (Cattell, 1963) refers to the capacity to engage in abstract reasoning and  
359 problem solving in a way that is minimally dependent on prior experience. It complements  
360 “crystallized intelligence” (the ability to apply learned procedures and knowledge) as one of  
361 the two factors (gf, gc) comprising so-called “general intelligence” (g). Fluid intelligence is  
362 closely related to the executive control of attention and manipulation of information held  
363 in working memory (Engle, 2018)(Engle 2018). It is hypothesized to support technological  
364 innovation (Coolidge & Wynn, 2005) and/or the intentional learning of new skills (Stout &  
365 Khreichsheh, 2015; Unsworth & Engle, 2005). We measured fluid intelligence using the short  
366 version (Bilker et al., 2012) of the classic Raven Progressive Matrices task, which requires  
367 participants to complete increasingly difficult pattern matching questions.
- 368 5) The use of social information for learning and decision making varies across individuals  
369 and societies (Molleman et al., 2019). Such variation is a key topic for understanding social  
370 learning and cultural evolutionary processes (Heyes, 2018; Kendal et al., 2018; Miu et al.,

371        2020) and represents a potential confound for assessing experimental effects of different  
372        social learning conditions. We measured participants' tendency to rely on social information  
373        vs. their own insights using the Berlin Estimate AdjuStment Task (BEAST) developed by  
374        Molleman et al. (2019). In this task, participants are present with large arrays of items on  
375        a screen and asked to estimate the number present. They are then provided with another  
376        person's estimate and allowed to provide a second estimate. The participants' average  
377        adjustment between first and second estimates provides a measure of their propensity to  
378        rely on social information.

379        **2.4 Stone Tool Making**

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



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.

393 **2.4.1 Raw Materials**

394 Each participant was provided with 9 cores for use over the 2-hour experiment. These cores were  
395 produced from larger chunks of a fine-grained basalt purchased from neolithics.com by fracturing  
396 them with a sledgehammer. This basalt has not been mechanically compared (e.g., rebound  
397 hardness, [Braun et al., 2009](#)) to East African basalts, but appears qualitatively similar to finer  
398 grained examples from sites around Lake Turkana (D. Braun, pers. comm.) and at Gona. Spalling  
399 produced irregular, angular chunks (**Figure 2**) for use in the experiment, weighing between 459g -  
400 1876g (mean = 975g). All cores were weighed, measured (Length, Width, Thickness), and painted  
401 white so that new fracture surfaces could be discriminated from those created during production.  
402 Cores were sorted by shape and weight and then distributed evenly to each participant. As a

403 result, there were no significant difference across participants in the mean weight (ANOVA, df  
404 = 22, F=0.3, p = 0.9; Levene test of homogeneity of variance = 1.04, df1=22, df2 = 184, p = 0.4) or  
405 shape (Length × Width/Thickness: ANOVA, df = 22, F=0.4, p = 0.9; Levene statistic = .6, df1=22,  
406 df2 = 184, p = 0.9) of cores provided. This was also true for cores provided to participants across  
407 the two experimental conditions (Taught vs. Untaught mean weight = 1001g vs. 956g, t = 1.24, df =  
408 205, p = 0.2, Levene's Test F = 0.6, p = 0.4; mean shape = 221.43 vs. 221.45, t = -0.003, df = 205, p =  
409 0.9, Levene's Test F = 3.8, p= 0.05). Participants were, however, allowed to choose which cores to  
410 work on so that differences in the weight and shape of cores actually used across participants and  
411 conditions could still emerge as a result of selection bias.

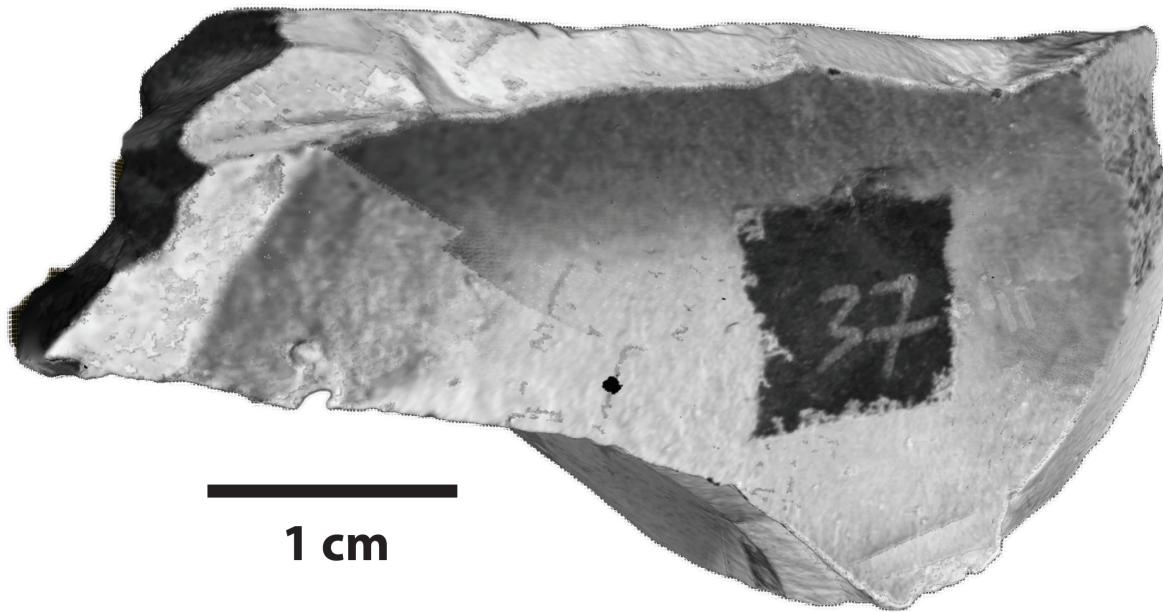


Figure 2: 3D scan of a basalt core prior to knapping. Cores were spray painted white to facilitate subsequent technological analyses of the flake scars and flaking intensity. Scale in the background is in 10mm blocks.

412 Sixty pounds of 3-to-5 inch basalt “Mexican Beach Pebbles” were purchased from a landscaping  
413 supply company for use as hammerstones in the experiment. Of these, 90 were selected as  
414 suitable for use. These weighed between 213g-1360g (mean = 425) and varied in elongation (L/W  
415 = 1.01 to 2.65) and relative thickness (LxW/T = 90.48 to 283.67). Forty-five stones were placed  
416 in the middle of the knapping area (Figure 2) for participants to freely choose from during the  
417 experiment. Broken hammerstones were replaced from the reserve to maintain a consistent  
418 number and range of choices. Each hammerstone was numbered and participants’ choices were

<sup>419</sup> recorded along with the number of the core(s) being worked on with a particular hammerstone.

<sup>420</sup> **2.4.2 Experimental Conditions**

<sup>421</sup> In both conditions, three researchers were present to record activities and collect materials.  
<sup>422</sup> Participants were seated in a circle ([Shea, 2015](#)) and experiments were video recorded using two  
<sup>423</sup> cameras. Participants were free to select hammerstones from the common pile and to work on  
<sup>424</sup> any or all of their nine assigned cores in any order they preferred ([Figure 3](#)). However, each core  
<sup>425</sup> and all associated debitage were collected before participants were allowed to start working on a  
<sup>426</sup> new core, so it was not possible to partially work and then return to a particular core later. The  
<sup>427</sup> order of cores used and associated hammerstones were recorded for each participant during the  
<sup>428</sup> experiment.



Figure 3: Subjects selecting hammerstones.

429 In the untaught condition, a researcher (DS) sat with the participants and made stone tools  
430 but remained silent and made no effort to facilitate learning (e.g., through gesture, modified  
431 performance, facial expression, attention direction, or verbal instruction). Over the 2-hour  
432 period, the researcher completely reduced four cores (one every ~30 minutes). Participants

433 were not restricted from talking to each other, as this would create an unnatural and potentially  
434 stressful social context that might affect learning. Participants were asked to avoid any form of  
435 communication about the tool making task specifically, and they complied with this request.  
436 Participants in this condition thus had the opportunity to observe tool making (**Figure 4**) by  
437 an expert and/or by other learners, should they choose to do so, but received no intentional  
438 instruction.



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

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

447 **2.5 Lithic Analysis**

448 All finished cores were weighed and measured (L, W, T). Delta weight was calculated as (Start  
449 weight-End weight)/Start weight. All detached pieces (DPs) were collected and weighed. We  
450 did not sort DPs into types (e.g., whole flakes, fragments) as this would have greatly increased  
451 processing time and it is not clear that such distinctions add relevant information regarding  
452 utility/desirability beyond that supplied by metrics ([Stout et al., 2019](#)). All DPs larger than 40mm  
453 in maximum dimension were photographed and measured. It is conventional in Early Stone  
454 Age lithic analysis to employ a 20 mm cut-off. We selected a higher threshold for both pragmatic  
455 (analysis time) and theoretical reasons. Flake use experiments have shown that flakes weighing  
456 less than 5–10 g or with a surface area below 7–10 cm<sup>2</sup> ([Prasciunas, 2007](#)) or with a maximum  
457 dimension <50–60 mm ([Key & Lycett, 2014](#)) become markedly inefficient for basic cutting tasks.  
458 Similarly, data from Oldowan replication experiments ([Stout et al., 2019](#)) show that the utility  
459 index (flake cutting edge/flake mass<sup>1/3</sup>) \* (1 - exp[-0.31 \* (flake maximum dimension – 1.81)])  
460 developed by Morgan et al. (2015) falls off rapidly below 40mm maximum dimension ( Mean  
461 Utility < 40mm = 0.508; >=40mm = 0.946; t= 11.99, df = 707, p < 0.000). By including weight in our  
462 cut-off criteria we also avoid skewing the flake shape distribution by selectively retaining long,  
463 thin pieces (i.e., MD > 40, weight < 5g) while discarding rounder pieces of similar (or greater)  
464 weight and area.

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

473 Photographs were post-processed using Equalight software to adjust for lens and lighting falloff  
474 that result from bending light through a lens and its aperture which can affect measurements  
475 taken from photographs. Each image was shot with a scale that was then used to rectify the  
476 photograph's pixel scale to a real-world measurement scale in Adobe Photoshop. Images were  
477 converted to binary black and white format and silhouettes of the tools were extracted in Adobe

478 Photoshop. We then used a custom ImageJ ([Rueden et al., 2017](#)) script ([Pargeter et al., 2019](#)) to  
479 measure DP length and take nine width measurements at 10% increments of length starting at  
480 the base of each DP. We used the built-in ImageJ tool to measure DP area. A “Proportion Larger  
481 DPs” was calculated per core as the combined weight of all DPs >40mm in maximum dimension  
482 and 5g in weight divided by the weight of all DPs. Higher values show cores with proportionally  
483 more large DPs.

484 **2.6 Statistical Analyses**

485 To evaluate the association between psychometric, motor-skill, and training measures and tech-  
486 nological outcomes, we adopted an information-theoretic approach ([Burnham & Anderson,](#)  
487 [2002](#)). Information-theoretic approaches provide methods for model selection using all possible  
488 combinations of variables while avoiding problems associated with significance-threshold step-  
489 wise selection. We used the corrected Akaike information criterion (AICc) to rate each possible  
490 combination of predictors on the balance between goodness of fit (likelihood of the data given  
491 the model) and parsimony (number of parameters). The AICc consists of the log likelihood (i.e.,  
492 how well does the model fit the data?) and a penalty term for the number of parameters that  
493 must be estimated in the model (i.e., how parsimonious is the model?), with a correction for small  
494 sample sizes (AICc converges to the standard AIC at large samples). A lower AICc indicates a  
495 more generalizable model and we used it to compare and rank various possible models. Each  
496 analysis begins with a full model that includes all predictors of interest. All possible combinations  
497 of predictors are then fit, and the resulting models are ranked and weighted based on their AICc.  
498 The “best” model is chosen because it has the lowest AICc score.

