Experimental Evidence for the Co-Evolution of Hominin Tool-Making Teaching and Language

T.J.H. Morgan[a,b], N. T. Uomini[c]*, L.E. Rendell[a], L. Chouinard-Thuly[a,d], S. E. Street[a,e], H. M. Lewis[a,f], C. P. Cross[a,e], C. Evans[a], R. Kearney[a], I. De la Torre[g], A. Whiten[e], K.N. Laland[a]*

[a] Centre for Social Learning and Cognitive Evolution, School of Biology, University of St Andrews, Fife, KY16 9AJ, U.K.
[b] Department of Psychology, University of California, Berkeley, 94720, United States
[c] Department of Archaeology, Classics & Egyptology, University of Liverpool, L69 3BX, U.K.
[c] Department of Linguistics and Department of Primatology, Max-Planck Institute for Evolutionary Anthropology, Leipzig
[d] Department of Biology, McGill University, H3A 1B1, Canada
[e] Centre for Social Learning and Cognitive Evolution, School of Psychology & Neuroscience, University of St Andrews, Fife, KY16, 9JP, U.K.
[f] Department of Anthropology, University College London, WC1E 6BT, U.K.
[g] Institute of Archaeology, University College London, WC1H 0PY, U.K.

* Correspondence:

K. N. Laland: knl1@st-andrews.ac.uk, 01334 463568
N. T. Uomini: N.Uomini@liverpool.ac.uk, 01517 945787

Keywords: tool-use || human evolution || social transmission || language evolution

Abstract

Hominin reliance on Oldowan stone tools – which appear from 2.5mya and are believed to have been socially transmitted – has been hypothesised to have led to the evolution of teaching and language. Here we present an experiment investigating the efficacy of transmission of Oldowan tool-making skills along chains of adult human participants (N=184) using 5 different transmission mechanisms. Across six measures, transmission improves with teaching, and particularly with language, but not with imitation or emulation. Our results support the hypothesis that hominin reliance on stone tool-making generated selection for teaching and language and imply that (*i*) low-fidelity social transmission, such as imitation/emulation, may have contributed to the ~700,000 year stasis of the Oldowan technocomplex, and (*ii*) teaching or proto-language may have been pre-requisites for the appearance of Acheulean technology. This work supports a gradual evolution of language, with simple symbolic communication preceding behavioural modernity by hundreds of thousands of years.

From 2.5 million years ago, early hominins were skilled stone knappers, capable of producing more than 70 sharp flakes from a single cobble core by striking it with a hammerstone (termed the Oldowan technocomplex[1–3]; **Figure 1a**, Supplementary Note 1). Existing remains show systematic flake detachment, maintenance of flaking angles and repair of damaged cores[4]. This complexity, along with present-day tool-making experiments[5], implies that Oldowan technology was learned and required considerable practice[1,6]. Furthermore, the technology's continual existence and wide geographic spread, along with hints of regional traditions[3,7] indicate that it was socially transmitted, although the underlying psychological mechanisms remain poorly understood[8].

Whether Oldowan stone tool making has implications for the evolution of human language and teaching (defined as active information donation[9]) is debated[10,11]. Positions range from the view that Oldowan tool making indicates a major development in hominin cognition[8], such as teaching or language[12], to the hypothesis that chimpanzee-like emulation or imitation (reproducing the object manipulations or motor patterns of others, respectively) is sufficient to transmit knapping technology[13]. Accordingly, accounts of the evolution of language range from a gradual emergence beginning 2mya[14,15], to a relatively sudden appearance 50-100kya[16]. However, a difficulty with positing complex Oldowan communication, is the apparent stasis in Oldowan technology for more than 700,000 years until Acheulean tools appear ~1.7mya[17,18]. The absence of clear cultural change during this window seems inconsistent with the presence of language, and remains an outstanding mystery more generally[19].

Across disciplines, researchers are increasingly turning to gene-culture co-evolutionary accounts to explain the evolution of human cognitive abilities, including teaching and language[10,13,20–31]. Central to such hypotheses is the idea that cultural traits can both shape, and be shaped by, genetic evolution, and a number of examples of gene-culture co-evolution are now known from human evolution[26–30]. Hominin stone tool manufacture is a particularly interesting candidate case as the appearance of such technology 2.5mya - at the dawn of *Homo* - and its continued deployment for millions of years, means it could have played a protracted role in human evolution. Furthermore, due to the challenging ecological niche that early hominins occupied[20,32] and the difficulty of acquiring tool-making skills[6], fitness benefits were likely associated with the ability to make and deploy effective cutting tools[32] as well as the ability to rapidly transmit the skills[33], and so a co-evolutionary relationship between tool making and cognition, specifically teaching and language, would seem plausible. Accordingly, Oldowan stone tool production could have generated selection for more complex forms of social transmission that enhanced the fidelity of information transmission. This could have resulted in a form of social transmission sufficient to transmit Acheulean technology reliably, and which would then generate selection for further increases in the complexity of social transmission, and so on. If this hypothesis is correct, changes in hominin cognition, including those underlying the appearance of Acheulean technology, could have depended upon selection generated by a reliance on Oldowan technology. In support of this hypothesis, archaeological remains show that changes to hominin morphology, including increased overall brain size, follow the advent of Oldowan tool making[3]. Other recent work has linked the cultural evolution of technologies to the capacity for high-fidelity social transmission[9,33–35]. However, hitherto such studies have either been theoretical or limited to somewhat artificial and abstract tasks. Accordingly, whether hominin lithic technology and social transmission genuinely represents a case of gene-culture co-evolution is currently unclear.

Experiments with contemporary humans have provided insights into the cognitive and motor processes supporting lithic technology[23,24], and could also establish which mechanisms support its transmission. However, research on the social transmission of tool making is very limited. For instance, a review of Acheulean tool-making found that reduction strategies were highly consistent across individuals[36]. The authors suggest "true imitation" (i.e. reproducing the motor pattern of another individual through observational learning) is the minimal form of social transmission that could produce such consistency[36]. Furthermore, an unpublished experimental study found that "demonstrative gestures" were sufficient for the co-operative procurement and initial reduction of bedrock slabs[37]. Only two studies have directly investigated the ability of contemporary adult humans to make tools following different means of social transmission, both comparing the efficacy of speech with symbolic gestural communication. One investigated the acquisition of Levallois technology[38] (a complex technology prevalent from 300-30kya) and reported no differences between the conditions. However, the measure of performance was a binary (yes/no) assessment by the experimenter, leaving the possibility that more subtle differences existed but were undetected. The second investigated bifacial knapping[39] (a technique associated with Acheulean technology). Whilst the tools produced in both conditions showed similar shape, symmetry and quality, the two groups used different techniques, with verbally taught participants more accurately replicating the technique of the instructor (even though they lacked the skill to enact it effectively)[39]. As verbal and gestural communication are both symbolic forms of communication, further differences may yet emerge if a wider range of social transmission mechanisms, including imitation, emulation, and subtle forms of pedagogy, are considered. This is particularly relevant to the manufacture of Oldowan technology, where the debate over the underlying transmission mechanisms is at its fiercest.

Here we present a large-scale experimental study testing the capability of five social learning mechanisms to transmit Oldowan stone knapping techniques across multiple transmission events. By establishing the relative rates of transmission resulting from different means of communication, we aimed to provide insights into which forms of communication might have been selected for as a result of reliance on tool use. The mechanisms investigated are summarised as (*i*) reverse engineering, (*ii*) imitation/emulation, (*iii*) basic teaching, (*iv*) gestural teaching and (*v*) verbal teaching (**Figure 1b-f**). In total, 184 participants took part, producing over 6000 pieces of flint, each of which was weighed, measured and assessed for quality using a novel metric that we developed and verified. We find that, across six measures, performance increases with teaching and, particularly, language. However, there is little evidence that imitation/emulation enhances transmission. Our findings support a gene-culture co-evolutionary account human evolution in which reliance on Oldowan tools would have generated selection favouring teaching and, ultimately, language. We suggest that Oldowan cultural evolution was limited, in part, by low-fidelity social transmission mechanisms. The appearance of Acheulean tools indicates the evolution of higher-fidelity social transmission, with teaching and/or some basic form of symbolic communication as plausible candidates. Accordingly, this work supports an early origin for language.

Results

Performance across conditions. Across numerous measures of individual performance we consistently found that teaching and language, but not imitation or emulation, enhanced the acquisition of stone knapping skills relative to reverse engineering (see **Table 1**). For instance, total flake quality only showed clear improvement with gestural or verbal teaching (**Figure 2a**), with language nearly doubling performance relative to reverse engineering, and also improving performance relative to imitation/emulation and basic teaching. The number of viable flakes produced shows a similar pattern (**Figure 2b**), with substantial increases relative to reverse engineering requiring gestural or verbal teaching. Moreover, unlike all forms of teaching, imitation/emulation did not increase the proportion of flakes that were viable (**Figure 2c**). Neither was there evidence for an increase in the rate of manufacture of viable flakes with imitation/emulation; only verbal teaching was clearly associated with an increase (**Figure 2d**). Similarly, only verbal teaching led to a clear increase (>30%) in the volume of core reduced (**Figure 2e**). Finally, whilst there was no evidence that imitation/emulation increased the probability of a viable flake per hit, gestural teaching doubled and verbal teaching quadrupled this probability (**Figure 2f**). Across the six measures there is strong evidence that verbal teaching increases performance relative to gestural teaching. Thus, teaching, but particularly verbal teaching, greatly facilitated the rapid transmission of flaking, whilst there is little evidence that imitation/emulation did so.

Performance along chains. In all conditions, as expected, performance decreased along chains relative to the trained experimenter as information was lost. However, with teaching, transmission was sufficiently improved that performance declined steadily along chains, whereas without teaching, the drop in performance along chains was so severe that performance immediately fell to floor levels (i.e., the minimal level of performance we observed, likely representing participants' intuitive understanding of stone knapping). For instance, with verbal teaching, the probability that each hit produced a viable flake (**Figure 2g**), the number of viable flakes produced, and the proportion of flakes that were viable (**Figure 2h**) all decreased steadily along chains, approaching the baseline performance observed with reverse engineering and imitation/emulation (see **Table 2**). Analyses of the utterances by participants in the verbal teaching condition showed that both the total number of utterances spoken and the proportion of teaching-related utterances that were correct also decreased along the chain (**Figure 2i**). The rate of decline varied with topic, with knowledge of both the exterior platform angle and force-carrying ridges rapidly lost, but information concerning the platform edge being preserved for longer and with greater accuracy.

For a full listing of all model estimates see Supplementary Tables 1-6.

Discussion

The central finding of this work is that the social transmission of Oldowan technology is enhanced by teaching, and in particular, by language. This is in line with a gene-culture co-evolutionary account of human evolution and supports the hypothesis that Oldowan stone tool manufacture generated selection favouring increasingly complex teaching and language[13,24,40]. Although the learning period in this experiment (at five mintues long) is clearly unrealistically short compared to the length of time that Oldowan hominins likely had available to learn, particularly given available data showing that precise control of conchoidal fracture can take decades to acquire[41] and anthropological data showing that knapping skills are acquired across an apprenticeship lasting several years[42], a short learning period is sufficient to examine the relative rates of transmission, which is the focus of this work. As such, we cannot rule out the possibility that with a longer learning period, performance across conditions would have converged. However, given that knapping skills are known to take years to develop fully[6,41], we suspect that increasing the time spent learning would initially only increase the differences in performance across conditions, with any convergence only occurring after extensive learning. Given their magnitude, the observed differences in performance between conditions would likely translate into significant fitness differences in the shorter term. Key to our findings' support of a gene-culture co-evolutionary account of human technology and cognition is the continuous improvement in the rate of transmission observed with increasingly complex forms of communication. For example, if verbal teaching provided transmission benefits, but simpler forms of teaching did not, then the co-evolutionary process would not be able to account for the evolution of these simpler forms of teaching. Likewise, if the transmission of tool technology benefitted from simple teaching, but gained no further benefit from verbal teaching, then the co-evolutionary process would stop with simpler forms of teaching and could not explain the evolution of verbal teaching.