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

502 **3 Results**

503 Following a recent protocol to enhance the reproducibility and data transparency of archaeo-  
504 logical research ([Marwick, 2017](#)), detailed results of all analyses and assessments of the data  
505 structure are available in our paper’s supplementary materials and through Github ([https:](https://)

506 //github.com/Raylc/PaST-pilot). Here we limit discussion to the major findings regarding  
507 flaking performance and individual differences. We were particularly interested in: 1) group  
508 level effects of experimental condition, 2) individual differences in aptitude and learning, and 3)  
509 potential interactions between learning conditions and individual differences. To address these  
510 questions, we employed data reduction (Principal Component Analysis) to derive two summary  
511 metrics of flaking performance, compared these factors across the two experimental condi-  
512 tions, and built multivariate models examining the relations between our various psychometric  
513 measures, subject's motor skill scores, and our two lithic performance factors.

514 **3.1 Principal Component analyses**

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

517 **3.1.1 Detached Piece size and shape**

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

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

533 **3.1.2 Lithic flaking performance measures**

534 To better understand the relationship between our various lithic performance measurements and  
535 to reduce data dimensionality, we conducted a second principal component analysis examining  
536 the study's six lithic performance measures (count of large pieces [ $>40\text{mm}$  and  $5\text{g}$ ], mass of large  
537 pieces relative to total detached mass, core delta mass, and the three flake shape factors). All of  
538 these measures were summarized for each core and unique factor scores were calculated from  
539 these core-specific measures. Bartlett's Test of Sphericity was significant ( $\chi^2 (6) = 3185$ ,  $p < .01$ )  
540 indicating that the set of variables are at least adequately related for factor analysis.

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

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

556 **3.2 Relationships between Performance Measures**

557 This approach also allowed us to compare the relationship between Total Productivity, Quantity,  
558 and Quality across our two experimental groups ([Figure 5](#) and [Figure 6](#)). As might be expected,  
559 we found that per-core Quantity and Total Productivity are positively correlated in both groups  
560 ([Figure 5a](#)), although this relationship is twice as strong in the trained ( $F[1, 9] = 33$ ,  $p < 0.01$ ,  
561 Adj.  $R^2 = 0.8$ ) compared to untrained ( $F[1, 8] = 8$ ,  $p = 0.02$ , Adj.  $R^2 = 0.4$ ) group. Interestingly,  
562 we also found evidence of a negative correlation between Total Productivity and Quality in the

563 untrained group ( $F[1, 8] = 28$ ,  $p = <0.01$ , Adj.  $R^2 = 0.7$ ), but no relation in the Trained group (**Figure**  
 564 **5b**). A qualitatively similar trend with respect to Quality vs. Quantity (**Figure 5c**) did not achieve  
 565 significance ( $F[1, 17] = 0.6$ ,  $p = 0.2$ ).

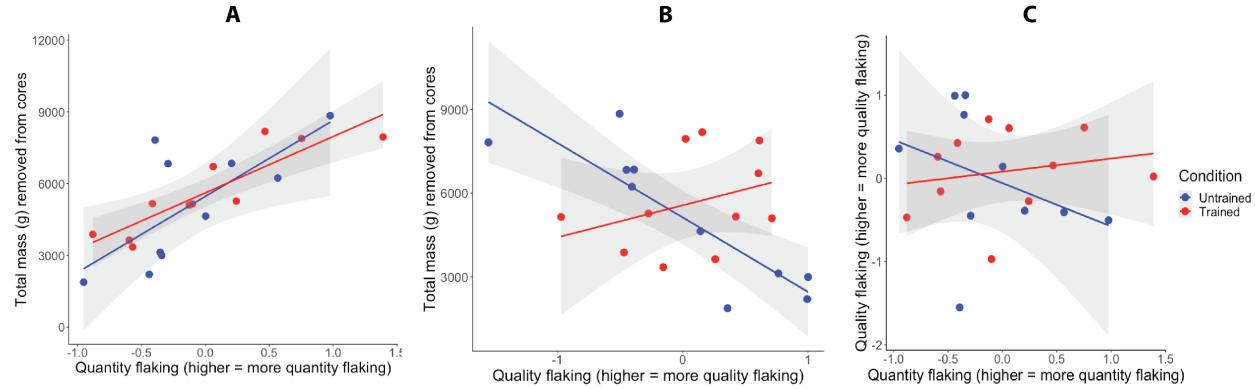


Figure 5: Relationships between flaking quantity, quality, productivity, and the two training conditions. Each dot represents a participant, colors represent training conditions.

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

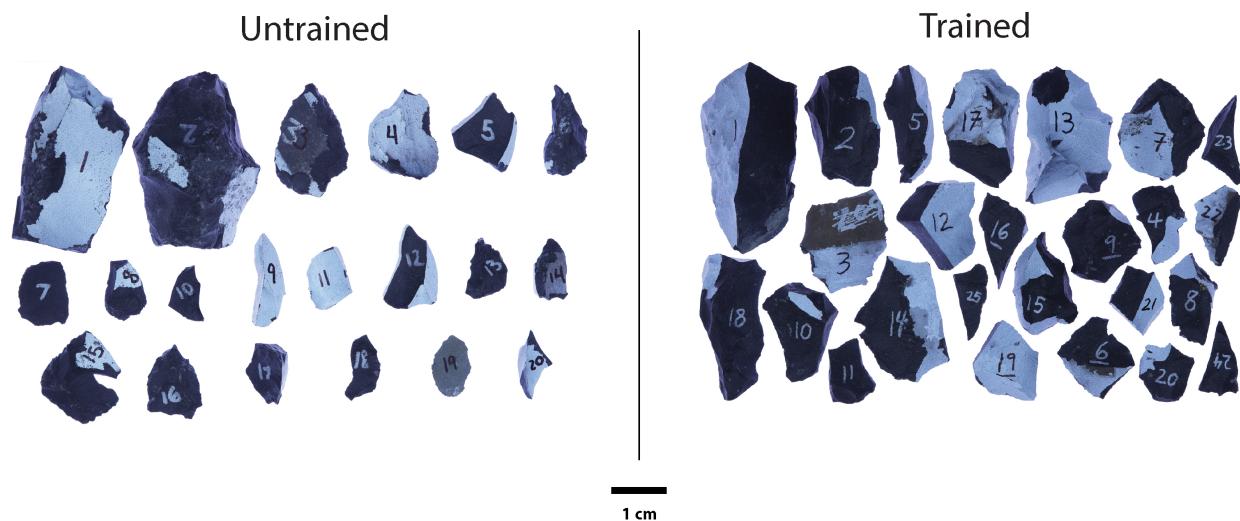


Figure 6: Comparison of untrained and trained detached pieces illustrating the ability of Trained participants to maximize Quality and Productivity at the same time.

571 **3.3 Do trained, untrained, and expert knappers perform differently?**

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

575 Table summarizes the results of ANOVA tests group level difference in central tendency on various  
576 performance measures. We found no significant differences between the trained and untrained  
577 groups on our flaking Quantity and Quality factors. In contrast, three-way flake size and shape  
578 comparisons between our expert knapper and the two novice groups showed that the expert  
579 knapper made significantly more large flakes (effect size = 0.14), had a significantly higher core  
580 delta mass signal than either of the novice groups (effect size = 0.26), and left significantly smaller  
581 finished cores (effect size = 0.27) (**Figure 7**). All three of these results show either medium or large  
582 effect sizes. In all three comparisons, the trained group's data distributions tended towards the  
583 expert sample although they were not significantly different from the untrained group (**Figure**  
584 **8**). We also observed a significant difference in shape factor 2 (splinters) driven by the expert's  
585 lower values, but with a very low (<0.01) effect size. These results show that mean core reduction  
586 intensity and large flake production rates distinguish expert and novice performance whereas  
587 novices in experimental groups produced pieces of similar mean size and shape as those of the  
588 expert trainer.

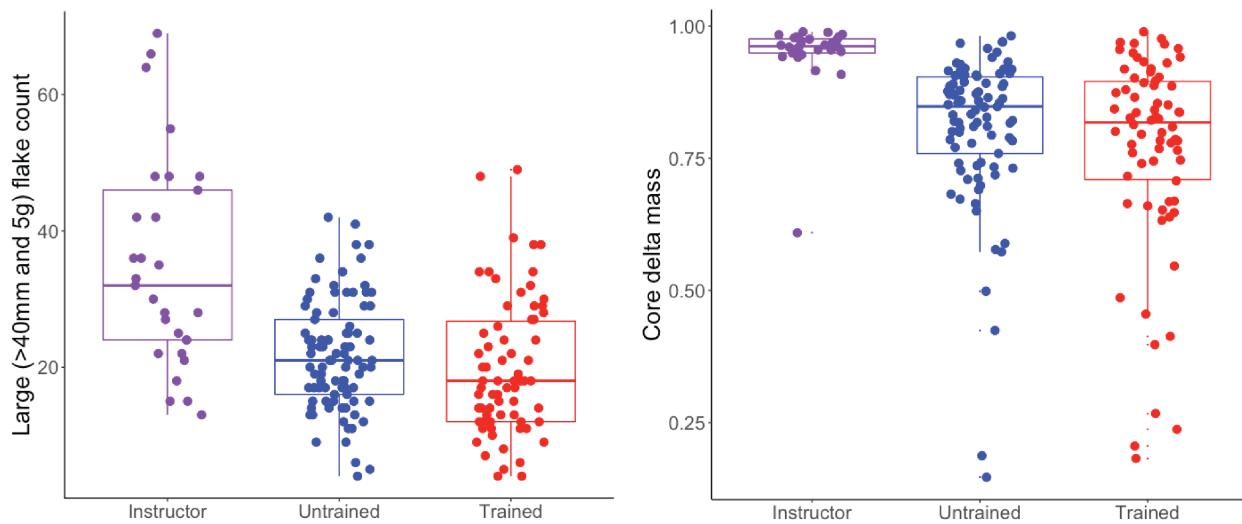


Figure 7: Results showing significant differences between the instructor (expert) and novice flaking performance.

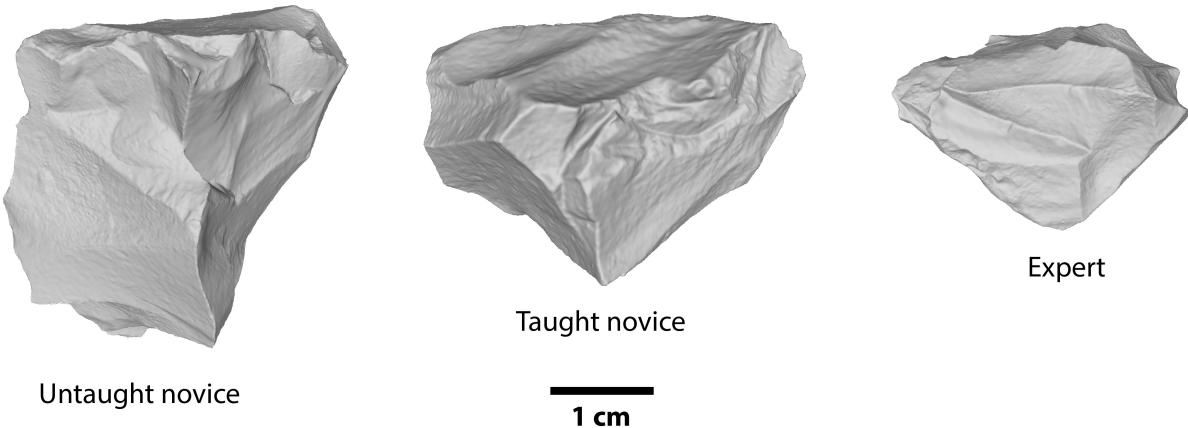


Figure 8: Comparison of untaught, taught, and expert cores.

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

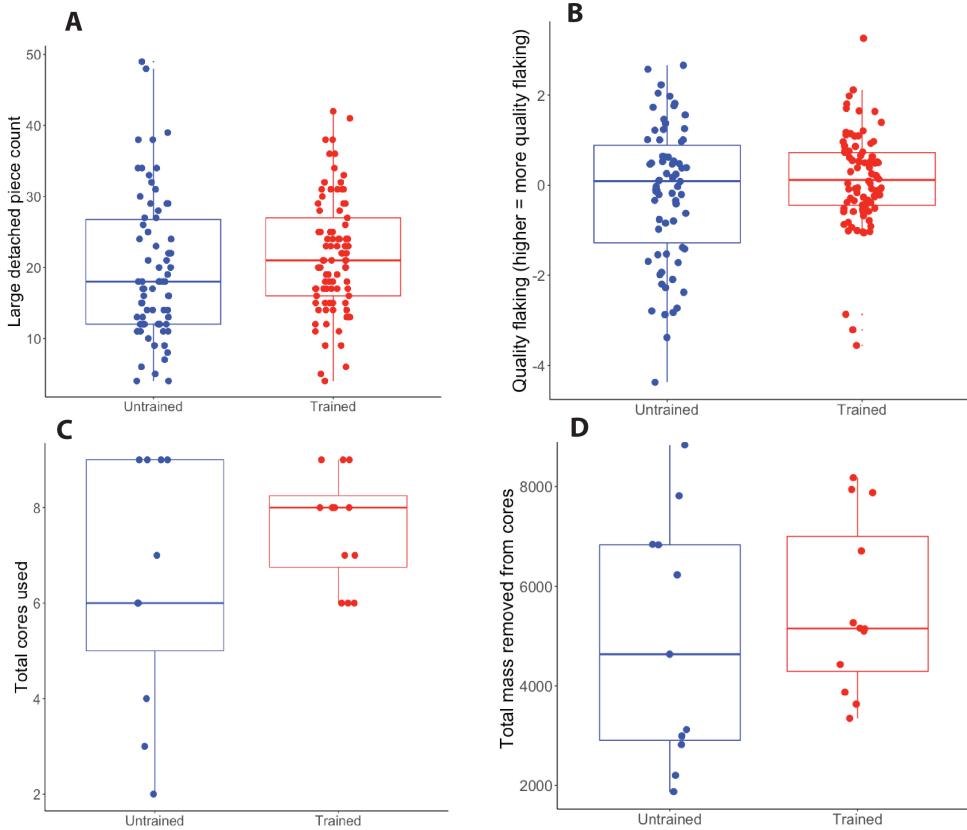


Figure 9: Variance comparisons between trained and untrained individuals across various flaking performance metrics. Note the lower variance of the trained group in all the plots.

### 599 3.4 Does performance change over time?

600 In addition to comparing overall performance during the two hour experiment, we also wanted to  
 601 determine if groups or individuals differed in learning (i.e., performance change) over the period.  
 602 For these analyses, we calculated the learning stage as the ordinal number of each core out of the  
 603 total number knapped by each subject (i.e., core 2 of 4 or 4 of 8 both equal 50% complete). These  
 604 relative core use-order percentages were then binned into 20 percent brackets for core-order  
 605 and group-level comparisons. Flaking outcomes were tracked using the two performance factors  
 606 (Quality and Quantity). We added the nodule starting mass to track whether training/practice  
 607 times impacted raw material selection.  
 608 Table shows no significant training effects across the two performance measures either as grouped  
 609 data or between individuals ([Figure 10](#) and [Figure 11](#)). This result demonstrated that flaking  
 610 outcomes did not change dramatically across the study interval. This lack of significant learning

611 effects is confirmed by an inspection of individual learning curves (Figure). The one significant  
 612 main training effect related to core starting mass (with a strong effect size = 0.25). On average,  
 613 core starting masses start low and increase, showing that participants selected smaller nodules  
 614 first. As the smaller nodules in their allotment were depleted, participants were left to knap larger,  
 615 less preferred nodules. This preference for smaller cores is somewhat less pronounced in the  
 616 untrained group, as indicated by a small main effect of learning condition and generally higher  
 617 starting nodule masses for the untrained group (Figure??).

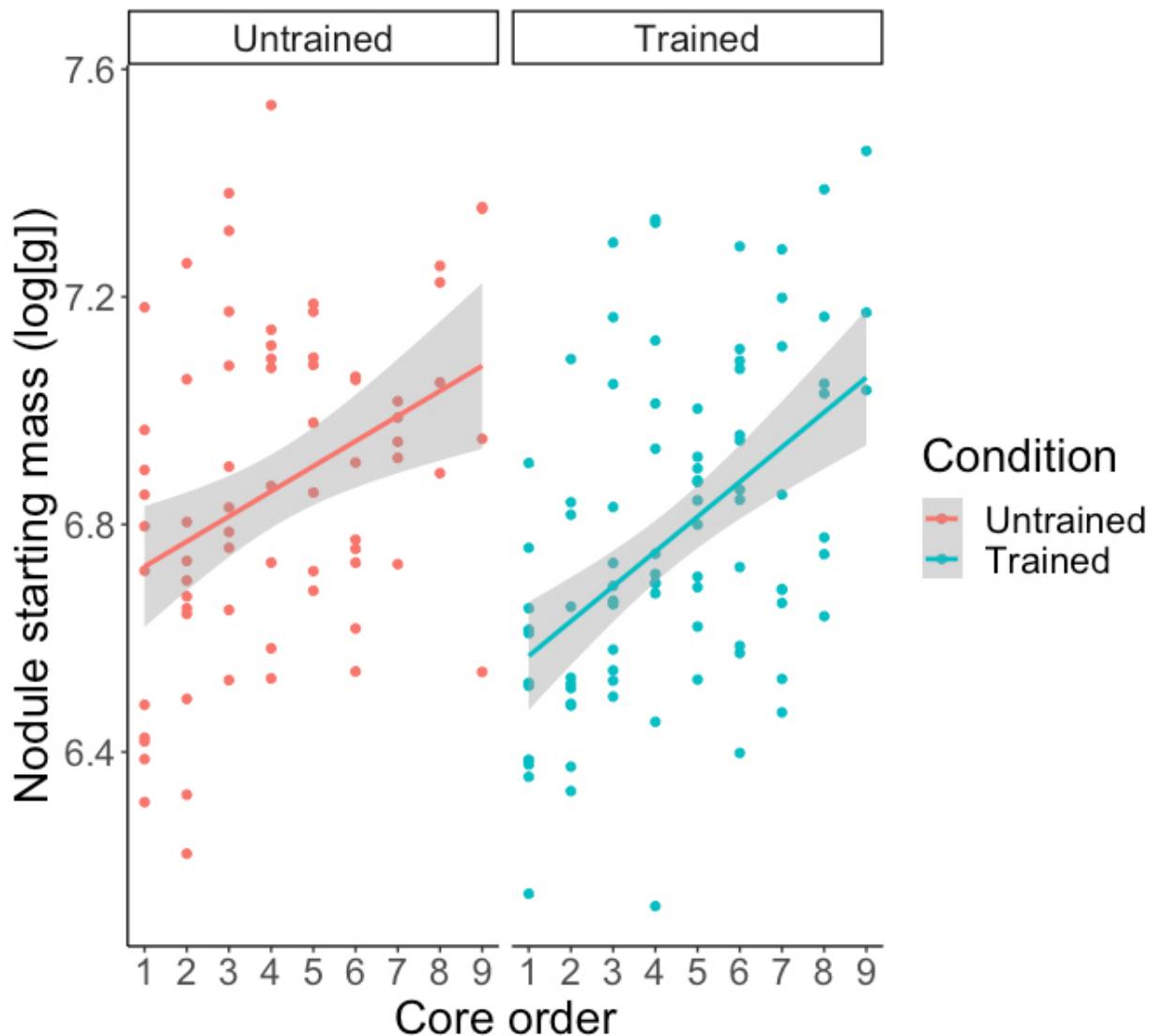


Figure 10: Comparisons of nodule starting mass across the study period by training condition. Results show a significant relationship between nodule starting mass and length of time in the study.

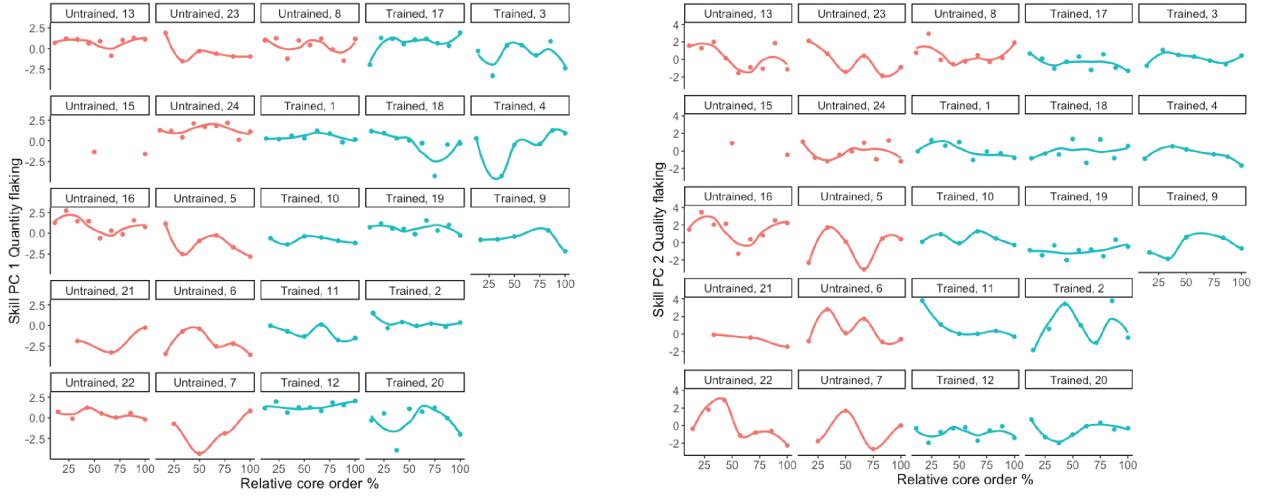


Figure 11: Individual learning curves derived from each subject's relative core use order and the two flake performance factors.

### 618 3.5 Do individual differences in motor skill and psychometric measures predict 619 flaking performance?

620 One of the experiment's primary goals was to test if measures of individual perceptual-motor  
621 and cognitive variation predict success in stone flaking across different training conditions. To  
622 address this goal, we built three multivariate models examining the relations between training  
623 conditions, individual difference measures, and our three lithic performance measures (overall  
624 productivity and average per-core Quantity and Quality). These models enabled us to determine  
625 which of the psychometric and motor skill factors are better predictors of a participant's flaking  
626 performance in the study.

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

636 The full models were fitted with the lm function in R 3.2.3, and we used the Glmulti package's

637 automated model selection algorithm to select the best performing model (lowest AICc score)  
638 (see methods for further details on the multimodal selection process). All three models follow the  
639 same complete model statement as follows:

640 *Flaking performance variable* ~ Training condition + Highest n-back level + Raven's Progressive  
641 Matrix score + BEAST score + Fitt's score + Grip strength

642 For our two per-core performance factors (Quantity and Quality) it is also relevant to consider how  
643 individual core features may have affected performance. We found no evidence of individual or  
644 group level practice effects over the two hours, so we did not include core order in the models. We  
645 did, however, find that subjects selected progressively larger nodules throughout the experiment.  
646 It is thus important to understand whether nodule variability had any impact on our flaking  
647 results. Because starting nodule size (mass) and shape were strongly correlated ( $F [1,157] = 186$ , p  
648  $< 0.01$ ,  $R^2 = 0.54$ ) we included nodule mass as a covariate to control for any variance in flaking  
649 performance that may be driven by nodule differences.

650 **3.5.1 Model 1: Individual differences and overall productivity**

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

657 *Total flaked mass* ~ Training condition + Grip strength + RPM × Highest n-back level + Fitts  
658 score × BEAST score + Grip strength × RPM

659 This model explains a statistically significant and substantial proportion of variance in flaking  
660 productivity ( $R^2 = 0.84$ ,  $F (6, 14) = 12.7$ ,  $p < 0.01$ , adj.  $R^2 = 0.77$ ). A model residuals normality  
661 test shows no significant differences with the normal distribution ( $p = 0.72$ ) indicating that this  
662 relationship is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (BP = 2, df  
663 = 6,  $p = 0.8$ ).

664 Table presents this model's coefficients and summary outputs, wherein baseline refers to the un-  
665 trained condition with all continuous predictors at the sample average. The parameter estimates

666 for the continuous predictors reflect the expected change in utility for 1 standard deviation change  
 667 in the predictor variable. We found significant ( $p < 0.05$ ) and substantial (Standardized Estimate  
 668  $\geq 0.50$ , i.e. a 50% change in variable) main effects of Grip Strength, Visuospatial nBack, and  
 669 BEAST. The main effect of Grip Strength (Figure 12), irrespective of learning condition, indicates  
 670 the basic importance of strength in generating higher production rates among naive knappers at  
 671 least when efficiency and quality are not considered.