Accordingly, our data imply that Oldowan tool-making would have created a continuous selective gradient leading from observational learning to much more complex verbal teaching. This process need not have taken place entirely within the Oldowan, but was probably already underway during the Oldowan and likely continued well after, as Oldowan tools continued to be made for hundreds of thousands of years beyond the Oldowan time period. Furthermore, assuming that the transmission of more complex technologies also benefits from more complex means of communication, later technologies would have reinforced the gene-culture co-evolutionary dynamic. Such a process could have lasted for millions of years (and may be ongoing[29]), with more complex communication allowing the stable and rapid transmission of increasingly complex technologies, which in turn generate selection for even more complex communication and cognition, and so forth. Whilst this places little necessary constraint on when teaching and language may have evolved, our central contribution is to provide evidence that Oldowan tools, produced by hominins since at least 2.5my, were involved in this dynamic.

A second significant finding of this work is that the rate of transmission of Oldowan tool making is, at best, minimally enhanced by the addition of imitation/emulation relative to reverse engineering. That the low level of performance with imitation/emulation and reverse engineering is stable along chains (and that performance with teaching and language collapses to this level) suggests a baseline level of performance reliant on little transmitted knowledge, and which could well be achieved through intuition and individual trial-and-error learning. We suggest that the rapid decline of performance with teaching and language to this baseline merely reflects the short learning time employed in this study. Previous transmission chain studies have established that periods of individual practice can bolster the stability of socially transmitted knowledge[43]. This suggests that with more time to learn, with bouts of teaching and language integrated with periods of individual practice, the benefits of teaching and language would likely have been preserved for longer. Likewise, a benefit of observational learning relative to reverse engineering may well appear over a longer learning period. However, our data suggest that any such benefit is likely to be less than the benefit that would be derived through teaching across a similar timespan due to the improved rate of transmission with teaching. Accordingly, while we do not suggest that imitation is insufficient to transmit the technology *per se*, our findings supports other recent work in implying that observation alone is an inefficient means to acquire stone tool making skills[23,44,45].

Limited information concerning tool manufacture can, no doubt, be rapidly acquired through imitation or emulation, for instance, the basics of core, hammerstone or flake selection[36], the requirement to strike the core with the hammerstone, and some idea of the force required. However, it seems plausible that the rapid striking action associated with tool manufacture hinders the transmission of the more subtle information crucial to knapping, such as details of the point of percussion or the platform edge and angle, through observation alone. It is here that teaching (e.g. slowing down the striking action, pointing to appropriate targets, demonstrating core rotation, manual shaping of pupil's grasp) and verbal instruction likely provide immediate benefits to the pupil. Indeed, transcripts from the verbal teaching condition show that abstract knapping concepts, such as the platform angle, were transmitted between individuals in the verbal teaching condition (see Supplementary Figure 3). It may well be the capacity for arbitrary labels such as "platform angle" that facilitates transmission with verbal teaching; such labels break the task into constituent parts, can be used to identify the important elements and provide a clear framework with which pupils can go on to teach others. Language not only allows transmission of the skill itself, but also the ability to transmit the skill to others effectively.

Thirdly, our findings have implications for one of the most enduring puzzles of human evolution; the apparent stasis of the Oldowan technocomplex, which lasted 700,000 years[8,11,19,45]. Our experiment suggests that Oldowan technological change could have been restricted by low-fidelity forms of social transmission that prevented the spread of innovations. This suggestion is supported by the slow spread of Oldowan technology across Africa which indicates that this technology was difficult for Oldowan hominins to transmit[3]. Furthermore, the acquisition of Oldowan knapping skills is not trivial even for modern humans, as shown by our finding that the benefits of teaching and language were rapidly lost in transmission. Whilst we cannot conclusively identify what form Oldowan transmission might have taken, our data indicate imitation or emulation as likely candidates. In naturalistic contexts, the relatively poor transmission that we observed with imitation and emulation could well be too slow and imprecise for innovations to be transmitted reliably, leaving the technology unable to increase in complexity until more effective communication had evolved.

The suggestion that low-fidelity social transmission is a limiting factor on technological development might contribute to an understanding of why human culture is so complex compared to the behavioural traditions of non-human animals[46,47]. Whilst human social transmission has allowed the cumulative elaboration of a vast number of technologies and behaviours, non-human animal social transmission has not. It seems possible that this is because non-human animal social transmission, which appears to be largely limited to forms of observational learning less sophisticated than those of humans[43], lacks the fidelity required to transmit more complex innovations, thus constraining cumulative cultural evolution[34,35,48]. Even the modest knapping ability of extensively trained bonobos[49,50] may rely on their prior training in symbolic communication[51]. Whilst it is plausible that a similar co-evolutionary process has operated to a lesser degree in some other species, such as other apes[52], it remains an open question as to why their tool use did not generate selection for the higher-fidelity social transmission (teaching, language) observed in humans. One possibility is that the technologies of other apes are either sufficiently simple that they can be acquired through more basic mechanisms or so hard to acquire that they can only rarely be transmitted successfully, removing the benefit to teaching[9]. Task difficulty might also explain a previous experimental finding that simple transmission mechanisms were sufficient for cumulative cultural evolution in the context of human paper-plane design[53]; this task may be sufficiently simple that teaching is of little benefit. Alternatively, ape reliance on tool use could be insufficient for the benefits of tool-use to outweigh the costs of complex social transmission, thus preventing teaching from increasing fitness[9]. Any of these constraints would undermine selection for higher-fidelity social transmission, hindering the co-evolutionary process.

Given that our findings support a co-evolution of Oldowan tool use and complex communication, it might seem puzzling that the Oldowan stasis should last so long. If the selective advantage was present, why did more complex communication not evolve for 700,000 years? A likely explanation is that more complex communication may well have evolved during the Oldowan, but that this alone was insufficient for the evolution of stone tool technology. The appearance of Acheulean tools may have additionally been contingent on the evolution of other aspects of cognition, such as technical comprehension or the hierarchical planning of actions[54–56], as well as demographic and socio-ecological factors[57,58]. Accordingly, the extraordinary length of the Oldowan stasis could indicate that a large number of limiting factors needed to be overcome before innovations could appear and spread.

Given this, our findings imply that the appearance of Acheulean tools 1.7mya[17,18] reflects, in part, the evolution of mechanisms of transmission that facilitated the more effective transmission of Oldowan tools, but also enabled the reliable transmission of the sub-goals and techniques required to make the distinctive and regularly-shaped Acheulean tools[59]. We cannot specify the form of this transmission with precision. However, given the observation that chimpanzees are capable of some form of observational learning, yet cannot produce stone tools approaching the quality of the earliest known Oldowan examples[13], combined with the complexity of Acheulean technology[36], we suggest that teaching in the form of facilitated observation (similar to our basic teaching condition) is the minimal plausible form of social transmission for Acheulean hominins and that rudimentary forms of language are a possibility. However, whilst our findings suggest that Oldowan hominins would have benefitted from modern language, the suggestion that modern language evolved during the Oldowan seems unlikely given how slowly technology evolved thereafter. This leaves open the possibility that the transmission of Acheulean technology was reliant on a form of (gestural or verbal) proto-language[12,60,61]. This need not imply that Acheulean hominins were capable of manipulating a large number of symbols or generating complex grammars. Our findings imply that simple forms of positive or negative reinforcement, or directing the attention of a learner to specific points (as was common in the gestural teaching condition), are considerably more successful in transmitting stone knapping than observation alone. This is supported by existing theoretical work that suggests positive and negative feedback greatly enhances the rate of transmission[33]. Whether or not simple symbolic communication was present during the Acheulean, we anticipate that the gene-culture co-evolutionary dynamic between tools and communication was, and that it would continue beyond the Acheulean, generating selection favouring the use of symbols for increasingly subtle and abstract concepts, and contributing to the eventual evolution of modern language capabilities.

In sum, our data support the hypothesis that a gene-culture co-evolutionary dynamic between tool use and social transmission was on-going in human evolution, starting at least 2.5mya and potentially continuing to the present. The simplicity and stasis of Oldowan technology is indicative of a limited form of social transmission, such as observational learning, that only allowed the transmission of the broadest concepts of stone knapping technology. Whatever its nature, this was sufficient to support limited transmission amongst individuals with prolonged contact, but insufficient to propagate innovations more rapidly than they were lost, and would have contributed to the stasis in the Oldowan technocomplex. However, hominin reliance on stone technology would have generated selection for increasingly complex communication that allowed the more effective spread of stone-tools. Under this continued selection, teaching, symbolic communication and eventually verbal language may have been favoured, allowing the ready transmission of abstract flaking concepts, such as the role of the exterior platform angle in choosing where to strike[38], which our findings show are effectively transmitted by language. Given the increased complexity of the later Acheulean and Mousterian lithic technologies, with their reliance on "long sequences of hierarchically organised actions"[36,38] and other abstract concepts, our results imply that hominins possessed a capacity for teaching - and potentially simple proto-language - as early as 1.7mya.

Methods

Participants and materials. 184 participants took part in the study. This sample size was chosen based on effect sizes observed in previous transmission chain studies. Participants were students at the University of St Andrews recruited through the University's experimental sign-up system. Across the experiment we used 2 tonnes of Brandon flint from Norfolk, UK, broken up into cores of roughly 1kg. We also used 100 granite hammerstones collected from the coastline near Stonehaven, Scotland.

Experimental design. Adult human participants (N=184) first learned, were tested on their ability, and then helped others to learn, to knap stone flakes using a granite hammerstone and flint core, across five cumulatively complex transmission conditions (see **Figure 1 b-f**): (1) **Reverse Engineering**; pupils were provided with a core and hammerstone for practice, but saw only the flakes manufactured by their tutor and not their tutor themselves; (2) **Imitation/Emulation**; in addition to having their own core and hammerstone, pupils also observed their tutor making flakes, but could not interact with them; (3) **Basic Teaching**; in addition to demonstrating tool production, tutors could also manually shape the pupil's grasp of their hammerstone or core, slow their own actions, and reorient themselves to allow the pupil a clear view (this condition replicates teaching reported in non-human primates[62]); (4) **Gestural teaching**; tutors and pupils could also interact using any gestures, but no vocalisations; and (5) **Verbal Teaching**; tutors and pupils were also permitted to speak. Participants were assigned to conditions at random and blinding was not possible. The test given to participants to assess their ability was to make as many good-quality flakes as possible from a single core. This reflected pressures on hominin knappers to make the most of the limited availability of high quality knapping materials.

Participants were arranged into transmission chains[63] in which information was passed along chains of participants, with each participant learning from the previous participant and acting as tutor to the next participant. For each condition we carried out four short chains (≤5 participants) and two long chains (≤10 participants) per condition (see **Figure 1g**). Experimenters trained in stone knapping (TM, NU) acted as tutor to the first participant.

To ensure participant motivation, we paid participants between £10 and £20, with the value dependent upon their performance when tested. In the teaching conditions (conditions 3-5) participants' payment was also dependent upon how well their pupils went on to perform, thus tutors were motivated to teach effectively. In the imitation/emulation condition (condition 2) participants' payment was also dependent upon how well they performed when demonstrating, this was to motivate demonstrators to focus on their own performance and not to teach the pupil.

Procedure. Upon arrival, participants were briefed on the experimental procedure and their consent was required to proceed (ethical approval was given by St Andrews UTREC, code: BL6376). Before they learnt to knap, and to ensure that participants understood what Oldowan tools were used for, participants were given an information sheet, flint flakes of varying quality, chamois leather and wooden sticks. They were then given 5 minutes to use these items to gain an understanding of what made a good-quality sharp cutting flake. The information sheet gave only very brief information on the history and uses of Oldowan stone tools, and not any information as to how to make them beyond striking a flint core with a hammerstone.

The learning/teaching period lasted for five minutes, after which participants were interrupted. After the learning phase, the pupil then advanced to the test phase. Participants were instructed to take as long as they needed for the test phase, however, if they had not stopped within 18 minutes the experimenter encouraged them to finish and after 20 minutes the experimenter instructed them to stop (only 12.5% of participants used the full 20 minutes). After the test phase (if applicable) participants went on to teach the next pupil. Once the procedure was complete, participants were debriefed and paid before leaving.