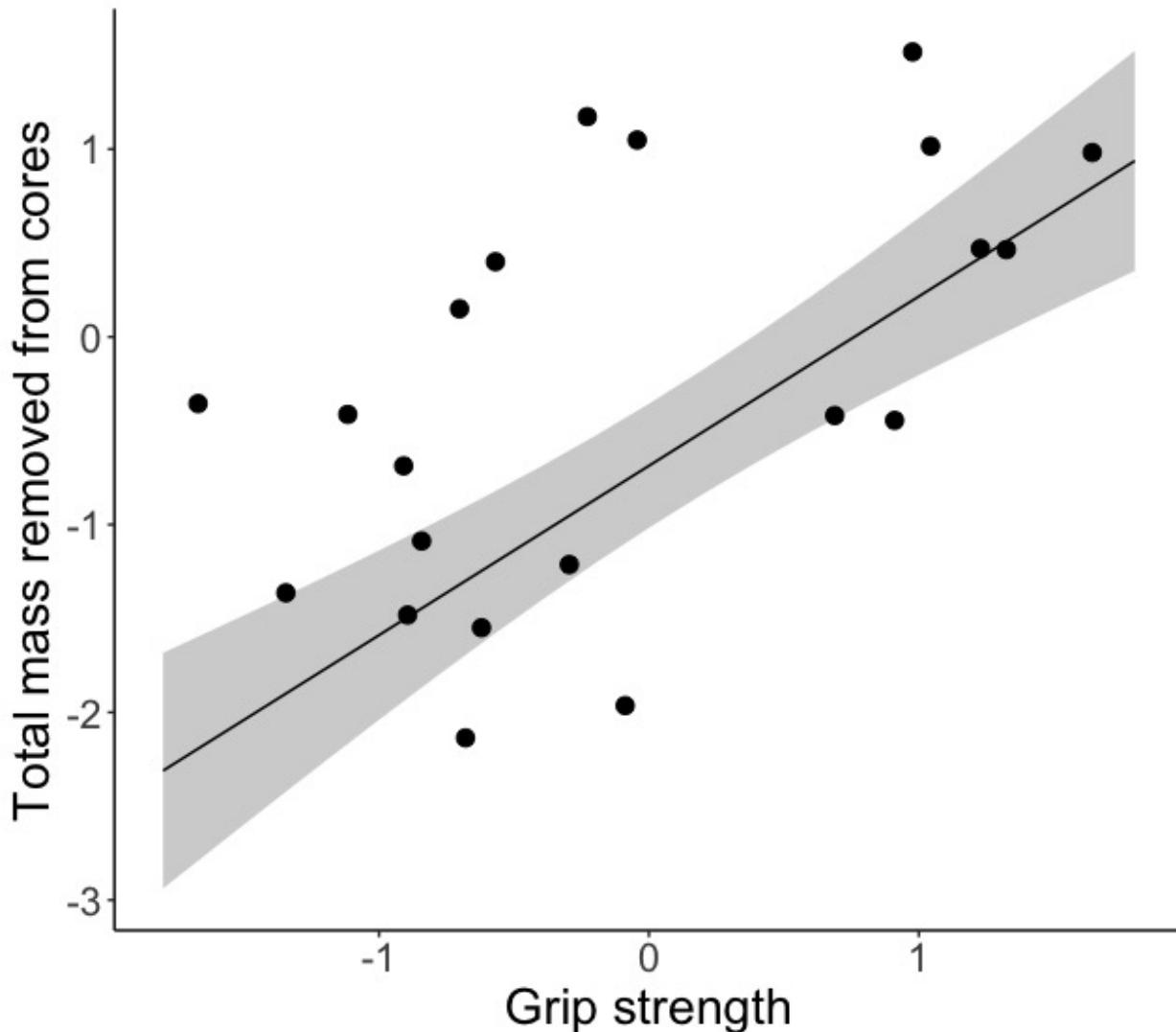


Figure 12: Significant main effect of grip strength on overall flaking productivity.

672 Effects of visuospatial working memory capacity and social information use are more complicated,  
 673 as indicated by strong interactions with learning condition (Figure 13). In each case, higher scores  
 674 were associated with better performance in the uninstructed group but worse performance in

675 the instructed group. Positive effects in the uninstructed group were as expected, given the  
 676 hypothesized importance of spatial cognition (Coolidge and Wynn 2005) and social learning  
 677 (Morgan et al. 2015) in the acquisition of knapping skills. Negative effects in the trained group are  
 678 unexpected but presumably reflect differences in learning strategies adopted under instruction.

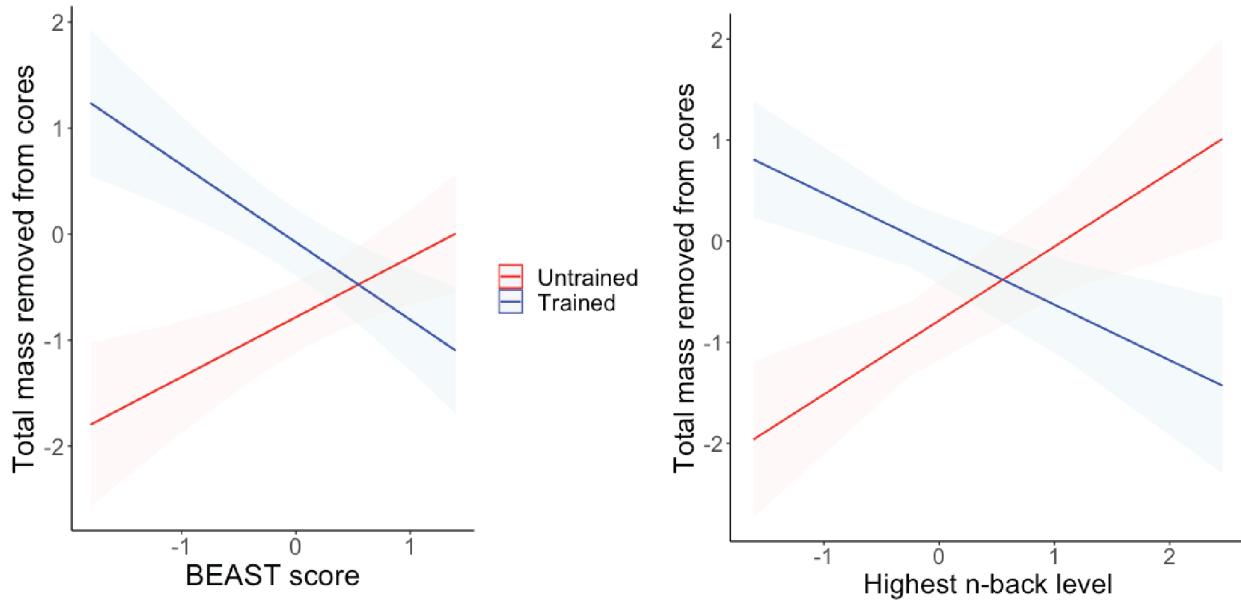


Figure 13: Two significant interaction effects of training condition, social information, and visuo-spatial working memory on overall flaking productivity.

### 679 3.5.2 Model 2: Individual differences and quality flaking

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

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

688 This model explains a statistically significant and substantial proportion of variance in quantity  
 689 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  
 690 significant differences with the normal distribution ( $p = 0.38$ ) indicating that this relationship  
 691 (as required) is linear. A Breusch-Pagan test showed no evidence for heteroskedasticity (whether

692 variance for all observations in our data set are the same) ( $\text{BP} = 4.4$ ,  $\text{df} = 8$ ,  $p = 0.8$ ).  
 693 The Quantity model roughly paralleled results for Total Production, yielding substantial and  
 694 significant interactions between training condition, n-back level, BEAST scores, and grip strength.  
 695 As with Total Production, higher visuospatial n-back levels and BEAST scores were associated with  
 696 lower Quantity scores in the trained group but higher or unchanged Quantity in the untrained  
 697 group (**Figure 14**).

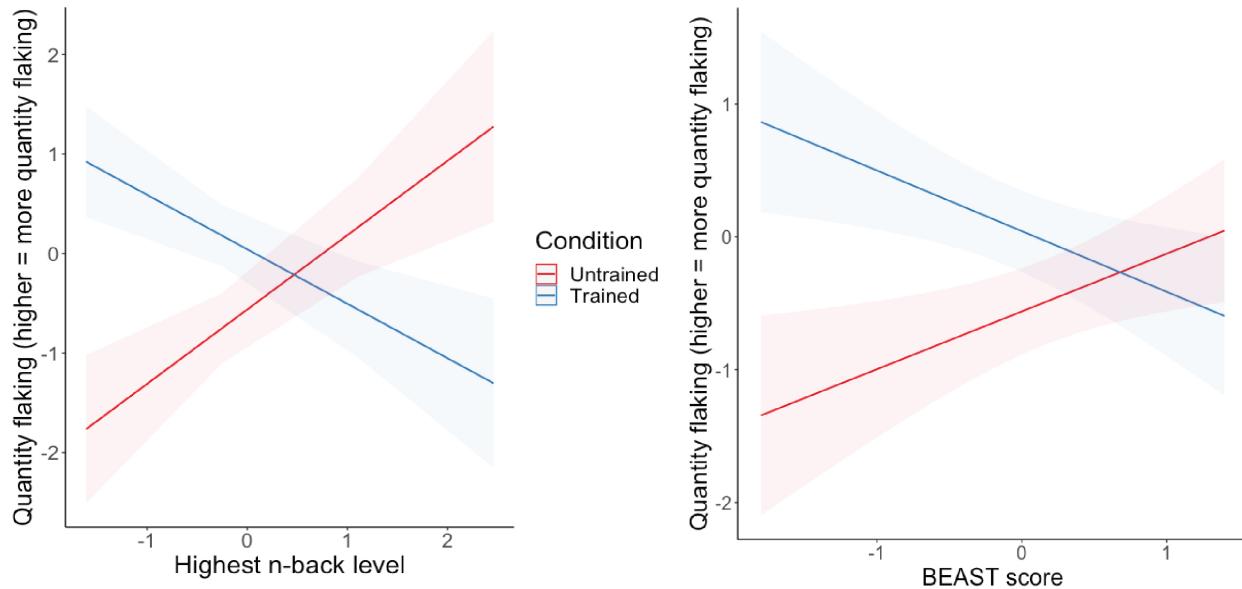


Figure 14: Two significant interaction effects of training condition, visuo-spatial working memory, and social information, on overall flaking quality.

698 Unlike Total Productivity, the effect of Grip Strength on per-core Quantity was mediated by an  
 699 interaction with learning condition (**Figure 15**). Thus, high Grip Strength enabled individuals  
 700 in both groups to produce more total debitage, but only Instructed individuals translated Grip  
 701 Strength into more intense reduction of individual cores, including not only delta mass, but also  
 702 number and proportion of larger pieces.