Data. All flint used by participants was bagged throughout the experiment. In total, participants produced 6214 pieces of flint greater than 2cm across. All of these pieces were weighed, measured, and assessed for viability (i.e., whether they had possible use as a cutting tool) and quality (using a novel metric, which we developed, that took into account flake mass, cutting edge length and diameter; see Supplementary Methods for details). Any pieces less than 2cm across were not coded, as 2cm was considered to be the minimum size for a flake to possibly have utility as a butchery tool[64]. We also weighed participants' cores both before and after knapping. Participants' behaviour during the experiment was recorded using video cameras and we subsequently measured the length of time participants spent knapping and the number of times participants struck their core with their hammerstone. We also transcribed everything participants said whilst in the verbal teaching condition and split it into utterances (N=1481) for analysis. In particular all utterances were coded as either "correct" or "incorrect" which was determined relative to established knapping practices. The robustness of flake viability ratings as well as video coding, were tested by triple and double coding, respectively, a subset of the data. In both cases the level of agreement between coders was very high (see Supplementary Methods for details of the double/triple coding procedure).

Analyses. We analysed the data using Bayesian GLMMs fitted using MCMC methods in OpenBUGS[65,66]. We modelled six different measures of individual performance: 1) the number of viable flakes produced, 2) the total quality of flakes produced, 3) the proportion of flakes that were viable, 4) the rate at which viable flakes were produced, 5) the probability of a viable flake per hit and 6) the proportion of their core successfully reduced. These measures were modelled as a function of condition, position along the chain, interactions between condition and position, initial core mass and random repeat-level effects.

For a full description of the experimental procedure and all analyses see Supplementary

Methods. For a comparison of the model results with the raw data see Supplementary Figures 1 and 2.

References

1. Roche, H. *et al.* Early hominid stone tool production and technical skill 2.34 Myr ago in West Turkana, Kenya. *Nature* **399,** 57–60 (1999).

2. Semaw, S., Renne, P., Harris, J. W. K. & Feibel, C. S. 2.5-million-year-old stone tools from Gona, Ethiopia. *Nature* **385,** 333–336 (1997).

3. Schick, K. & Toth, N. in *Oldowan Case Stud. into earliest stone age* (Toth, N. & Schick, K.) (Gosport: Stone Age Institute, 2006).

4. Delagnes, A. & Roche, H. Late Pliocene hominid knapping skills: the case of Lokalalei 2C, West Turkana, Kenya. *J. Hum. Evol.* **48,** 435–72 (2005).

5. Toth, N. Behavioral inferences from early stone artifact assemblages: an experimental model. *J. Hum. Evol.* **16,** 763–787 (1987).

6. Callahan, E. *The basics of biface knapping in the eastern fluted point tradition: A manual for flintknappers and lithic analysts*. (Eastern States Archaeological Federation, 1979).

7. Braun, D. R., Plummer, T., Ditchfield, P. W., Bishop, L. C. & Ferraro, J. V. in *Interdiscip. Approaches to Oldowan* (Hovers, E. & Braun, D. R.) 99–110 (Springer, 2009).

8. Hovers, E. Invention, reinvention, and innovation: The makings of Oldowan lithic technology. *Orig. Hum. Innov. Creat.* **16,** 51–68 (2012).

9. Fogarty, L., Strimling, P. & Laland, K. N. The evolution of teaching. *Evolution.* **65,** 2760–2770 (2011).

10. Gibson, K. & Ingold, T. *Tools, language and cognition in human evolution*. (Cambridge University Press, 1993).

11. Ambrose, S. H. Paleolithic Technology and Human Evolution. *Science (80-.).* **291,** 1748–1753 (2001).

12. Bickerton, D. *Adam's Tongue*. (Hill and Wang, 2009).

13. Wynn, T., Hernandez-Aguilar, A., Marchant, L. F. & McGrew, W. C. "An ape's view of the Oldowan" revisited. *Evol. Anthropol.* **20,** 181–97 (2011).

14. Belfer-Cohen, A. & Goren-Inbar, N. Cognition and communication in the Levantine Lower Palaeolithic. *World Archaeol.* **26,** 144–157 (1994).

15. D'Errico, F. *et al.* Archaeological Evidence for the Emergence of Language, Symbolism, and Music — An Alternative Multidisciplinary Perspective. *J. World Prehistory* **17,** 1–70 (2003).

16. Mellars, P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl. Acad. Sci.* **103,** 9381–9386 (2006).

17. Beyene, Y. *et al.* The characteristics and chronology of the earliest Acheulean at Konso, Ethiopia. *Proc. Natl. Acad. Sci.* **110,** 1584–1591 (2013).

18. Lepre, C. J. *et al.* An earlier origin for the Acheulian. *Nature* **477,** 82–5 (2011).

19. De la Torre, I. The origins of stone tool technology in Africa: a historical perspective. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366,** 1028–37 (2011).

20. Blumenschine, R. J. *Early hominid scavenging opportunities: implications of carcass availability in the Serengeti and Ngorongoro ecosystems*. (B. A. R., 1986).

21. Enquist, M., Ghirlanda, S., Jarrick, A. & Wachtmeister, C.-A. Why does human culture increase exponentially? *Theor. Popul. Biol.* **74,** 46–55 (2008).

22. Sterelny, K. Language, gesture, skill: the co-evolutionary foundations of language. *Philos. Trans. R. Soc. B* **367,** 2141–2151 (2012).

23. Stout, D., Toth, N., Schick, K., Stout, J. & Hutchins, G. Stone tool-making and brain activation: position emission tomography (PET) studies. *J. Archaeol. Sci.* **27,** 1215–1223 (2000).

24. Uomini, N. T. & Meyer, G. F. Shared brain lateralization patterns in language and Acheulean stone tool production: a functional transcranial Doppler ultrasound study. *PLoS One* **8,** e72693 (2013).

25. Boyd, R., Richerson, P. J. & Henrich, J. The cultural niche: why social learning is essential for human adaptation. *Proc. Natl. Acad. Sci.* **108 Suppl,** 10918–25 (2011).

26. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39,** 31–40 (2007).

27. Durham, W. H. *Coevolution: Genes, Culture and Human Diversity*. (Stanford University Press, 1991).

28. Hünemeier, T. *et al.* Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. *PLoS One* **7,** e38862 (2012).

29. Laland, K. N., Odling-Smee, J. & Myles, S. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat. Rev. Genet.* **11,** 137–148 (2010).

30. Richerson, P. J., Boyd, R. & Henrich, J. Gene-culture coevolution in the age of genomics. *Proc. Natl. Acad. Sci. U. S. A.* **107 Suppl,** 8985–92 (2010).

31. Feldman, M. W. & Laland, K. N. Gene-culture coevolutionary theory. *Trends Ecol. Evol.* **5347,** 453–457 (1996).

32. Potts, R. Hominin evolution in settings of strong environmental variability. *Quat. Sci. Rev.* **73,** 1–13 (2013).

33. Castro, L. & Toro, M. a. The evolution of culture: from primate social learning to human culture. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 10235–40 (2004).

34. Dean, L. G., Kendal, R. L., Schapiro, S. J., Thierry, B. & Laland, K. N. Identification of the social and cognitive processes underlying human cumulative culture. *Science* **335,** 1114–8 (2012).

35. Lewis, H. M. & Laland, K. N. Transmission fidelity is the key to the build-up of cumulative culture. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 2171–80 (2012).

36. Shipton, C. B. K., Petraglia, M. & Paddayya, K. Stone tool experiments and reduction methods at the Acheulean site of Isampur Quarry, India. *Antiquity* **83,** 769–785 (2009).

37. Petraglia, M., Shipton, C. B. K. & Paddayya, K. in *Hominid Individ. Context Archaeol. Investig. Low. Middle Palaeolithic landscapes, locales artefacts.* (Gamble, C. & Porr, M.) (Routledge, 2005).

38. Ohnuma, K., Aoki, K. & Akazawa, T. Transmission of Tool-making through Verbal and Non-verbal Communication--Preliminary Experiments in Levallois Flake Production. *Anthropol. Sci.* **105,** 159–168 (1997).

39. Putt, S. S., Woods, A. D. & Franciscus, R. G. The Role of Verbal Interaction During Experimental Bifacial Stone Tool Manufacture. *Lithic Technol.* **39,** 96–112 (2014).

40. Stout, D. in *Stone tools Evol. Hum. Cogn.* (Nowell, A. & Davidson, I.) 159–184 (University Press of Colorado, 2010).

41. Nonaka, T., Bril, B. & Rein, R. How do stone knappers predict and control the outcome of flaking? Implications for understanding early stone tool technology. *J. Hum. Evol.* **59,** 155–167 (2010).

42. Stout, D. Skill and Cognition in Stone Tool Production: An Ethnographic Case Study from Irian Jaya. *Curr. Anthropol.* **43,** 693–723 (2002).

43. Hoppitt, W. J. E. & Laland, K. N. *Social Learning: An introduction to mechanisms, methods, and models*. 320 (Princeton University Press, 2013).

44. Uomini, N. T. The prehistory of handedness: archaeological data and comparative ethology. *J. Hum. Evol.* **57,** 411–9 (2009).

45. Stout, D., Semaw, S., Rogers, M. J. & Cauche, D. Technological variation in the earliest Oldowan from Gona, Afar, Ethiopia. *J. Hum. Evol.* **58,** 474–491 (2010).

46. Laland, K. N. & Galef Jr, B. G. The Question of Animal Culture. 320 (2009).

47. Whiten, A. The scope of culture in chimpanzees, humans and ancestral apes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366,** 997–1007 (2011).

48. Tomasello, M. in *Chimpanzee Cult.* (Wrangham, R. W., McGrew, W. C., de Waal, F. B. M. & Heltne, P. G.) (Harvard University Press, 1994).

49. Roffman, I., Savage-Rumbaugh, S., Rubert-Pugh, E., Ronen, A. & Nevo, E. Stone tool production and utilization by bonobo-chimpanzees (Pan paniscus). *Proc. Natl. Acad. Sci.* 1–4 (2012). doi:10.1073/pnas.1212855109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1212855109

50. Toth, N. & Schick, K. The Oldowan: the tool making of early hominins and chimpanzees compared. *Annu. Rev. Anthropol.* (2009).

51. Savage-Rumbaugh, S., Fields, W. M. & Spircu, T. The emergence of knapping and vocal expression embedded in a *Pan*/*Homo* culture. *Biol. Philos.* **19,** 541–575 (2004).

52. Whiten, A. & van Schaik, C. P. The evolution of animal "cultures" and social intelligence. *Philos. Trans. R. Soc. B* **362,** 603–20 (2007).

53. Caldwell, C. a. & Millen, A. E. Social Learning Mechanisms and Cumulative Cultural Evolution: Is Imitation Necessary? *Psychol. Sci.* **20,** 1478–1483 (2009).

54. Stout, D. Stone toolmaking and the evolution of human culture and cognition. *Philos. Trans. R. Soc. B* **366,** 1050–9 (2011).

55. Pelegrin, J. in *Use Tools by Hum. Non-human Primates* (Berthelet, A. & Chavaillon, J.) (Oxford University Press, 1993). doi:10.1093/acprof

56. Dennett, D. *Darwin's Dangerous Idea*. (Simon & Schuster, 1995).

57. Powell, A., Shennan, S. J. & Thomas, M. G. Late Pleistocene demography and the appearance of modern human behavior. *Science* **324,** 1298–301 (2009).

58. Potts, R. Environmental hypotheses of hominin evolution. *Am. J. Phys. Anthropol.* **27,** 93–136 (1998).

59. Gowlett, J. in *Axe Age Acheulian Tool-making from Quarr. to Discard* (Goren-Inbar, N. & Sharon, G.) (Equinox, 2006).

60. Donald, M. *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. (Harvard University Press, 1991).

61. Corballis, M. C. *The Lopsided Ape: Evolution of the Generative Mind*. (Oxford University Press, 1993).

62. Boesch, C. Teaching among wild chimpanzees. *Anim. Behav.* **41,** 530–532 (1991).

63. Mesoudi, A. & Whiten, A. The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **363,** 3489–501 (2008).

64. Key, A. J. M. & Lycett, S. J. Are bigger flakes always better? An experimental assessment of flake size variation on cutting efficiency and loading. *J. Archaeol. Sci.* **41,** 140–146 (2014).