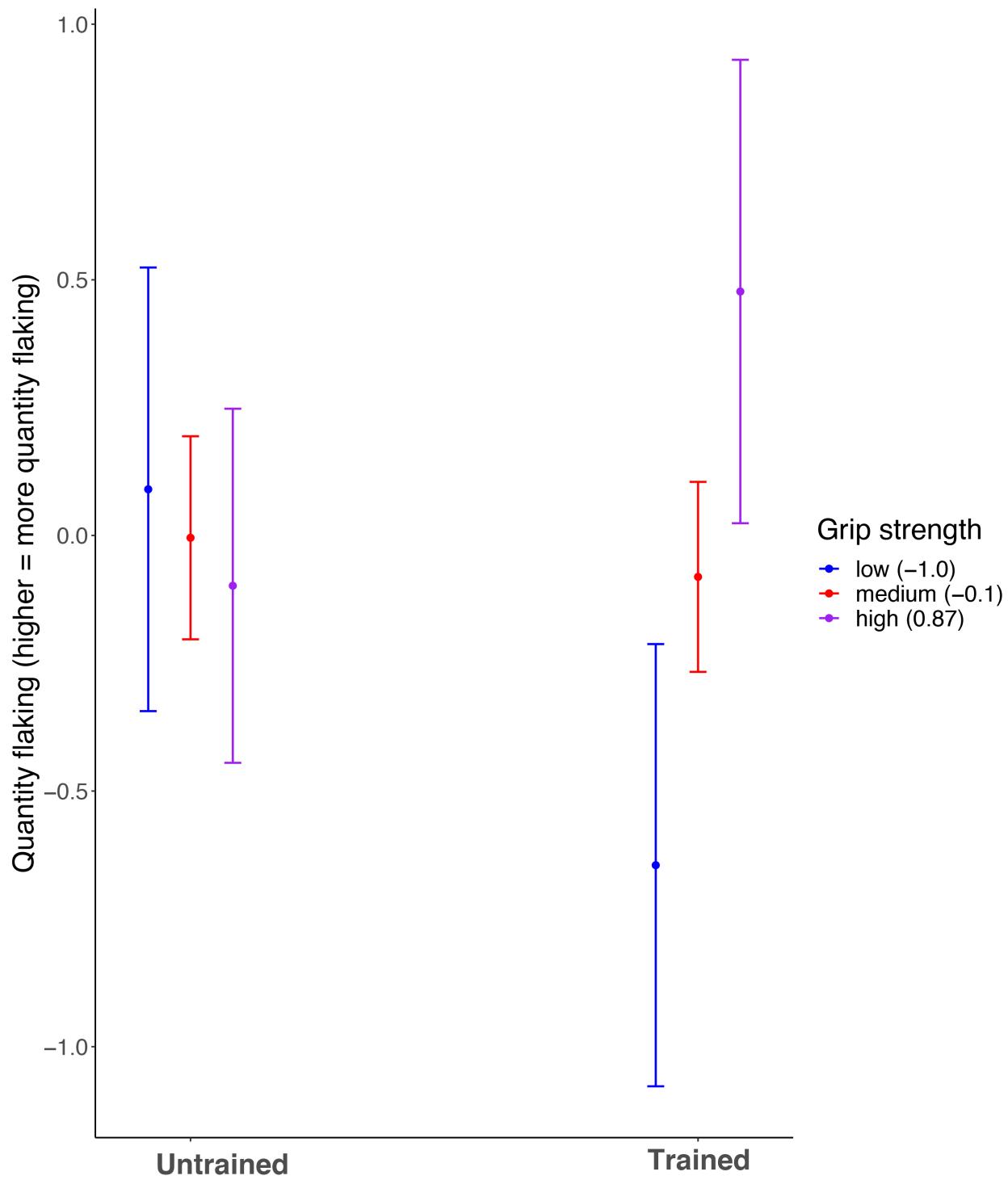


Figure 15: Three significant interaction effects of training condition, visuo-spatial working memory and social information, and training and grip strength use on quantity flaking.

703 **3.5.3 Model 3: Individual differences and quality flaking**

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

708 *Quality flaking* ~ Highest n-back level + Fitt's score + Grip strength + Fitt's score  $\times$  BEAST score  
 709 + Grip strength  $\times$  BEAST score + Grip strength  $\times$  Fitt's score + Training condition  $\times$  Grip strength +  
 710 Nodule mass (as control)

711 This model explains a statistically significant and substantial proportion of variance in Quality  
 712 ( $R^2 = 0.75$ ,  $F(8, 12) = 4.6$ ,  $p < 0.01$ , adj.  $R^2 = 0.6$ ) in the absence of any main training effects. A  
 713 model residuals normality test shows no significant differences with the normal distribution ( $p$   
 714 = 0.41) indicating that this relationship is linear. A Breusch-Pagan test showed no evidence for  
 715 heteroskedasticity ( $BP = 7$ ,  $df = 8$ ,  $p = 0.5$ ).

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

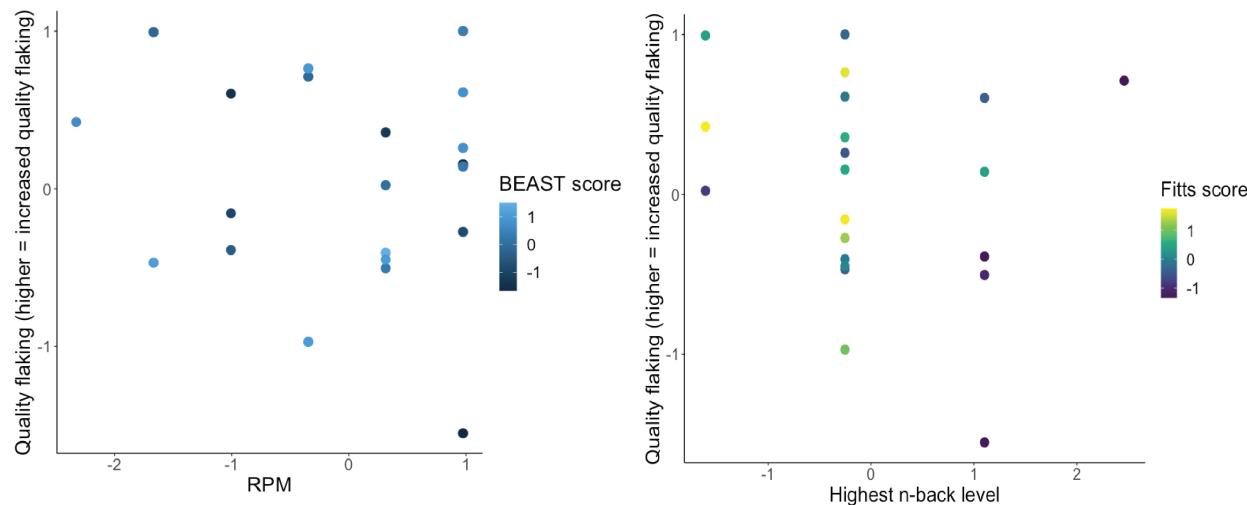


Figure 16: Significant interactions with low estimated effect sizes from the quality flaking model.

721 Without a larger sample, it is not possible to determine if these are anomalous outliers or simply

<sup>722</sup> represent a poorly sampled part of the broader population.

### <sup>723</sup> 3.6 Behavioral observations

<sup>724</sup> We designed this exploratory study primarily to trial experimental design elements such as train-  
<sup>725</sup> ing time, conditions, and raw materials and to collect preliminary data on the effect of individual  
<sup>726</sup> differences and training on knapping outcomes. We thus focused on collecting quantitative  
<sup>727</sup> psychometric and lithic data. However, we also considered that quantifying participant knapping  
<sup>728</sup> behaviors as well as products could be important for future studies. To support methods develop-  
<sup>729</sup> ment in this regard, we made ad hoc notes on observed behaviors during the experiments and  
<sup>730</sup> video-recorded all experiments to enable later, more systematic analyses yet to be completed.  
<sup>731</sup> However, even casual behavior observation was sufficient to reveal an unexpected effect. Whereas  
<sup>732</sup> all trained participants copied the general posture and technique of the expert (free hand knap-  
<sup>733</sup> ping seated in a chair) fully half (6) of the uninstructed participants experimented with or even  
<sup>734</sup> knapped all of their cores using the floor as a support (**Figure 17**). Three of these participants  
<sup>735</sup> were in the same session, which is also the only group composed of just three individuals. In this  
<sup>736</sup> group, knapping on the ground appears to have been transmitted from one participant to the  
<sup>737</sup> other two based on appearance order and the point of gaze of participants.

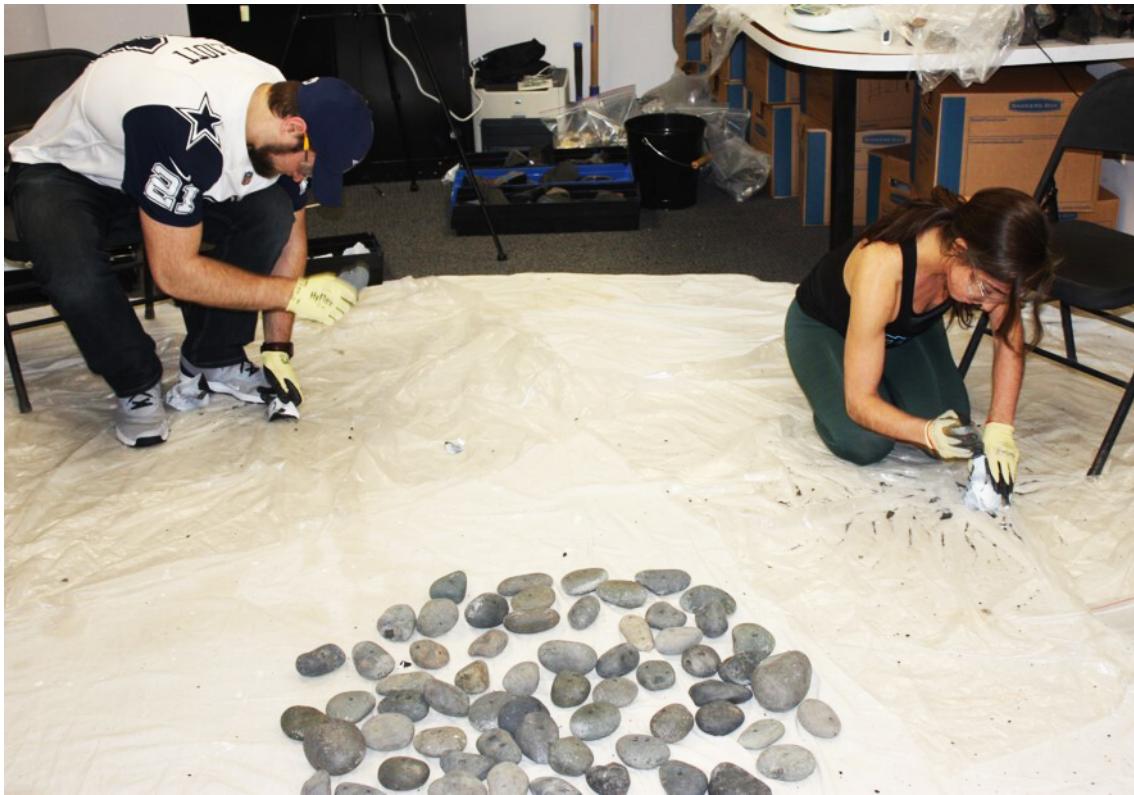


Figure 17: Novices in the untaught condition developing and transmitting a form of bipolar technology involving core reduction on the floor.

## 738 4 Discussion

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

## 748 4.1 Variance Reduction

749 In our experiment, the strongest effects of teaching were to reduce variance (Figure, Table)  
750 rather than shift mean values. In particular, teaching acted as a “safety net” that homogenized  
751 performance by reducing the frequency of extremely poor outcomes (i.e., learning failures). This  
752 finding provides additional support for the hypothesis that teaching would have increased the  
753 reliability of Oldowan skill reproduction ([Morgan et al., 2015](#)) while simultaneously corroborating  
754 the view that basic flaking competence can be achieved in its absence ([Tennie et al., 2017](#)). Our  
755 results thus do not imply that teaching was required or even present during Oldowan times (but  
756 see [Gärdenfors & Höglberg \(2017\)](#)), but rather serve to reinforce the plausibility of co-evolutionary  
757 scenarios positing the cost/reliability of technological skill acquisition as a selection pressure  
758 favoring the evolution of teaching and language ([Morgan et al., 2015](#); [Stout, 2010](#); [Stout & Hecht,  
759 2017](#)).

## 760 4.2 Knapping Behaviors

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

768 We found that a key impact of teaching was to alter basic flake production strategies, as reflected  
769 in the relationship between Total Productivity and detached piece Quality (Fig.Xb). Whereas  
770 Untrained participants achieved greater Productivity at the expense of Quality and vice versa,  
771 these dimensions were unrelated across Trained participants. Thus, even though Untrained  
772 participants achieved the highest values on each metric, only trained individuals managed to  
773 maximize both simultaneously. Indeed, a core function of teaching is to reduce the search space  
774 that learners must explore and increase the likelihood of discovering globally as opposed to  
775 locally optimum solutions (cf. [Hinton & Nowlan, 1996](#); [Stout, 2013](#)). In our study, Untrained  
776 individuals explored a greater range of basic behavioral variations not seen in the Trained group,  
777 including knapping on the floor, concentrating on working just a few cores (2-4, Figure) over the

778 practice period, and showing less constrained nodule size preferences (Figure). It is notable that  
779 this variation occurred even in the presence of an observable expert example, suggesting it may  
780 be interesting for future experiments to address the impact of social context, expectations, and  
781 relationships on observational learning strategies ([Kendal et al., 2018](#)).