65. Lunn, D. & Spiegelhalter, D. The BUGS project: Evolution, critique and future directions. *Stat. Med.* **28,** 3049–3067 (2009).

66. Ntzoufras, I. *Bayesian Modeling Using WinBUGS*. (Wiley, 2009).

Acknowledgements

Research supported in part by an ERC Advanced Grant to KNL (EVOCULTURE, Ref: 232823) and grants to NTU from the British Academy (Centenary Project "Lucy to Language: the Archaeology of the Social Brain") and The Leverhulme Trust (ECF 0298). We are grateful to Gillian Brown, Richard Byrne, Jane Rees and Stephen Shennan for helpful comments on earlier drafts. We would like to thank John and Val Lord for supplying us with flint.

Authors' contributions

TM, NU, LR and KL designed the experiment; TM, NU, LR, LT, SS, HL, CC and CE executed the experiment; TM, NU, IdlT and RK coded the data; TM carried out the analyses; all authors contributed to the preparation of the manuscript.

Conflict of financial interests

All authors declare that they have no conflicts of interest concerning the publication of this work.

Figures

Figure 1. Experimental design and structure. (*a*) A diagram of the stone knapping process. The hammerstone strikes the core with the goal of producing a flake. The platform edge and angle are important to the success of knapping. (*b*-*f*) The five learning conditions. (*g*) The structure of the experiment. For each condition 6 chains were carried out (4 short and 2 long); one of two trained experimenters started each chain (equally within each condition).

Figure 2. Performance across conditions and along chains. Values shown are the median model estimates and the corresponding 95% central credible intervals. More complex forms of communication, in particular verbal teaching, increased several measures of participant performance, including (*a*) the total quality of all flakes, (*b*) the number of viable flakes, (*c*) the proportion of flakes that were viable, (*d*) the rate at which viable flakes were made, (*e*) the proportion of the core knapped and (*f*) the probability that each hit resulted in a viable flake. The brackets marked with double asterisks indicate contrasts for which there is strong evidence of a difference (95% credible interval excluding 0), single asterisks indicate cases for which there is weak evidence of a difference (90% credible interval excluding 0). The red bracket in panel (*c*) indicates that the increase in performance from imitation/emulation to basic teaching is greater than the increase between all other adjacent conditions. (*g*,*h*) Although verbal and gestural teaching

increased the probability of a viable flake per hit and the proportion of flakes that were viable, performance in these conditions decreased along chains such that across conditions performance was similar by position 5. With reverse engineering, performance did not decline along chains, suggesting it was already at floor levels. Position 1 corresponds to the first participant, not the trained experimenter. (*i*) With verbal teaching, both the total number of utterances (left hand bars) and the probability a teaching utterance was correct (right hand bars) decreased along chains. Key: reverse engineering-blue (n=37), imitation/emulation-green (n=34), basic teaching-yellow (n=38), gestural teaching-orange (n=37), verbal teaching-red (n=38).

Tables

Table 1. Effects of different transmission mechanisms on performance.

Variable	Condition				
	RE	IE	BT	GT	VT
Total quality	13.0, [9.2, 17.9]	15.7, [11.1, 21.4]	15.4, [11.1, 20.7]	19.8, [14.6, 26.7]	23.6, [17.0, 31.9]
Number of viable flakes	15.76, [12.1,0.47]	18.31, [14.07,23.56]	19.56, [15.08,25.37]	21.73, [16.77,28.32]	25.22, [19.42,33.02]
Proportion of flakes that are viable	0.55, [0.48,0.62]	0.58, [0.52,0.64]	0.72, [0.66,0.77]	0.72, [0.67,0.77]	0.73, [0.68,0.78]
Viable flakes per minute	1.96, [1.33,2.87]	1.98, [1.35,2.85]	2.55, [1.78,3.69]	2.95, [2.03,4.36]	3.37, [2.26,5.19]
Proportion of core knapped	0.44, [0.35,0.54]	0.46, [0.37, 0.56]	0.53, [0.43, 0.63]	0.51, [0.43,0.62]	0.59, [0.48, 0.71]
Probability of a viable flake per hit	0.03, [0.02,0.05]	0.04, [0.03,0.06]	0.06, [0.04,0.08]	0.07, [0.05,0.10]	0.10, [0.07,0.16]

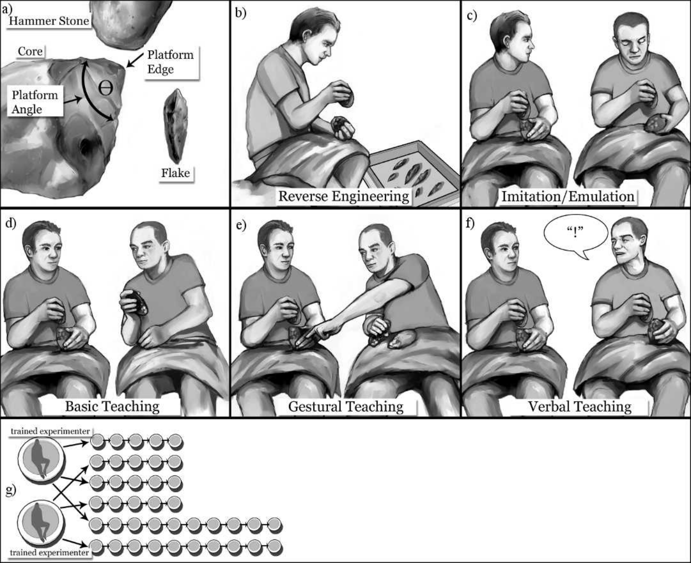
Estimated values for parameters at the first position in the chain for different conditions. Quoted values are median model estimates and their 95% central credible intervals. RE = Reverse Engineering, IE = Imitation/Emulation, BT = Basic Teaching, GT = Gestural Teaching, VT = Verbal Teaching.

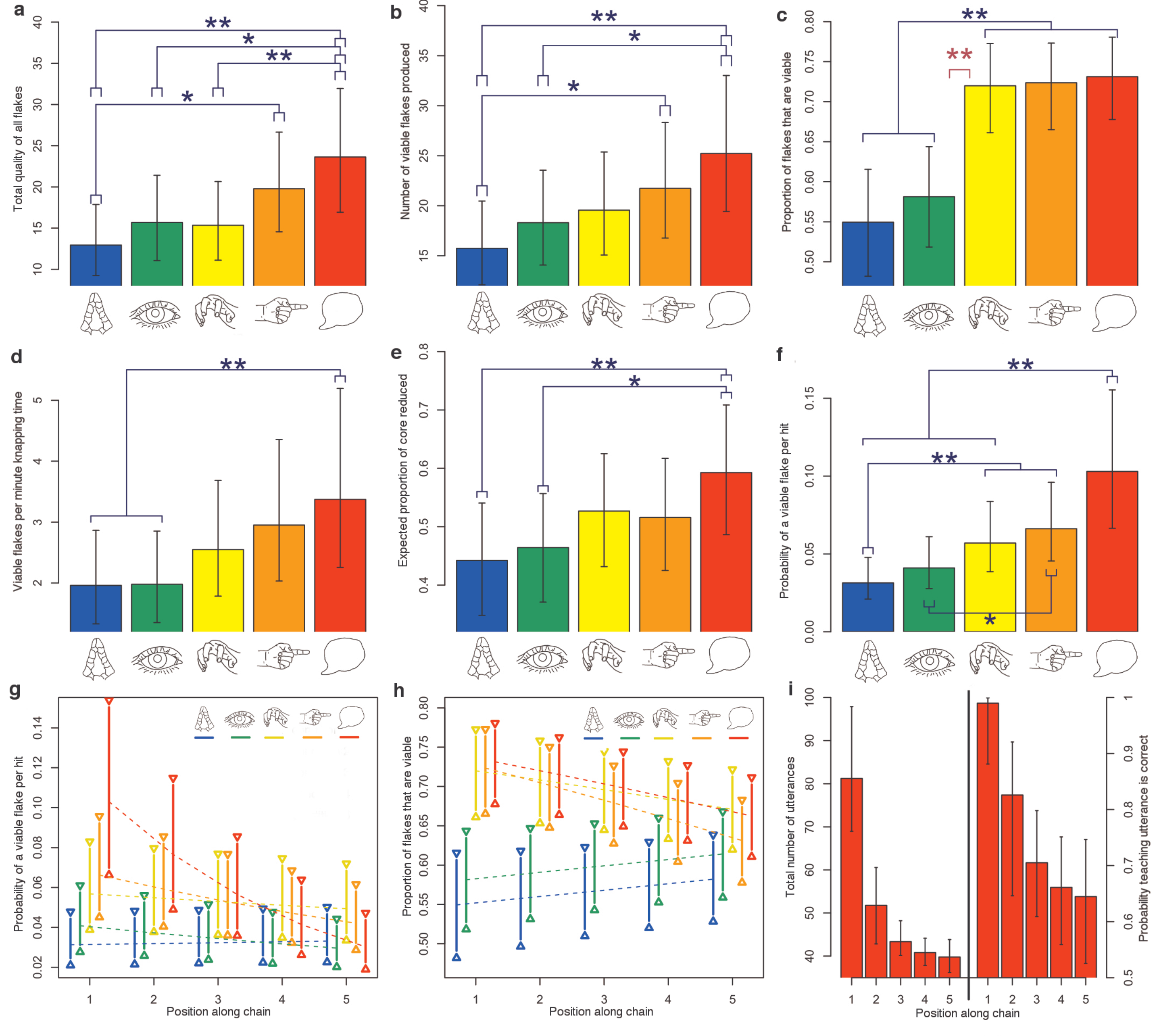
Table 2. Effects of position along chains on performance.

Variable	Condition	Gradient/rate of change	Extent of change
Number of viable flakes	VT	-0.07, [-0.10, -0.04]	-
Proportion of flakes that are viable	BT	-0.06, [-0.10, -0.01]	-
	GT	-0.11, [-0.15, -0.06]	-
	VT	-0.08, [-0.13, -0.03]	-
Probability of a viable flake per hit	IE	-0.08, [-0.12, -0.05]	-
	BT	*-0.04, [-0.08, 0.00]*	-
	GT	-0.12, [-0.16, -0.08]	-
	VT	-0.33, [-0.38, -0.28]	-
Total Utterances	VT	1.2, [0.63, 14.0]	-42.2, [-29.3, -58.9]
Proportion of teaching utterances correct	VT	1.4, [0.56, 45.8]	-4.0, [-1.4, -6.9]
Platform angle teaching accuracy	VT	3.99, [0.0, 128.1]	-0.75, [3.21, -1.91]
Ridge teaching accuracy	VT	0.42, [0.1766, 1.10]	-3.69, [-1.95, -6.75]
Platform edge teaching accuracy	VT	0.00, [0.0, 0.09]	1.18, [4.78, -4.12]
Force required teaching accuracy	VT	0.00, [0.0, 0.03]	0.53, [4.73, -3.489]

Quoted values are median model estimates and their 95% central credible intervals.

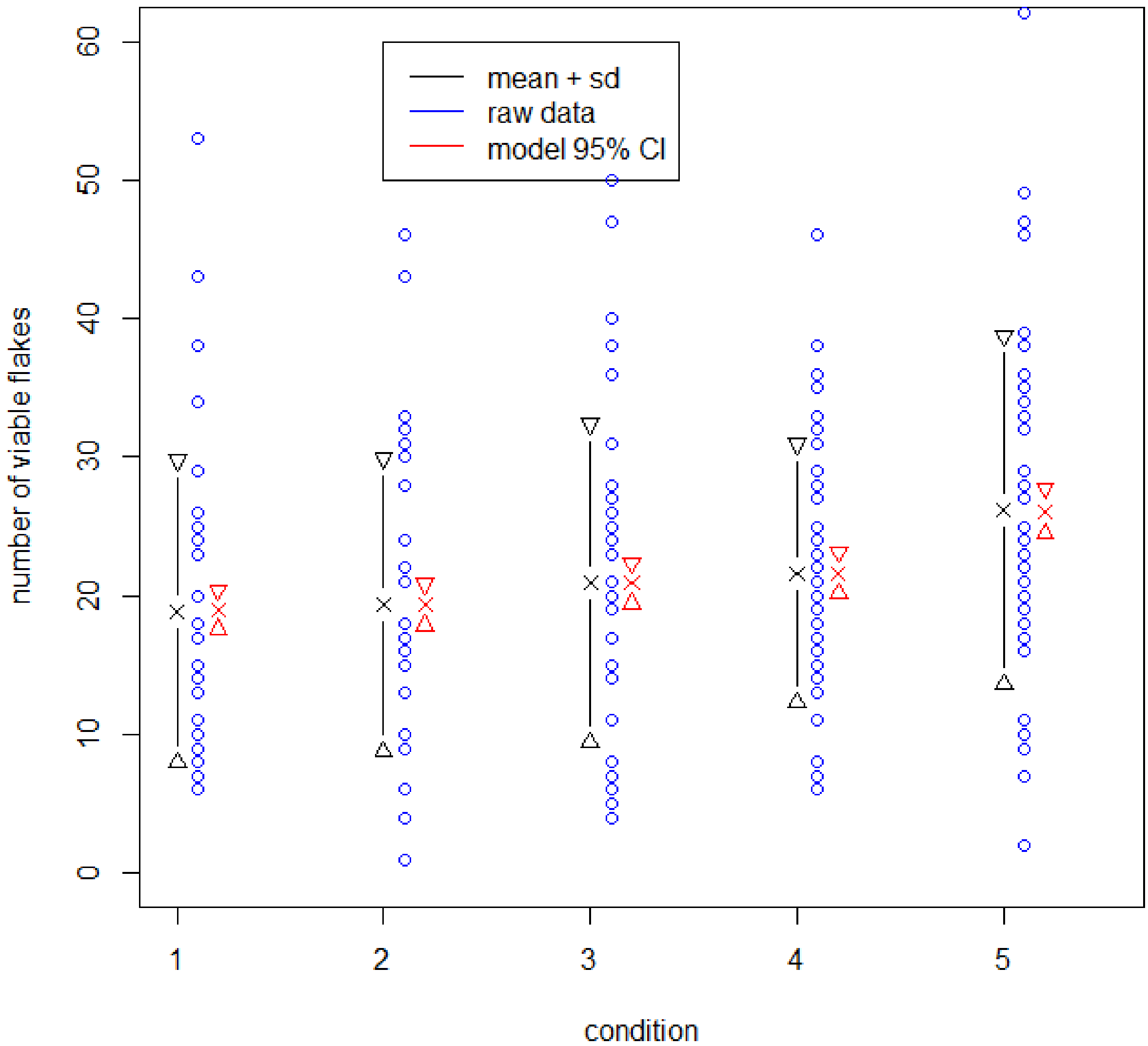
Where only the gradient is given, a negative change corresponds to a decrease along chains; where both rate and extent are given, the rate is a scalar quantity and a negative extent corresponds to a decrease along chains. Values in italics represent cases where the 95% credible interval did not exclude 0, but the 90% interval did (i.e., weak, but not strong evidence). RE = Reverse Engineering, IE = Imitation/Emulation, BT = Basic Teaching, GT = Gestural Teaching, VT = Verbal Teaching.


a)
Hammer Stone
Core
Platform Edge
Platform Angle
Θ
Flake
b)
Reverse Engineering
c)
Imitation/Emulation
d)
Basic Teaching
e)
Gestural Teaching
f)
"!"
Verbal Teaching
g)
trained experimenter
trained experimenter

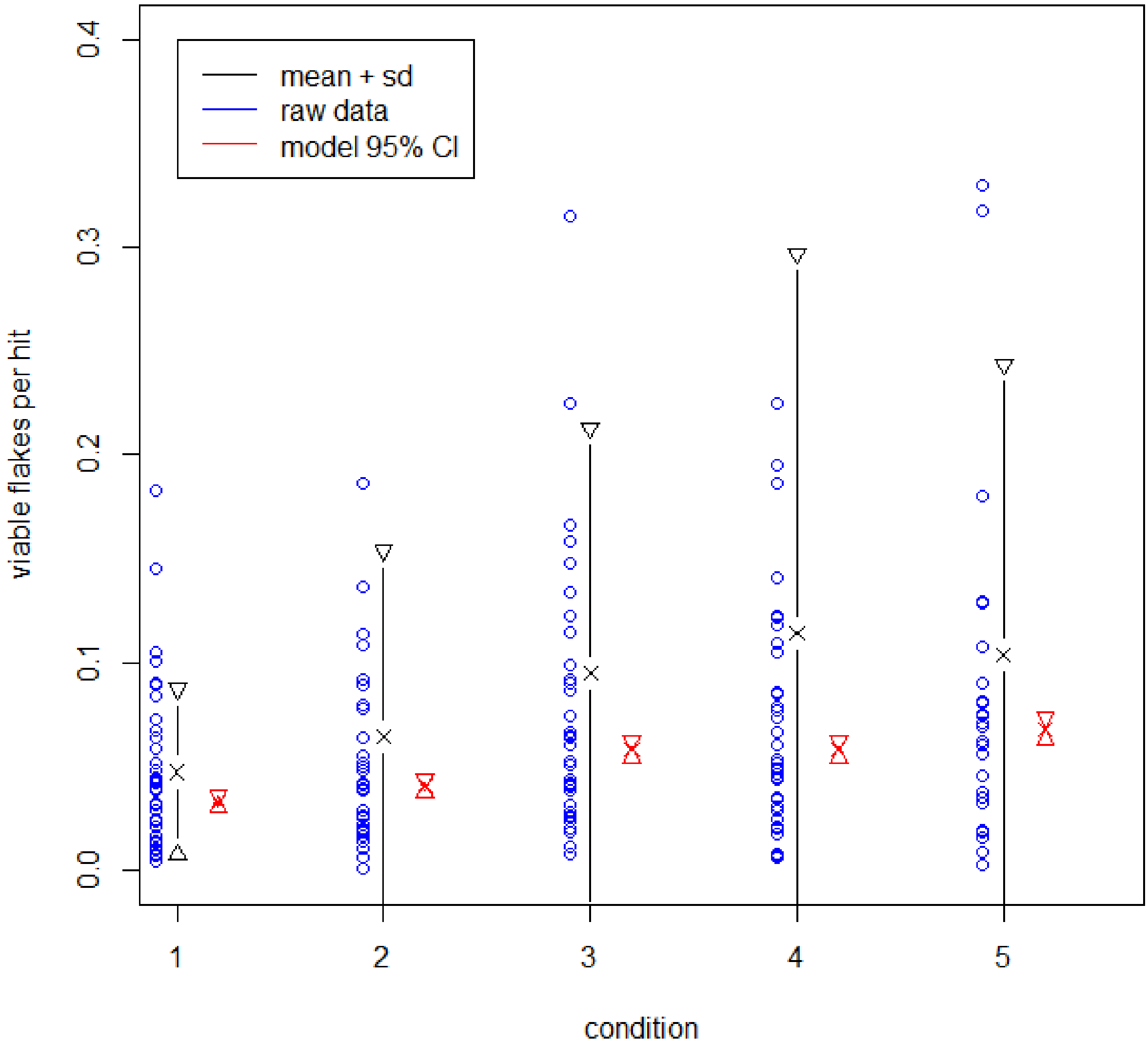


a
Total quality of all flakes
b
Number of viable flakes produced
c
Proportion of flakes that are viable
d
Viable flakes per minute knapping time
e
Expected proportion of core reduced
f
Probability of a viable flake per hit
g
Probability of a viable flake per hit
Position along chain
h
Proportion of flakes that are viable
Position along chain
i
Total number of utterances
Probability teaching utterance is correct
Position along chain

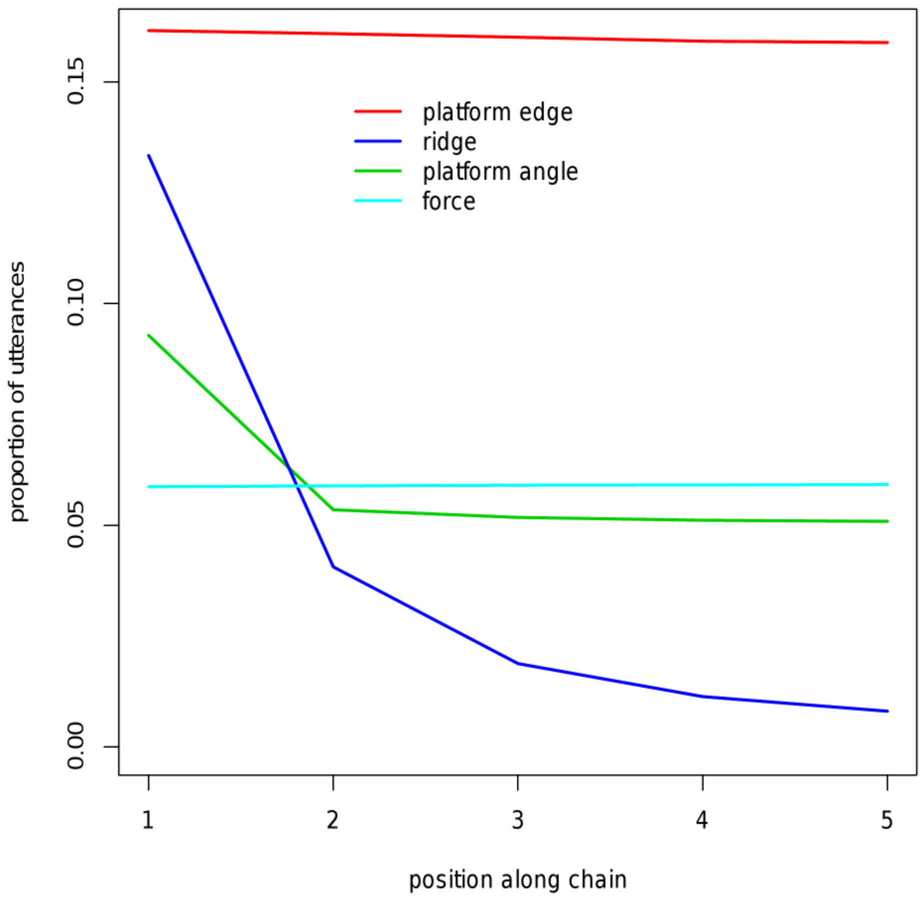
Supplementary Figures



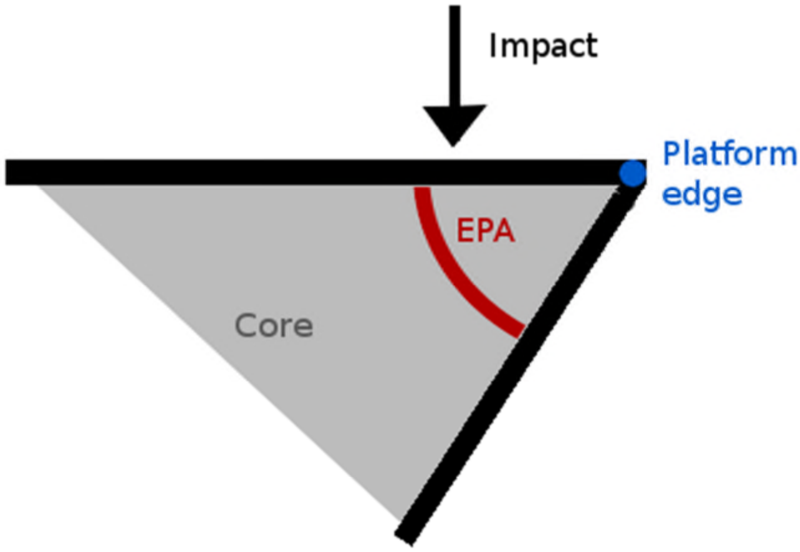
Supplementary Figure 1: A comparison of the raw data and model estimates. This figure shows the raw data (blue dots), raw data average +/- one standard deviation (black interval) and median model estimate with 95% central credible interval of the raw data average (red interval) for the total number of viable flakes produced by participants across the five conditions. As can be seen the model is very accurate at estimating the raw data average and does so with a high degree of certainty as the model intervals are much narrower than the standard deviation interval. This can give us high confidence in the ability of the model to fit the data.