782 We also found a strong positive effect of Grip Strength on Total Productivity independent of  
783 learning condition (Figure). While it is tempting to interpret this with respect to the demands for  
784 hand strength specifically, it is important to remember that grip strength is strongly correlated  
785 with total muscle strength ([Wind et al., 2010](#)) and overall fitness ([Sasaki et al., 2007](#)). Thus, it is  
786 best taken to indicate some importance of fitness generally in increasing the rate and intensity  
787 of core reduction by naïve knappers, potentially affecting rate of work and the kinetic energy  
788 of the swing as well as the handling of core and hammerstone. It thus provides further support  
789 for hypotheses positing stone tool making as a selection pressure on the functional anatomy  
790 of hand, arm, and shoulder (e.g., [Williams-Hatala et al., 2018](#)), but initially appears orthogonal  
791 to variations in learning condition and knapping behaviors in our study. However, we also  
792 found that the effect of Grip Strength on per-core knapping Quantity is dependent on teaching  
793 (Figure). The absence of this effect in the uninstructed group reflects the weaker association  
794 between Total Productivity and per-core Quality across these participants (Fig.Xa) and shows  
795 Grip Strength increased uninstructed Productivity specifically by allowing them to knap more  
796 cores rather than to reduce individual cores more heavily. In keeping with this, uninstructed  
797 Grip Strength is positively correlated with Total Cores knapped ( $R^2 = 0.54$ ,  $p = 0.01$ ). Conversely,  
798 strength allowed instructed participants to increase their average Quantity per core without  
799 affecting the total number of cores knapped ( $R^2 = 0.18$ ,  $p = 0.165$ ). Thus, strength appears to  
800 have achieved its effects on core reduction rate and intensity in different ways, depending on  
801 teaching. This difference is likely related to the homogenizing effect of teaching on knapping  
802 rate (all instructed participants knapped 6 or more cores) and methods. Subjectively, knapping  
803 behaviors of uninstructed participants often appeared more physically demanding (e.g., greater  
804 number of non-productive blows, rapid and unregulated battering) which would imply different  
805 demands on both strength and aerobic fitness ([Mateos et al., 2019](#); [Williams-Hatala et al., 2021](#)).  
806 However, this remains to be systematically investigated.

807 In this respect, it is also important to note that we do not know how well the knapping objectives  
808 and strategies communicated by the expert in our experiment correspond to actual Oldowan

809 goals and behaviors. The instructor has successfully replicated assemblage-level patterning at  
810 Gona ([Stout et al., 2019](#)) but Oldowan behavior is variable across space and time (e.g., [Braun et al.,](#)  
811 [2019](#)) and alternative knapping methods might maximize different values (productivity, quality,  
812 effort), especially in novices ([Putt, 2015](#)) ([Putt 2015](#)). Nevertheless, the effect of instruction to  
813 constrain behavioral exploration and homogenize outcomes is clear. We expect that this effect  
814 would generalize to the teaching of alternative knapping goals and behaviors, although this  
815 remains to be tested.

816 **4.3 Learning Strategies**

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

833 At a basic level, learners of any new task must balance investment in task exploration vs. ex-  
834 ploitation of knowledge and skills already in hand ([Sutton & Barto, 2018](#)). Premature exploitation  
835 risks settling for a sub-optimal local solution whereas continued exploration sacrifices more  
836 immediate payoffs. Managing this trade-off is especially challenging for complex, real-world  
837 tasks like stone knapping, and is thought to depend on the interplay of uncertainty and reward  
838 expectation ([Wilson et al., 2021](#)). Teaching and social learning generally have the potential to

839 provide low-cost information about task structure and payoffs (Kendal et al., 2018; Rendell et al.,  
840 2010), which if adopted, would be expected to affect exploration/exploitation decisions. Such  
841 adoption is itself known to be influenced by individual cognitive differences, for example if higher  
842 fluid intelligence allows observers to better understand observed tasks (Vostroknutov et al., 2018)  
843 or if individuals vary in their tendency to use and value social information (Molleman et al., 2019;  
844 Toelch et al., 2014).

845 In our study, we did not observe any effect of fluid intelligence (RPM) on knapping outcomes but  
846 did find strong interactions of learning condition with participant visuospatial working memory  
847 and social information use tendency (Figure). As expected, uninstructed individuals with higher  
848 scores on these dimensions displayed higher Total Productivity and average per-core flaking  
849 Quantity (although the effect on n-Back on Quantity did not achieve significance). We attribute  
850 these effects to increased ability to hold relevant morphological/spatial information in mind and a  
851 tendency to benefit from observing successful strategies of others, including the expert model. In  
852 contrast, instructed individuals with higher scores tended to have lower Productivity and Quantity.  
853 We interpret this unexpected effect to an increased tendency to privilege exploratory learning  
854 behavior over exploitation. In particular, we suggest that trained participants might knap more  
855 slowly and less productively if higher working memory capacity inclined them to experiment  
856 more with morphological/spatial variables highlighted by the instructor or if a predisposition to  
857 use social information use caused them to invest greater time and effort attending to and trying  
858 out observed actions and/or instructions. These suggestions remain to be tested by further work.  
859 Unfortunately, the training period in our current experiment was insufficient to capture learning  
860 effects and so we have no evidence of the effects of these individual differences and putative  
861 exploration/exploitation tradeoffs to the ultimate achievement of expertise. A similar negative  
862 effect of instruction on knapping outcomes during early stage learning was reported by Putt et  
863 al. (2014), and has been interpreted to reflect learners experimenting with advanced techniques  
864 before they have the perceptual-motor skill to execute them (Stout & Khriesheh, 2015; Whiten,  
865 2015). Such effects might be further explored with more detailed behavioral data, as opposed to  
866 purely lithic data, and with longer learning periods.

867 **4.4 Limitations and Prospects**

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

885 Another major limitation that our study shares with all other published experiments on knapping  
886 skill acquisition is that we do not address variation in social and cultural context or in teaching  
887 style. Currently, we have little basis other than personal experience/tradition (Callahan, 1979;  
888 Shea, 2015; Whittaker, 1994) and theoretical speculation (Stout, 2013; Whiten, 2015) from which to  
889 assess which pedagogical techniques are most effective even in WEIRD contexts. No study to date  
890 has considered how variation in teacher skill (Shea, 2015) or social relationship to participants  
891 might impact learning under different conditions. To properly address these questions would  
892 require a major research program, including both cross-cultural comparative studies (Barrett,  
893 2020) and more naturalistic study designs. While costly, such research would produce results  
894 of broad relevance to anthropologists, biologists, psychologists, and sociologists interested in  
895 teaching and learning.

896 **5 Conclusions**

897 **6 Acknowledgments**

898 **7 Figures**

899 **References**

- 900 Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation  
901 supports flexible tool use and physical reasoning. *Proceedings of the National Academy of  
902 Sciences*, 117(47), 29302–29310. <https://doi.org/10.1073/pnas.1912341117>
- 903 Barrett, H. C. (2020). Towards a Cognitive Science of the Human: Cross-Cultural Approaches and  
904 Their Urgency. *Trends in Cognitive Sciences*, 24(8), 620–638. [https://doi.org/10.1016/j.tics.202  
0.05.007](https://doi.org/10.1016/j.tics.202<br/>905 0.05.007)
- 906 Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Develop-  
907 opment of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test.  
908 *Assessment*, 19(3), 354–369. <https://doi.org/10.1177/1073191112446655>
- 909 Boogert, N. J., Madden, J. R., Morand-Ferron, J., & Thornton, A. (2018). Measuring and under-  
910 standing individual differences in cognition. *Philosophical Transactions of the Royal Society B:  
911 Biological Sciences*, 373(1756), 20170280. <https://doi.org/10.1098/rstb.2017.0280>
- 912 Boyette, A. H., & Hewlett, B. S. (2017). Autonomy, Equality, and Teaching among Aka Foragers and  
913 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/<br/>914 s12110-017-9294-y)
- 915 Braun, D. R., Aldeias, V., Archer, W., Arrowsmith, J. R., Baraki, N., Campisano, C. J., Deino, A.  
916 L., DiMaggio, E. N., Dupont-Nivet, G., Engda, B., Feary, D. A., Garello, D. I., Kerfelew, Z.,  
917 McPherron, S. P., Patterson, D. B., Reeves, J. S., Thompson, J. C., & Reed, K. E. (2019). Earliest  
918 known Oldowan artifacts at >2.58 Ma from Ledi-Geraru, Ethiopia, highlight early technological  
919 diversity. *Proceedings of the National Academy of Sciences*, 116(24), 11712–11717. <https://doi.org/10.1073/pnas.1820177116>
- 921 Braun, D. R., Plummer, T., Ferraro, J. V., Ditchfield, P., & Bishop, L. C. (2009). Raw material quality

- 922 and Oldowan hominin toolstone preferences: Evidence from Kanjera South, Kenya. *Journal of*  
923 *Archaeological Science*, 36(7), 1605–1614. <https://doi.org/10.1016/j.jas.2009.03.025>
- 924 Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical*  
925 *Information-Theoretic Approach* (2nd ed.). Springer-Verlag. <https://doi.org/10.1007/b97636>
- 926 Caldwell, C. A., Atkinson, M., Blakey, K. H., Dunstone, J., Kean, D., Mackintosh, G., Renner, E., &  
927 Wilks, C. E. H. (2020). Experimental assessment of capacities for cumulative culture: Review  
928 and evaluation of methods. *WIREs Cognitive Science*, 11(1), e1516. <https://doi.org/10.1002/wcs.1516>
- 930 Callahan, E. (1979). The basics of biface knapping in the eastern fluted point tradition: A manual  
931 for flintknappers and lithic analysts. *Archaeology of Eastern North America*, 7(1), 1–180.  
932 <https://www.jstor.org/stable/40914177>
- 933 Cataldo, D. M., Migliano, A. B., & Vinicius, L. (2018). Speech, stone tool-making and the evolution  
934 of language. *PLOS ONE*, 13(1), e0191071. <https://doi.org/10.1371/journal.pone.0191071>
- 935 Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of*  
936 *Educational Psychology*, 54(1), 1–22. <https://doi.org/10.1037/h0046743>
- 937 Coolidge, F. L., & Wynn, T. (2005). Working Memory, its Executive Functions, and the Emergence  
938 of Modern Thinking. *Cambridge Archaeological Journal*, 15(1), 5–26. <https://doi.org/10.1017/S0959774305000016>
- 940 Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or, The Preservation of*  
941 *Favoured Races in the Struggle for Life* (1st ed.). John Murray.
- 942 Darwin, C. (1871). *The descent of man, and selection in relation to sex* (1st ed.). John Murray.
- 943 Duke, H., & Pargeter, J. (2015). Weaving simple solutions to complex problems: An experimental  
944 study of skill in bipolar cobble-splitting. *Lithic Technology*, 40(4), 349–365. <https://doi.org/10.1179/2051618515Y.0000000016>
- 946 Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psycho-*  
947 *logical Science*, 13(2), 190–193. <https://doi.org/10.1177/1745691617720478>
- 948 Engles, F. (2003). The part played by labour in the transition from ape to man. In R. C. Scharff &  
949 V. Dusek (Eds.), *Philosophy of Technology – The Technological Condition: An Anthology* (pp.