Supplementary Figure 2: A comparison of the raw data and model estimates. This figure shows the raw data (blue dots), raw data average +/- one standard deviation (black interval) and median model estimate with 95% central credible interval of the raw data average (red interval) for the probability that each time a participant struck the flint core with their hammerstone a viable flake would be produced. In this case the model predictions are consistently below the raw data average, although well within the standard deviation interval. This is because the data has a high positive skew (there are several raw data points well above the upper limit of the figure) and so the raw data average has been increased. That the model estimate is lower shows that the model is better able to deal with skewed data than the raw data average. Indeed, observation of the blue raw data points indicates that the model estimate sits much closer to the densest area of the raw data points than the raw data average does. Furthermore the size of the model estimate interval is much less than the standard deviation interval indicating the greater precision afforded by the model. Again, this plot can give us great confidence that the model was able to fit the data well.



Supplementary Figure 3: The transmission of concepts along chains in the verbal teaching condition. This figure shows the proportion of teaching utterances than covered particular topics contingent on position along the chain in the verbal teaching condition. It illustrates how some concepts were more successfully transmitted along chains than others. Knowledge of the platform edge and force required were transmitted effectively, with no evidence of a decrease, whilst the extent to which teachers talked about the platform angle decreased and utterances concerning a ridge to carry force had virtually disappeared by position 5. The values shown are median model estimates.



Supplementary Figure 4: A labelled diagram of the stone knapping process. The angle subtended by the rock between the point of impact and the nearest edge is the Exterior Platform Angle (EPA) and the nearby edge is referred to as the platform edge.

Supplementary Tables

Supplementary Table 1: Estimated values for parameters at the first position in the chain for different conditions.

Variable		Condition				
		RE	IE	BT	GT	VT
Number of flakes	All	28.0, [21.9,36.0]	31.7, [24.9,40.5]	27.9, [21.8,35.3]	30.1, [23.5,38.4]	34.3, [26.9,43.8]
	Viable	15.8, [12.1, 0.5]	18.3, [14.1, 3.6]	19.6, [15.1, 5.4]	21.7, [16.8, 8.3]	25.2, [19.4,33.0]
	Non-viable	12.0, [9.1, 15.9]	13.1, [10.0,17.1]	8.1, [6.1, 10.9]	8.6, [6.5, 11.3]	9.6, [7.2, 12.7]
	Selected	12.5, [9.4, 16.4]	13.3, [10.1,17.4]	16.3, [12.5,21.1]	14.8, [11.3,19.4]	23.0, [17.5,30.4]
	Non-selected	14.7, [11.3,19.3]	17.6, [13.6,23.0]	11.3, [8.6, 14.7]	14.6, [11.3,19.0]	13.1, [10.1,17.1]
Proportion of flakes	Viable	0.55, [0.48,0.62]	0.58, [0.52,0.64]	0.72, [0.66,0.77]	0.72, [0.67,0.77]	0.73, [0.68,0.78]
	selected	0.46, [0.39,0.53]	0.45, [0.38,0.51]	0.62, [0.55,0.68]	0.48, [0.42,0.55]	0.61, [0.54,0.67]
Total cutting edge (cm)		52.6, [37.3,72.3.]	61.3, [43.5,84.0]	62.3, [46.2,83.2]	81.2, [59.7,109.5]	98.1, [72.0,133.3]

Total flake mass (g)		40.6, [28.2,55.8]	45.1, [31.1,62.2]	57.1, [41.2,76.3]	59.7, [42.8,80.9]	59.3, [42.3,79.9]
Total quality		13.0, [9.2, 17.9]	15.7, [11.1,21.4]	15.4, [11.1,20.7]	19.8, [14.6, 26.7]	23.6, [17.0, 31.9]
Proportion of core remaining		0.56, [0.46,0.65]	0.54, [0.44,0.63]	0.47, [0.37,0.57]	0.49, [0.38,0.57]	0.41, [0.29,0.52]
Hits per minute knapping		43.2, [32.7,57.5]	39.7, [30.1,52.5]	34.5, [26.1,45.2]	34.3, [26.0,45.5]	28.8, [20.9,39.3]
Flakes per minute	All	3.28, [2.31,4.62]	3.13, [2.21,4.36]	3.56, [2.56,5.00]	4.04, [2.87,5.77]	4.52, [3.15,6.69]
	viable	1.96, [1.33,2.87]	1.98, [1.35,2.85]	2.55, [1.78,3.69]	2.95, [2.03,4.36]	3.37, [2.26,5.19]
Probability of a viable flake per hit		0.03, [0.02,0.05]	0.04, [0.03,0.06]	0.06, [0.04,0.08]	0.07, [0.05,0.10]	0.10, [0.07,0.16]

Quoted values are medians and 95% central credible intervals.

Supplementary Table 2: Estimated values for effects of position along the chain on different variables and for different conditions.

Variable		Condition				
		Reverse Engineering	Imitation/ Emulation	Basic Teaching	Gestural Teaching	Verbal Teaching
Number of flakes	All	0.05, [0.03, 0.08]	-0.02, [-0.05, 0.01]	0.03, [0.00, 0.05]	0.03, [0.00, 0.06]	-0.04, [-0.06,-0.01]
	Viable	0.07, [0.03, 0.10]	0.00, [-0.03, 0.04]	0.00, [-0.03,0.03]	-0.01, [-0.04, 0.02]	-0.07, [-0.10, -0.04]
	Non-viable	0.04, [-0.00, 0.08]	-0.05, [-0.09,-0.01]	0.07, [0.03,0.11]	0.09, [0.05, 0.14]	0.02, [-0.02,0.06]
	Selected	0.05, [0.01,0.09]	0.02, [-0.02,0.06]	-0.03, [-0.06,0.01]	-0.01, [-0.05,0.03]	-0.11, [-0.14,-0.07]
	Non-selected	0.06, [0.02,0.10]	-0.05, [-0.08,-0.01]	0.08, [0.04,0.11]	0.07, [0.03,0.11]	0.02, [-0.01,0.05]
Proportion of flakes	Viable	0.03, [-0.01,0.08]	0.03, [-0.01,0.08]	-0.06, [-0.10,-0.01]	-0.11, [-0.45,-.06]	-0.08, [-0.13,-0.03]
	Selected	0.02, [-0.03,0.06]	0.02, [-0.02,0.07]	-0.12, [-0.16,-0.07]	-0.03, [-0.08,0.02]	-0.15, [-0.19,-0.10]
Total cutting edge (cm)		0.06, [-0.01,0.14]	-0.02, [-0.11,0.07]	0.02, [-0.05,0.08]	-0.04, [-0.12,0.04]	-0.06, [-0.13,0.01]
Total flake mass		0.01,	0.01,	-0.01,	0.00,	-0.01,

(g)		[-0.08,0.08]	[-0.08,0.09]	[-0.08,0.05]	[-0.08,0.08]	[-0.08,0.06]
Total quality		0.06, [-0.02,0.14]	-0.02, [-0.12,0.06]	0.01, [-0.05,0.07]	-0.04, [-0.12,0.04]	-0.07, [-0.14,0.01]
Proportion of core remaining		-0.02, [-0.13,0.08]	-0.06, [-0.16,0.04]	0.00, [-0.09,0.09]	*-0.09, [-0.20,0.01]*	-0.04, [-0.16,0.08]
Hits per minute knapping		0.06, [-0.01,0.13]	*0.06, [-0.01,0.13]*	0.01, [-0.05,0.07]	0.05, [-0.02,0.12]	0.15, [0.06,0.24]
Flakes per minute	All	0.02, [-0.07,0.11]	0.03, [-0.05,0.12]	-0.00, [-0.08,0.08]	0.00, [-0.09,0.09]	-0.09, [-0.21,0.02]
	viable	0.02, [-0.08,0.12]	0.02, [-0.07,0.12]	-0.02, [-0.11,0.07]	-0.03, [-0.13,0.07]	*-0.12, [-0.25,0.00]*
Probability of a viable flake per hit		0.01, [-0.02,0.05]	-0.08, [-0.12,-0.05]	*-0.04, [-0.08,0.00]*	-0.12, [-0.16,-.08]	-0.33, [-0.38,-0.28]

Quoted values are medians and 95% central credible intervals. If the 95% central credible interval excludes 0 this is considered strong evidence for an effect. Values in italics correspond to cases where the 95% central credible interval includes 0, but the 90% central credible interval excludes 0, thus it can be considered weak or moderate evidence for an effect.

Supplementary Table 3: Estimated values for effects of core mass on different variables.

Variable		Effect of core mass
Number of flakes	All	0.13, [0.09, 0.17]
	Viable	0.13, [0.08, 0.17]
	Non-viable	0.11, [0.04, 0.17]
	Selected	-0.03, [-0.08, 0.02]
	Non-selected	0.26, [0.21, 0.31]
Total cutting edge (cm)		0.04, [-0.06, 0.15]
Total flake mass (g)		*0.09, [-0.00, 0.18]*
Total quality		0.05, [-0.05, 0.16]
Proportion of core remaining		-1.82, [-3.42, -0.60]

Quoted values are medians and 95% central credible intervals. If the 95% central credible interval excludes 0 this is considered strong evidence for an effect. Values in italics correspond to cases where the 95% central credible interval includes 0, but the 90% central credible interval excludes 0, thus it can be considered weak or moderate evidence for an effect.

Supplementary Table 4: Estimated values for rate and extent of change for variables along chains, and, where appropriate, accuracy of topics.

Variable/Category/Topic	Rate of change along chains	Extent of change along chains	Accuracy
Total Utterances	1.2, [0.63, 14.0]	-42.2, [-29.3, -58.9]	-
Proportion of teaching utterances correct	1.4, [0.56, 45.8]	-4.0, [-1.4, -6.9]	-
Said by the teacher	0.00, [0.0, 0.00]	-0.76, [-3.57, 5.19]	-
Teaching	0.00, [0.0, 0.01]	-0.28, [-5.76, 3.87]	-
Feedback	0.00, [0.0, 0.06]	-0.28, [-3.90, 3.25]	-
Confirmation of understanding	13.3, [1.89, 163.5]	-0.88, [-1.77, -0.09]	-
Watch this	0.00, [0.0, 0.30]	2.35, [-2.99, 6.47]	-
This/that	0.40, [0.00, 91.57]	-0.56, [-3.35, 3.56]	-
Requesting Information	10.9, [0.86, 149.5]	0.96, [-0.04, 2.23]	-
Conveying uncertainty	7.18, [1.63, 159.0]	3.88, [1.95, 6.69]	-
Abstract	0.00, [0.0, 0.00]	-0.52, [-4.40, 3.15]	-
Correct	4.03, [1.38, 6.90]	-4.03, [-6.90, -1.38]	-
Incorrect	2.36, [0.83, 98.85]	4.00, [-1.33, 7.39]	-
Knapping	0.11, [0.00, 111.0]	-0.74, [-4.07, 2.08]	-
Knapping site	0.09, [0.02, 7.82]	-2.31, [-5.65, -0.54]	0.55, [0.34, 0.76]
Platform edge	0.00, [0.0, 0.09]	1.18, [-4.13, 4.78]	0.93, [0.79, 0.98]

Platform angle	3.99, [0.0, 128.1]	-0.75, [-1.91, 3.21]	0.72, [0.36, 0.93]
Ridge	0.42, [0.18, 1.10]	-3.69, [-6.75, -1.95]	1.0, [0.96, 1.0]
force	0.00, [0.0, 0.03]	0.53, [-3.49, 4.37]	0.38, [0.20, 0.60]
How to hit	0.00, [0.0, 0.0]	1.01, [-4.01, 5.52]	0.80, [0.57, 0.93]
Hot to hold	0.00, [0.0, 0.00]	0.68, [-3.93, 4.68]	0.83, [0.52, 0.97]
Hammerstones	0.00, [0.0, 1.51]	1.72, -1.81, 6.25]	0.73, [0.47, 0.90]
Cortex	0.00, [0.0, 0.65]	1.79, [-2.16, 6.72]	0.94, [0.77, 0.99]
Choosing flakes	9.97, [0.00, 161.8]	0.82, [-1.73, 3.73]	-
Size of flakes	0.00, [0.0, 0.00]	2.01, [-1.94, 6.15]	0.68, [0.39, 0.89]
Cutting edge of flakes	0.00, [0.0, 0.09]	1.09, [-2.64, 6.09]	0.91, [0.80, 0.97]

Quoted values are medians and 95% central credible intervals. A negative value for the extent of change corresponds to a decrease along the chain. To aid interpretation of the rate parameter; a value greater than 2 is very rapid change such that ~90% of any change is achieved in the first step. A value below 0.5 corresponds to a more gentle change with ~90% of the change occurring over the first 5 steps, and lower values correspond to even gentler change. Values between these correspond to intermediate rates of change.