- 950        71–77). Blackwell.
- 951    Eren, M. I., Bradley, B. A., & Sampson, C. G. (2011). Middle Paleolithic Skill Level and the Individual  
952        Knapper: An Experiment. *American Antiquity*, 76(2), 229–251. <https://doi.org/10.7183/0002-7316.76.2.229>
- 953
- 954    Eren, M. I., Lycett, S. J., Patten, R. J., Buchanan, B., Pargeter, J., & O'Brien, M. J. (2016). Test, model,  
955        and method validation: The role of experimental stone artifact replication in hypothesis-  
956        driven archaeology. *Ethnoarchaeology: Journal of Archaeological, Ethnographic and Experi-  
957        mental Studies*, 8(2), 103–136. <https://doi.org/10.1080/19442890.2016.1213972>
- 958    Eren, M. I., Roos, C. I., Story, B. A., von Cramon-Taubadel, N., & Lycett, S. J. (2014). The role of raw  
959        material differences in stone tool shape variation: an experimental assessment. *Journal of  
960        Archaeological Science*, 49, 472–487. <https://doi.org/10.1016/j.jas.2014.05.034>
- 961    Faisal, A., Stout, D., Apel, J., & Bradley, B. (2010). The Manipulative Complexity of Lower Paleolithic  
962        Stone Toolmaking. *PLOS ONE*, 5(11), e13718. <https://doi.org/10.1371/journal.pone.0013718>
- 963    Fitts, P. M. (1954). The information capacity of the human motor system in controlling the  
964        amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–391. <https://doi.org/10.1037/h0055392>
- 965
- 966    García-Medrano, P., Ollé, A., Ashton, N., & Roberts, M. B. (2019). The Mental Template in Handaxe  
967        Manufacture: New Insights into Acheulean Lithic Technological Behavior at Boxgrove, Sussex,  
968        UK. *Journal of Archaeological Method and Theory*, 26(1), 396–422. <https://doi.org/10.1007/s10816-018-9376-0>
- 969
- 970    Gärdenfors, P., & Höglberg, A. (2017). The archaeology of teaching and the evolution of homo  
971        docens. *Current Anthropology*, 58(2), 188–208. <https://doi.org/10.1086/691178>
- 972    Geribàs, N., Mosquera, M., & Vergès, J. M. (2010). What novice knappers have to learn to become  
973        expert stone toolmakers. *Journal of Archaeological Science*, 37(11), 2857–2870. <https://doi.org/10.1016/j.jas.2010.06.026>
- 974
- 975    Gowlett, J. A. J. (1984). Mental abilities of early man: A look at some hard evidence. *Higher  
976        Education Quarterly*, 38(3), 199–220. <https://doi.org/10.1111/j.1468-2273.1984.tb01387.x>
- 977    Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting  
978        to new responses in a weigl-type card-sorting problem. *Journal of Experimental Psychology*,

- 979        38(4), 404–411. <https://doi.org/10.1037/h0059831>
- 980        Hecht, E. E., Gutman, D. A., Bradley, B. A., Preuss, T. M., & Stout, D. (2015). Virtual dissection and  
981        comparative connectivity of the superior longitudinal fasciculus in chimpanzees and humans.  
982        *NeuroImage*, 108, 124–137. <https://doi.org/10.1016/j.neuroimage.2014.12.039>
- 983        Hecht, E. E., Gutman, D. A., Khriesheh, N., Taylor, S. V., Kilner, J. M., Faisal, A. A., Bradley, B. A.,  
984        Chaminade, T., & Stout, D. (2015). Acquisition of Paleolithic toolmaking abilities involves  
985        structural remodeling to inferior frontoparietal regions. *Brain Structure & Function*, 220(4),  
986        2315–2331. <https://doi.org/10.1007/s00429-014-0789-6>
- 987        Hecht, E. E., Gutman, D. A., Preuss, T. M., Sanchez, M. M., Parr, L. A., & Rilling, J. K. (2013). Process  
988        versus product in social learning: Comparative diffusion tensor imaging of neural systems  
989        for action executionobservation matching in macaques, chimpanzees, and humans. *Cerebral  
990        Cortex*, 23(5), 1014–1024. <https://doi.org/10.1093/cercor/bhs097>
- 991        Hecht, E. E., Murphy, L. E., Gutman, D. A., Votaw, J. R., Schuster, D. M., Preuss, T. M., Orban,  
992        G. A., Stout, D., & Parr, L. A. (2013). Differences in neural activation for object-directed  
993        grasping in chimpanzees and humans. *The Journal of Neuroscience*, 33(35), 14117–14134.  
994        <https://doi.org/10.1523/JNEUROSCI.2172-13.2013>
- 995        Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302),  
996        29–29. <https://doi.org/10.1038/466029a>
- 997        Hewes, G. W. (1993). A history of speculation on the relation between tools and language. In K.  
998        R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution* (pp. 20–31).  
999        Cambridge University Press.
- 1000        Heyes, C. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transac-  
1001        tions of the Royal Society B: Biological Sciences*, 373(1743), 20170051. <https://doi.org/10.1098/rstb.2017.0051>
- 1003        Hinton, G. E., & Nowlan, S. J. (1996). How learning can guide evolution. In R. K. Belew & M.  
1004        Mitchell (Eds.), *Adaptive individuals in evolving populations: Models and algorithms* (pp.  
1005        447–454). Addison-Wesley Publishing Company.
- 1006        Isaac, G. L. (1976). Stages of Cultural Elaboration in the Pleistocene: Possible Archaeological  
1007        Indicators of the Development of Language Capabilities. *Annals of the New York Academy of*

- 1008      *Sciences*, 280(1), 275–288. <https://doi.org/10.1111/j.1749-6632.1976.tb25494.x>
- 1009      Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and*  
1010      *instruction*. Lawrence Erlbaum.
- 1011      Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social  
1012      Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, 22(7),  
1013      651–665. <https://doi.org/10.1016/j.tics.2018.04.003>
- 1014      Key, A. J. M., & Dunmore, C. J. (2015). The evolution of the hominin thumb and the influence  
1015      exerted by the non-dominant hand during stone tool production. *Journal of Human Evolution*,  
1016      78, 60–69. <https://doi.org/10.1016/j.jhevol.2014.08.006>
- 1017      Key, A. J. M., & Dunmore, C. J. (2018). Manual restrictions on Palaeolithic technological behaviours.  
1018      *PeerJ*, 6, e5399. <https://doi.org/10.7717/peerj.5399>
- 1019      Key, A. J. M., & Lycett, S. J. (2014). Are bigger flakes always better? An experimental assessment of  
1020      flake size variation on cutting efficiency and loading. *Journal of Archaeological Science*, 41,  
1021      140–146. <https://doi.org/10.1016/j.jas.2013.07.033>
- 1022      Key, A. J. M., & Lycett, S. J. (2019). Biometric variables predict stone tool functional performance  
1023      more effectively than tool-form attributes: a case study in handaxe loading capabilities.  
1024      *Archaeometry*, 61(3), 539–555. <https://doi.org/10.1111/arcm.12439>
- 1025      Khreisheh, N. N., Davies, D., & Bradley, B. A. (2013). Extending Experimental Control: The Use of  
1026      Porcelain in Flaked Stone Experimentation. *Advances in Archaeological Practice*, 1(1), 38–46.  
1027      <https://doi.org/10.7183/2326-3768.1.1.37>
- 1028      Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of  
1029      teaching behavior in humans and other animals. *The Behavioral and Brain Sciences*, 38, e31.  
1030      <https://doi.org/10.1017/S0140525X14000090>
- 1031      Laland, K. N. (2017). The origins of language in teaching. *Psychonomic Bulletin & Review*, 24(1),  
1032      225–231. <https://doi.org/10.3758/s13423-016-1077-7>
- 1033      Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philoso-*  
1034      *sophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130302. <https://doi.org/10.1098/rstb.2013.0302>

- 1036 Lombao, D., Guardiola, M., & Mosquera, M. (2017). Teaching to make stone tools: new experi-  
1037 mental evidence supporting a technological hypothesis for the origins of language. *Scientific*  
1038 *Reports*, 7(1), 1–14. <https://doi.org/10.1038/s41598-017-14322-y>
- 1039 Marwick, B. (2017). Computational Reproducibility in Archaeological Research: Basic Principles  
1040 and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*,  
1041 24(2), 424–450. <https://doi.org/10.1007/s10816-015-9272-9>
- 1042 Marzke, M. W., Toth, N., Schick, K., Reece, S., Steinberg, B., Hunt, K., Linscheid, R. L., & An, K.-N.  
1043 (1998). EMG study of hand muscle recruitment during hard hammer percussion manufacture  
1044 of Oldowan tools. *American Journal of Physical Anthropology*, 105(3), 315–332. <https://doi.or>  
1045 [g/10.1002/\(SICI\)1096-8644\(199803\)105:3%3C315::AID-AJPA3%3E3.0.CO;2-Q](g/10.1002/(SICI)1096-8644(199803)105:3%3C315::AID-AJPA3%3E3.0.CO;2-Q)
- 1046 Mateos, A., Terradillos-Bernal, M., & Rodríguez, J. (2019). Energy Cost of Stone Knapping. *Journal*  
1047 *of Archaeological Method and Theory*, 26(2), 561–580. <https://doi.org/10.1007/s10816-018-9382-2>
- 1049 Miu, E., Gulley, N., Laland, K. N., & Rendell, L. (2020). Flexible learning, rather than inveterate  
1050 innovation or copying, drives cumulative knowledge gain. *Science Advances*, 6(23), eaaz0286.  
1051 <https://doi.org/10.1126/sciadv.aaz0286>
- 1052 Molleman, L., Kurvers, R. H. J. M., & van den Bos, W. (2019). Unleashing the BEAST: a brief  
1053 measure of human social information use. *Evolution and Human Behavior*, 40(5), 492–499.  
1054 <https://doi.org/10.1016/j.evolhumbehav.2019.06.005>
- 1055 Montagu, A. (1976). Toolmaking, Hunting, and the Origin of Language. *Annals of the New York*  
1056 *Academy of Sciences*, 280(1), 266–274. <https://doi.org/10.1111/j.1749-6632.1976.tb25493.x>
- 1057 Morgan, T. J. H., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M.,  
1058 Cross, C. P., Evans, C., Kearney, R., de la Torre, I., Whiten, A., & Laland, K. N. (2015). Experi-  
1059 mental evidence for the co-evolution of hominin tool-making teaching and language. *Nature*  
1060 *Communications*, 6(1), 6029. <https://doi.org/10.1038/ncomms7029>
- 1061 Nonaka, T., Bril, B., & Rein, R. (2010). How do stone knappers predict and control the outcome  
1062 of flaking? Implications for understanding early stone tool technology. *Journal of Human*  
1063 *Evolution*, 59(2), 155–167. <https://doi.org/10.1016/j.jhevol.2010.04.006>
- 1064 Oakley, K. P. (1949). *Man the toolmaker*. Trustees of the British Museum.

- 1065 Ohnuma, K., Aoki, K., & Akazawa, A. T. (1997). Transmission of tool-making through verbal  
1066 and non-verbal communication: Preliminary experiments in levallois flake production.  
1067 *Anthropological Science*, 105(3), 159–168. <https://doi.org/10.1537/ase.105.159>
- 1068 Pargeter, J., Khreisheh, N., Shea, J. J., & Stout, D. (2020). Knowledge vs. know-how? Dissecting  
1069 the foundations of stone knapping skill. *Journal of Human Evolution*, 145, 102807. <https://doi.org/10.1016/j.jhevol.2020.102807>
- 1071 Pargeter, J., Khreisheh, N., & Stout, D. (2019). Understanding stone tool-making skill acquisition:  
1072 Experimental methods and evolutionary implications. *Journal of Human Evolution*, 133,  
1073 146–166. <https://doi.org/10.1016/j.jhevol.2019.05.010>
- 1074 Pelegrin, J. (1990). Prehistoric Lithic Technology : Some Aspects of Research. *Archaeological  
1075 Review from Cambridge*, 9(1), 116–125. [/paper/Prehistoric-Lithic-Technology-%3A-Some-  
1076 Aspects-of-Pelegrin/5e02fc2a5280ac128727275ab6b833756e6a6056](#)
- 1077 Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to  
1078 large-scale decoding. *Neuron*, 72(5), 692–697. <https://doi.org/10.1016/j.neuron.2011.11.001>
- 1079 Prascunas, M. M. (2007). Bifacial Cores and Flake Production Efficiency: An Experimental Test of  
1080 Technological Assumptions. *American Antiquity*, 72(2), 334–348. [https://doi.org/10.2307/40035817](https://doi.org/10.2307/40<br/>1081 035817)
- 1082 Putt, S. S. (2015). The origins of stone tool reduction and the transition to knapping: An experi-  
1083 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<br/>1084 6/j.jasrep.2015.01.004)
- 1085 Putt, S. S., Wijekumar, S., Franciscus, R. G., & Spencer, J. P. (2017). The functional brain networks  
1086 that underlie Early Stone Age tool manufacture. *Nature Human Behaviour*, 1(6), 1–8. <https://doi.org/10.1038/s41562-017-0102>
- 1088 Putt, S. S., Wijekumar, S., & Spencer, J. P. (2019). Prefrontal cortex activation supports the  
1089 emergence of early stone age toolmaking skill. *NeuroImage*, 199, 57–69. [https://doi.org/10.1016/j.neuroimage.2019.05.056](https://doi.org/10.101<br/>1090 016/j.neuroimage.2019.05.056)
- 1091 Putt, S. S., Woods, A. D., & Franciscus, R. G. (2014). The role of verbal interaction during  
1092 experimental bifacial stone tool manufacture. *Lithic Technology*, 39(2), 96–112. <https://doi.org/10.1179/0197726114Z.00000000036>

- 1094 Rein, R., Nonaka, T., & Bril, B. (2014). Movement Pattern Variability in Stone Knapping: Im-  
1095 plications for the Development of Percussive Traditions. *PLOS ONE*, 9(11), e113567. <https://doi.org/10.1371/journal.pone.0113567>
- 1096  
1097 Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., Fogarty, L.,  
1098 Ghirlanda, S., Lillicrap, T., & Laland, K. N. (2010). Why Copy Others? Insights from the  
1099 Social Learning Strategies Tournament. *Science*, 328(5975), 208–213. <https://doi.org/10.1126/science.1184719>
- 1100  
1101 Reti, J. S. (2016). Quantifying Oldowan Stone Tool Production at Olduvai Gorge, Tanzania. *PLOS  
1102 ONE*, 11(1), e0147352. <https://doi.org/10.1371/journal.pone.0147352>
- 1103 Roux, V., Bril, B., & Dietrich, G. (1995). Skills and learning difficulties involved in stone knapping:  
1104 The case of stone-bead knapping in khambhat, india. *World Archaeology*, 27(1), 63–87. <https://doi.org/10.1080/00438243.1995.9980293>
- 1105  
1106 Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W.  
1107 (2017). ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*,  
1108 18(1), 529. <https://doi.org/10.1186/s12859-017-1934-z>
- 1109 Sasaki, H., Kasagi, F., Yamada, M., & Fujita, S. (2007). Grip Strength Predicts Cause-Specific  
1110 Mortality in Middle-Aged and Elderly Persons. *The American Journal of Medicine*, 120(4),  
1111 337–342. <https://doi.org/10.1016/j.amjmed.2006.04.018>
- 1112 Schillinger, K., Mesoudi, A., & Lycett, S. J. (2014). Copying Error and the Cultural Evolution  
1113 of “Additive” vs. “Reductive” Material Traditions: An Experimental Assessment. *American  
1114 Antiquity*, 79(1), 128–143. <https://doi.org/10.7183/0002-7316.79.1.128>
- 1115 Shallice, T., Broadbent, D. E., & Weiskrantz, L. (1982). Specific impairments of planning. *Philoso-  
1116 sophical Transactions of the Royal Society of London. B, Biological Sciences*, 298(1089), 199–209.  
1117 <https://doi.org/10.1098/rstb.1982.0082>
- 1118 Shea, J. J. (2015). Making and using stone tools: Advice for learners and teachers and insights for  
1119 archaeologists. *Lithic Technology*, 40(3), 231–248. [https://doi.org/10.1179/2051618515Y.0000000011](https://doi.org/10.1179/2051618515Y.0000<br/>1120 000011)
- 1121 Shea, J. J. (2016). *Stone tools in human evolution: Behavioral differences among technological  
1122 primates*. Cambridge University Press. <https://doi.org/10.1017/9781316389355>

- 1123 Sherwood, C. C., & Gómez-Robles, A. (2017). Brain plasticity and human evolution. *Annual Review  
1124 of Anthropology*, 46(1), 399–419. <https://doi.org/10.1146/annurev-anthro-102215-100009>
- 1125 Stout, D. (2002). Skill and cognition in stone tool production: An ethnographic case study from  
1126 irian jaya. *Current Anthropology*, 43(5), 693–722. <https://doi.org/10.1086/342638>
- 1127 Stout, D. (2010). Possible relations between language and technology in human evolution. In A.  
1128 Nowell & I. Davidson (Eds.), *Stone tools and the evolution of human cognition* (pp. 159–184).  
1129 University Press of Colorado.
- 1130 Stout, D. (2013). Neuroscience of technology. In P. J. Richerson & M. H. Christiansen (Eds.),  
1131 *Cultural evolution: Society, technology, language, and religion* (pp. 157–173). The MIT Press.
- 1132 Stout, D., Apel, J., Commander, J., & Roberts, M. (2014). Late Acheulean technology and cognition  
1133 at Boxgrove, UK. *Journal of Archaeological Science*, 41, 576–590. <https://doi.org/10.1016/j.jas.2013.10.001>
- 1135 Stout, D., & Chaminade, T. (2007). The evolutionary neuroscience of tool making. *Neuropsycholo-  
1136 gia*, 45(5), 1091–1100. <https://doi.org/10.1016/j.neuropsychologia.2006.09.014>
- 1137 Stout, D., & Chaminade, T. (2012). Stone tools, language and the brain in human evolution.  
1138 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585), 75–87. <https://doi.org/10.1098/rstb.2011.0099>
- 1140 Stout, D., & Hecht, E. E. (2017). Evolutionary neuroscience of cumulative culture. *Proceedings of  
1141 the National Academy of Sciences*, 114(30), 7861–7868. <https://doi.org/10.1073/pnas.1620738114>
- 1143 Stout, D., Hecht, E., Khreisheh, N., Bradley, B., & Chaminade, T. (2015). Cognitive Demands of  
1144 Lower Paleolithic Toolmaking. *PLOS ONE*, 10(4), e0121804. <https://doi.org/10.1371/journal.pone.0121804>
- 1146 Stout, D., & Khreisheh, N. (2015). Skill Learning and Human Brain Evolution: An Experimental  
1147 Approach. *Cambridge Archaeological Journal*, 25(4), 867–875. <https://doi.org/10.1017/S0959774315000359>
- 1149 Stout, D., Passingham, R., Frith, C., Apel, J., & Chaminade, T. (2011). Technology, expertise and  
1150 social cognition in human evolution. *The European Journal of Neuroscience*, 33(7), 1328–1338.  
1151 <https://doi.org/10.1111/j.1460-9568.2011.07619.x>

- 1152 Stout, D., Quade, J., Semaw, S., Rogers, M. J., & Levin, N. E. (2005). Raw material selectivity of the  
1153 earliest stone toolmakers at Gona, Afar, Ethiopia. *Journal of Human Evolution*, 48(4), 365–380.  
1154 <https://doi.org/10.1016/j.jhevol.2004.10.006>
- 1155 Stout, D., Rogers, M. J., Jaeggi, A. V., & Semaw, S. (2019). Archaeology and the origins of hu-  
1156 man cumulative culture: A case study from the earliest oldowan at gona, ethiopia. *Current*  
1157 *Anthropology*, 60(3), 309–340. <https://doi.org/10.1086/703173>
- 1158 Stout, D., & Semaw, S. (2006). Knapping skill of the earliest stone toolmakers: Insights from the  
1159 study of modern human novices. In N. Toth & K. Schick (Eds.), *The Oldowan: Case studies into*  
1160 *the earliest Stone Age* (pp. 307–320). Stone Age Institute Press.
- 1161 Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT  
1162 Press.
- 1163 Tehrani, J. J., & Riede, F. (2008). Towards an archaeology of pedagogy: Learning, teaching and  
1164 the generation of material culture traditions. *World Archaeology*, 40(3), 316–331. <https://doi.org/10.1080/00438240802261267>
- 1166 Tennie, C., Premo, L. S., Braun, D. R., & McPherron, S. P. (2017). Early stone tools and cultural  
1167 transmission: Resetting the null hypothesis. *Current Anthropology*, 58(5), 652–672. <https://doi.org/10.1086/693846>
- 1169 Toelch, U., Bruce, M. J., Newson, L., Richerson, P. J., & Reader, S. M. (2014). Individual consistency  
1170 and flexibility in human social information use. *Proceedings of the Royal Society B: Biological*  
1171 *Sciences*, 281(1776), 20132864. <https://doi.org/10.1098/rspb.2013.2864>
- 1172 Toth, N., & Schick, K. (1993). Early stone industries and inferences regarding language and  
1173 cognition. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution*  
1174 (pp. 346–362). Cambridge University Press.
- 1175 Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and  
1176 learning: Evidence from the serial reaction time task. *Memory & Cognition*, 33(2), 213–220.  
1177 <https://doi.org/10.3758/BF03195310>
- 1178 Vostroknutov, A., Polonio, L., & Coricelli, G. (2018). The Role of Intelligence in Social Learning.  
1179 *Scientific Reports*, 8(1), 6896. <https://doi.org/10.1038/s41598-018-25289-9>
- 1180 Washburn, S. L. (1960). Tools and human evolution. *Scientific American*, 203(3), 62–75. <https://doi.org/10.1038/s41598-018-25289-9>

- 1181        [//doi.org/10.1038/scientificamerican0960-62](https://doi.org/10.1038/scientificamerican0960-62)
- 1182    Whiten, A. (2015). Experimental studies illuminate the cultural transmission of percussive tech-  
1183        nologies in homo and pan. *Philosophical Transactions of the Royal Society B: Biological*  
1184        *Sciences*, 370(1682), 20140359. <https://doi.org/10.1098/rstb.2014.0359>
- 1185    Whittaker, J. C. (1994). *Flintknapping: Making and Understanding Stone Tools*. University of Texas  
1186        Press.
- 1187    Wilkins, J. (2018). The Point is the Point: Emulative social learning and weapon manufacture  
1188        in the Middle Stone Age of South Africa. In M. J. O'Brien, B. Buchanan, & M. I. Eren (Eds.),  
1189        *Convergent Evolution in Stone-Tool Technology* (pp. 153–174). The MIT Press.
- 1190    Williams-Hatala, E. M., Hatala, K. G., Gordon, M., Key, A., Kasper, M., & Kivell, T. L. (2018). The  
1191        manual pressures of stone tool behaviors and their implications for the evolution of the human  
1192        hand. *Journal of Human Evolution*, 119, 14–26. <https://doi.org/10.1016/j.jhevol.2018.02.008>
- 1193    Williams-Hatala, E. M., Hatala, K. G., Key, A., Dunmore, C. J., Kasper, M., Gordon, M., & Kivell, T.  
1194        L. (2021). Kinetics of stone tool production among novice and expert tool makers. *American*  
1195        *Journal of Physical Anthropology*, 174(4), 714–727. <https://doi.org/10.1002/ajpa.24159>
- 1196    Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploita-  
1197        tion with information and randomization. *Current Opinion in Behavioral Sciences*, 38, 49–56.  
1198        <https://doi.org/10.1016/j.cobeha.2020.10.001>
- 1199    Wind, A. E., Takken, T., Helders, P. J. M., & Engelbert, R. H. H. (2010). Is grip strength a predictor for  
1200        total muscle strength in healthy children, adolescents, and young adults? *European Journal of*  
1201        *Pediatrics*, 169(3), 281–287. <https://doi.org/10.1007/s00431-009-1010-4>
- 1202    Wynn, T. (1979). The intelligence of later acheulean hominids. *Man*, 14(3), 371–391. <https://doi.org/10.2307/2801865>
- 1203
- 1204    Wynn, T. (2017). Evolutionary cognitive archaeology. In T. Wynn & F. Coolidge (Eds.), *Cognitive*  
1205        *models in Palaeolithic archaeology* (pp. 1–20). Oxford University Press.
- 1206    Wynn, T., & Coolidge, F. L. (2004). The expert Neandertal mind. *Journal of Human Evolution*,  
1207        46(4), 467–487. <https://doi.org/10.1016/j.jhevol.2004.01.005>
- 1208    Wynn, T., & Coolidge, F. L. (2016). Archeological insights into hominin cognitive evolution.

<sub>1209</sub>      *Evolutionary Anthropology: Issues, News, and Reviews*, 25(4), 200–213. <https://doi.org/10.1002/evan.21496>

<sub>1211</sub>    Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. <https://doi.org/10.1017/S0140525X20001685>