Supplementary Table 5: Contrasts between conditions for different variables.

Variable	First condition		Second condition		Contrast
Number of viable flakes	VT		RE		9.4, [2.1, 18.1]
			IE		*6.9, [-0.8, 18.1]*
	GT		RE		*6.0, [-0.7, 13.5]*
Proportion of flakes that are viable	VT		RE		0.18, [0.12, 0.25]
			IE		0.15, [0.09, 0.21]
	GT		RE		0.17, [0.11, 0.24]
			IE		0.14, [0.08, 0.20]
	BT		RE		0.17, [0.11, 0.23]
			IE		0.14, [0.08, 0.20]
	BT	IE	VT	GT	0.57, [0.20, 0.95]
			GT	BT	0.60, [0.13, 1.08]
			IE	RE	0.49, [0.05, 0.94]
Number of non-viable flakes	GT		RE		*-3.4, [-7.7, 0.5]*
			IE		-4.5, [-9.1, -1.0]
	BT		RE		*-3.8, [-8.3, 0.1]*
			IE		-4.9, [-9.6, -0.8]
Number of selected flakes	VT		GT		8.1, [1.2, 16.3]
			BT		*6.7, [-0.5, 14.8]*
			IE		9.6, [2.7, 17.6]
			RE		10.5, [3.6, 18.5]

Proportion of flakes that were selected	VT	GT	0.12, [0.06, 0.18]
		IE	0.16, [0.10, 0.22]
		RE	0.15, [0.08, 0.22]
	BT	GT	0.13, [0.07, 0.20]
		IE	0.17, [0.11, 0.23]
		RE	0.16, [0.09, 0.22]
Number of non-selected flakes	BT	IE	-6.3, [-12.6, -1.0]
Total quality	VT	BT	*8.2, [-0.1, 17.4]*
		IE	*7.9, [-1.1, 17.5]*
		RE	10.6, [2.2, 20.0]
	GT	RE	*6.7, [-0.4, 14.7]*
Total cutting edge	VT	BT	36.0, [2.7, 72.9]
		IE	36.6, [2.9, 76.4]
		RE	45.7, [12.0, 85.4]
	GT	RE	*28.4, [-0.3, 61.3]*
Total mass	RE	VT	*-18.6, [-41.6, 2.0]*
		GT	*-18.9, [-40.8, 0.29]*
		BT	*-16.2, [-36.1, 1.9]*
Proportion of core remaining	VT	RE	-0.15, [-0.31, -0.00]
		IE	*-0.13, [-0.29, 0.01]*
Hits per minute knapping	VT	RE	-14.3, [-30.8, -0.11]

Viable flakes per minute knapping	VT	RE	1.39, [0.03, 3.35]
		IE	1.37, [0.03, 3.34]
Probability of a viable flake with each hit	VT	BT	0.05, [0.00, 0.10]
		IE	0.06, [0.02, 0.12]
		RE	0.07, [0.03, 0.12]
	GT	IE	*0.02, [-0.00, 0.06]*
		RE	0.03, [0.01, 0.07]
	BT	RE	0.03, [0.00, 0.05]
Topic Accuracy	Ridge	Knapping site	0.44, [0.22, 0.66]
		Platform edge	0.07, [0.01, 0.20]
		Platform angle	0.28, [0.06, 0.63]
		How to hit	0.20, [0.06, 0.42]
		How to hold	0.16, [0.03, 0.47]
		Hammerstones	0.27, [0.09, 0.52]
		Cortex	*0.06, [-0.00, 0.23]*
		Flake size	0.31, [0.11, 0.60]
		Cutting edge	0.08, [0.02, 0.19]
		Force	0.61, [0.39, 0.79]
	Cortex	Knapping site	0.37, [0.09, 0.62]
		Force	0.54, [0.28, 0.74]
	Platform edge	Knapping site	0.37, [0.07, 0.62]
		Flake size	0.24, [0.00, 0.53]
		Force	0.53, [0.30, 0.73]

		Hammerstones	*0.19, [-0.01, 0.40]*
	Cutting edge	Knapping site	0.35, [0.12, 0.58]
		Hammerstones	0.18, [-0.01, 0.44]
		Flake size	0.23, [-0.02, 0.54]
		Force	0.52, [0.29, 0.72]
	Force	How to hit	-0.41, [-0.66, -0.08]
		How to hold	-0.43, [-0.68, -0.08]
		Hammerstones	-0.33, [-0.60, -0.02]
		Flake size	*-0.29, [-0.56, 0.04]*

Quoted values are medians and 95% central credible intervals. Numbers given in italics correspond to cases where the 95% central credible interval included 0, but the 90% central credible interval did not. i.e., cases where strong evidence was not reached, but there is still some evidence for such a difference. Key: RE = reverse engineering, IE = imitation/emulation, BT = basic teaching, GT = gestural teaching, VT = verbal teaching.

Supplementary Figure 6: Differences in performance between gestural and verbal teaching.

Variable		Model Estimate
Probability that average performance with verbal teaching > with gestural teaching		0.9, [0.57, 1.00]
Probability of strong evidence that performance > than with reverse engineering, imitation/emulation or basic teaching	verbal teaching	0.6, [0.38, 0.8]
	gestural teaching	0.19, [0.06, 0.41]
	difference between verbal and gestural teaching	0.41, [0.12, 0.65]

Quoted values are medians and 95% central credible intervals. I no case do we find strong evidence that performance according to a particular measure was greater with verbal teaching than with gestural teaching. Nonetheless, there is strong evidence that across multiple measure, performance was better with verbal teaching than with gestural teaching.

Supplementary Methods

General Methods

Across two weeks 184 participants learnt and taught others to make flint flakes using a granite hammerstone and flint core. We used a transmission chain design in which the first participant in a chain was taught by a skilled experimenter and subsequent participants were taught by the previous participant. Participants gained asocial information through access to the materials themselves. The social information was from a demonstrator or teacher and varied across five learning conditions detailed below. For each of the learning conditions we ran four short chains (≤5 participants long) and two long chains (≤10 participants long), totalling 30 chains across all conditions. Each participant was involved for ~90 minutes and was paid between £10 and £20 depending on their performance.

Apparatus & Set-up

We used 2 tonnes of Brandon flint from a chalk quarry (Norfolk, UK), broken up into cores of roughly 1kg in weight. We collected around 100 granite hammerstones, of a range of shapes and sizes from the coastline near Stonehaven, Scotland.

The knapping room contained a 4x4m square knapping area, the floor of which was covered in cardboard or black plastic sheeting, divided into two 2x4m sections by a 1m tall clear perspex screen. In each section was a chair on which participants could sit and a large piece of Hessian that participants could use to protect their clothing whilst knapping. When only one participant was present they were free to use either section, but when a teacher and learner were both present they each used one section. Participants were free to enter each other's sections during the pupil/tutor phases, but were only allowed to knap in their own section. The screen ensured that flakes from each participant did not enter the other participant's section. Thus, it was clear who had produced any flakes found in each section. The screen also prevented flakes produced hitting another participant. Immediately to the side of the knapping area was a large pile of hammerstones from which participants were free to choose. For safety, all participants were required to wear a pair of safety glasses and latex coated cotton gloves. We additionally provided breathing masks for participants in case they found the dust produced to be irritating. Two experimenters were present, at all times, sitting at a desk outside of the knapping area. A small number of flint cores were stored behind the desk and the experimenters chose cores from this supply at random for each participant.

Procedure

Upon arrival, participants were briefed on the experimental procedure and given the opportunity to ask any questions. Participants then began the **introductory phase** of the experiment. Participants were provided with some pre-knapped flint flakes, some chamois leather and some sticks. They were given an information sheet containing superficial information on the emergence of such technology in the archaeological record, the tasks that flakes were used for, and that flakes were produced by striking pieces off a larger stone. They were then given 5 minutes to use the flakes to cut the leather and to sharpen the sticks. They were encouraged to try a range of flakes to achieve an understanding of what properties made a useful (henceforth "viable") flake. The introductory phase took part in a different room to the other phases of the experiment.

After this, the **pupil phase** began. Participants were given five minutes to practice making their own flint flakes. Additionally participants were provided with social information, the form of which varied depending on the learning condition, as detailed further below.

Next, participants entered the **test phase**. They were instructed to make as many high quality flakes from the core as they could. They were not told of a time-limit, although the experimenter called it to an end if the participant took over 20 minutes.

If applicable, the participants next continued to the **tutor phase** where they provided social information to the next participant in the chain, just as they had experienced in their pupil phase. After this, participants were debriefed and were paid according to their performance.

In all phases of the experiment that involved knapping, participants were provided with a flint core and could choose a hammerstone. At the end of the phase we asked participants to separate out their flint into three categories; what remained of the core, viable flakes, and non-viable flakes. Flakes the participant selected as viable will henceforth be referred to as “selected”, whilst those they did not selected as viable will be referred to as “non-selected”.

Conditions

The experiment involved 5 different learning conditions that dictated the form of the social information by placing limits on the ways in which learner and teacher could interact. The conditions were as follows:

1. **Reverse Engineering** - The learner had access only to the flakes produced by their teacher and no access to the teacher themselves. In this condition there was no teaching as the tutor was not present. Thus once participants had completed the test phase they proceeded immediately to debriefing. The flakes available to the pupil were those produced by the previous participant in the previous participant's test phase that the previous participant had categorized as viable.
2. **Imitation/Emulation** - The pupil was able to watch a tutor making flakes, but no forms of direct interaction were permitted. As the tutor produced flakes they categorized them as viable or non-viable and the flakes were available for the pupil to examine.
3. **Basic Teaching –** Communication between the pupil and tutor was permitted but was limited to some simple forms of non-symbolic teaching. The permitted interactions were manual shaping (where the tutor could adjust how the pupil was holding the core and hammerstone), slowing of actions, and reorientation to allow the pupil a clear view. These forms of teaching were chosen as they are the forms of teaching for which there is some evidence in non-human animals.
4. **Gestural Teaching** - Communication between the tutor and pupil was permitted but was limited to gestural (i.e., non-verbal) communication. This included, but was not limited to, mutual touching of tools, pointing, miming and nodding.
5. **Verbal teaching –** All forms of communication between the tutor and pupil were permitted, including use of language.

In all teaching conditions the tutor was provided with their own flint core and hammerstone and could make their own flakes. Once flakes had been made the pupil was allowed to examine them.

Payment

Participants were informed in advance of the payment scheme for the experiment, which varied by condition. In all conditions, we paid participants according to the number of viable flakes they were able to produce, divided by the initial mass of their core, during their test phase. We included any flakes that we considered viable, regardless of whether the participant had categorized them as such, as otherwise participants would have been motivated to categorise everything they produced as viable to increase their payment. We chose this payment scheme as it reflects pressures on early hominin tool makers to produce as many flakes as possible from a limited supply of knapping material.

In teaching conditions, tutors were also evaluated on their pupil's subsequent test phase performance; this was to ensure tutors were motivated to teach effectively. With imitation/emulation, participants were evaluated on their own test and tutor phase performance; this was to motivate them to focus on their own performance during the tutor phase, instead of teaching the pupil.

Recorded Variables

We used digital video cameras to record the entirety of the experiment (although video recording failed for one of the long chains in the VT condition). Additionally, we recorded the initial weight of all the flint cores given to participants. Finally, at the end of each phase and for each participant we separately bagged (i) what remained of the core, (ii) any selected flakes and (iii) any non-selected flakes.

Coding

Flakes

All flakes greater than 2cm in diameter were coded, totalling 6214 flakes. This lower limit of 2cm was considered to be the minimum for a useful butchery tool[2]. Any flakes that had an edge deemed sharp enough to be of use were coded as viable, otherwise they were coded as non-viable. Prior to the full coding, a subset of 317 flakes were triple coded by TM, NU and IT. All of this subset were coded first as viable or non-viable, and if viable they were then rated on a 10-point scale of quality that took into account the efficiency with which the raw material had been used. A latent variable analysis of flake viability was carried out to estimate the accuracy of the viability coding decisions of each of the coders. The viability of each flake was modelled as a latent variable with a Bernoulli error structure. Additionally the viability ratings of each coder were modelled with a Bernoulli error structure and a logit link function. The linear predictors for coders' ratings took separate values for each coder and for each value of the latent variable (viable or non-viable). The only constraint placed upon the model was that all coders performed above chance, such that they had a >50% chance of identifying a flake correctly. The model then used the coders' decisions to estimate the viability of each flake and in turn the accuracy of each coder. All three coders were estimated to have similarly high levels of accuracy (estimated probabilities of accurate identification; TM = 0.81 [0.75, 0.87], NU = 0.89 [0.83, 0.94], IT =

0.82, [0.74, 0.88]). The imperfect viability coding likely reflects the inherent difficulty in the coding decisions, as many flint fragments were of debatable value. The remaining flakes were coded by TM. In addition to viability we also recorded flake cutting edge length, flake diameter and flake mass.

Flake quality

Based upon the 10-point quality ratings by the triple coders, a metric for flake quality was developed such that all flakes could be assigned a numerical quality rating that could be subject to analysis. Following Braun & Harris[1], the metric began with:

$$\textit{quality} = \textit{flake cutting edge}/\textit{flake mass}^{(1/3)} \qquad (1)$$

This scores flakes according to how much cutting edge they had, but the cube root function prevents larger flakes from being penalised by their large size (when scaled up by length, a flakes mass will increase by the scaling factor cubed). However, this formula does not take into account size, which is clearly of relevance to flake quality, as excessively small flakes will be unusable and excessively large flakes will be wasteful of raw material. To include flake diameter the metric was extended to

$$\textit{quality} = (\textit{flake cutting edge}/\textit{flake mass}^{(1/3)})*f(\textit{flake diameter}), \qquad (2)$$

where *f(flake diameter)* was an unknown function, with the constraint that $f(x) \geq 0$. To estimate the shape of *f(x)* the quality ratings of the three triple coders were modelled with a binomial error structure (where n was 10 as the ratings were on a 10 point scale). The probability of a success was transformed into the positive continuous variable “quality”, which was modelled with the above formula. The unknown diameter function was modelled as categorical such that it could take independent values for diameters at intervals of one centimetre. Visual inspection of the estimated values of this function at each centimetre interval strongly suggested a cumulative exponential function was appropriate and so the model was re-run with the function of flake diameter as a cumulative exponential distribution such that

$$quality = (flake\ cutting\ edge/flake\ mass\wedge^{(1/3)})*(1-exp(-lambda*(flake\ diameter - offset))), \quad (3)$$

where *lambda* is a positive continuous variable that sets the gradient of the cumulative exponential function and *offset* is the minimum possible diameter of a flake to have any quality whatsoever. Offset was given a uniform prior ranging between 0 and 2 as flakes cannot be less than 0cm across and it was already decided that flakes over 2 could have some quality. The model estimates of these two parameters were: *lambda* = 0.31 [0.28, 0.35]; *offset* = 1.81, [1.69, 1.90]. The posterior distribution for *offset* sat comfortably within the interval specified by the prior, suggesting that it was an appropriate prior distribution. Given this, the final flake quality metric is:

$$quality = (flake\ cutting\ edge/flake\ mass^{(1/3)})*(1-exp(-0.31*(flake\ diameter - 1.81))) \quad (4)$$

This function rewarded flakes for a high cutting edge length and penalised flakes for being excessively small. Around a size of 2cm flakes were very heavily penalised; however, the effect of flake diameter flattens above 6cm such that further increases in size do not greatly increase quality. It is of note that the diameter function does not penalise flakes for being excessively large. This is presumably because most flakes produced by participants were small and so very few flakes were large enough to receive any penalisation.

Videos

The participants' behaviour, as video recorded at all points in the learning, testing and teaching phases, was coded into one of the following categories:

1. Knapping - when the participant directs their attention toward their own core and hammerstone with the aim of making flakes for their own ends e.g., knapping, looking, turning in hands.
2. Observing - when the participant directs their attention to their tutor or their tutor's flakes
3. Teaching - when the participant directs their attention to their pupil or knaps for the benefit of their pupil
4. Choosing - when the participant directs their attention to flakes they have produced as if considering the quality or nature of them. If the participant proceeds to try to knap the flake this no longer counts as choosing and instead counts as knapping.
5. Other - any behaviours that do not fit into the above categories.

Additionally, the time of every strike of the core with the hammerstone was recorded. As a test of coding accuracy, ten participants were randomly chosen (2 from each condition, 10% of all participants) and their videos were coded by TM and RK. We modelled the absolute magnitude of the disagreement between total time spent knapping and total number of hits for each of the coders as these were the variables used in further analyses. In the case of time spent knapping we used a gamma error structure and the expected difference is 20.4s, [14.0, 31.2]. As a proportion of the average time for which participants were present this is 0.04, [0.03, 0.07] which is a very low proportion of disagreement. In the case of total hits we used a poisson error structure and the expected disagreement is 7.7 hits [6.7, 8.8], as a proportion of the average number of times each participant hit the core with their hammerstone this is 0.04, [0.04, 0.05]. Given this high level of agreement RK went on to code all the remaining videos.

Language

Whilst coding the videos as described above, RK also transcribed everything that was said by participants. This was then coded by TM as follows. Initially, each transcript was split into utterances, defined as a single stretch of verbal communication by a single participant. Thus an utterance ends with a pause or when the other participant says something. Each utterance was scored according to the following categories which are not mutually exclusive in that a single utterance could (in theory) score positively for every category:

1. Said by the tutor – was the utterance said by the teaching participant.
2. Teaching – did the utterance transmit knapping relevant information to the other participant (note, this could be from the learner to the demonstrator) e.g. “You want to rest the flint core on your left leg” which transfers knowledge of how to hold the core.
3. Feedback - was the utterance giving feedback on performance, in terms of encouraging good behaviour or vice-versa. Note, feedback is a type of teaching. e.g. “So that's the sort

of thing you want to, that's brilliant"

4. Confirmation of understanding - was the purpose of the utterance to confirm that the speaker had understood something. Note, most instances of the word "yes" were coded in this category and not as a "yes/no". e.g. "Ok, of course", but not "So you're always trying to hit above a ridge then?" which would be coded as a question
5. Watch this - was the utterance directing attention to the speaker it order to demonstrate something. e.g. "just..." followed by the speaker knapping
6. This/that - did the utterance use words such as this or that to indicate objects or locations. e.g. "That one's no good, is it?"
7. Requesting Information - was the utterance a request for knapping relevant information. e.g. "So you're always trying to hit above a ridge then?" which requests information on where to hit
8. Conveying uncertainty - did the utterance include an expression of uncertainty. e.g. "Maybe that bit's kind of hanging over and there's kind of an under-hang, try that", note use of maybe, kind of and try that.
9. Abstract - did the utterance use abstract descriptions that gave general information not specific to a single case. e.g. "Find an edge, do you have an edge with black stuff on the other side as well?" which describes the general procedure for identifying an edge without cortex, as opposed to "Emm this is probably going to be your hit" where a participant simply points out a specific point with no generalisable information.
10. Correct – was information in the utterance factually correct.
11. Incorrect – was information in the utterance factually incorrect.

In addition to the above categories the topic of the utterances (as opposed to their nature/purpose) was also categorized according to the following topics:

1. knapping (a broad category)
2. knapping site
3. platform edge
4. platform angle
5. ridge
6. force
7. how to hit
8. how to hold
9. hammerstones
10. cortex
11. choosing flakes (a broad category)
12. size of flakes
13. cutting edge of flakes
14. safety whilst knapping

As with the previous categories, the topics are not mutually exclusive. Additionally topics 1 and 10 (knapping and choosing flakes) are very broad with the other topics falling as sub-topics within these. For example, the topic "platform edge" is a sub-topic within "knapping" as by talking about the platform edge you are also talking about knapping.

Analyses

We analysed **the number of total flakes, viable flakes, non-viable flakes, selected flakes and non-selected flakes** that each participant produced with a poisson error structure. We also analysed **the proportion of flakes that are viable** and **the proportion of flakes that are selected** using a binomial error structure. The total number of flakes produced was used as the number of trials and the number of viable or selected flakes was the number of successes. The proportion of flakes that were non-viable and not selected was not analysed as they are the inverse of the proportion of flakes that are viable and selected respectively. Using a gamma error structure we also analysed **the sum of the cutting edge length, the sum of the mass and the sum of the quality of all flakes** produced by participants. All of these models used a logarithmic link function, except for the binomial models that used a logit link function, and the linear predictor contained categorical effects of condition that interacted with a linear effect of position along the chain and a linear effect of core mass. Individual level effects were not included as each individual only contributed a single data point to each analysis.

Using a hurdle model we analysed **the proportion (by mass) of the participant's core remaining** after knapping. First the model analysed whether a participant had any of their core remaining at all with a bernoulli error structure and logit link function, then in the cases where there was some core left it analysed the proportion left with a beta error structure and logit link function. These two elements could then be combined to produce an estimate of the expected core remaining. In both parts of the model the linear predictor contained categorical effects of condition that interacted with a linear effect of position along the chain. Individual level effects were not included as each individual contributed only a single data point to each analysis.

We modelled **the number of hits per minute spent knapping** and **the number of flakes produced per minute** (both all flakes and viable flakes) with a lognormal model, and **the probability each hit produces a viable flake** with a binomial model and logit link function. In these cases the linear predictor contained categorical effects of condition that interacted with a linear effect of position. There were no effects of core mass as it was deemed implausible that this could have an effect on the variables investigated.

The total number of utterances said was analysed with a poisson error structure. The model incorporated chain length with a function that set a baseline number of utterances, an initial deviation to this number that set the initial value and then a rate parameter that set the rate at which the value approached the baseline from the initial value. The shape of the function was that of a cumulative exponential function. The model included a random effect of repeat for the initial value and did not need to include condition as only VT allowed language. We also analysed **the probability a given utterance satisfied each of the above categories or covered each of the above topics** with bernoulli error structures and logit link functions. The linear predictor used the same function as the model for the total number of utterances. We also investigated whether different topics were transmitted with greater accuracy by modelling **whether an utterance was scored as correct or incorrect** with a bernoulli error structure and logit link function. The linear predictor contained categorical effects of all the topics (other than knapping and choosing flakes as the sub-topics were included instead).

As a test of robustness, the analyses of the numbers of flakes produced (all/viable/nonviable/selected/nonselected) and the probability that each hit produces a viable flake, were repeated with a subset of the dataset such that only flakes > 5cm in diameter were included. This did not qualitatively change results and so below we present the results of the analyses where the minimal limit on size was 2cm.

As the relationship between gestural teaching and verbal teaching was of particular interest we carried out two further analyses comparing the two. Firstly we modelled the probability that the median aggregate performance estimates was greater with verbal teaching than with gestural teaching with a Bernoulli error structure (no link function was needed). The data consisted of 6 measures of aggregate performance: the total quality of all flakes, the number of viable flakes, the proportion of flakes that are viable, the number of viable flakes produced per minute spent knapping, the proportion of core reduced and the probability of a viable flake per hit. Secondly we modelled the probability that the main analyses found strong evidence of a difference between verbal teaching or gestural teaching and the three other conditions (reverse engineering, imitation/emulation and basic teaching). The analyses used the six aggregate measures of performance and used a binomial error structure, where strong evidence of a difference counted as a success and the number of trials was 18 (6 measures of performance x 3 comparison conditions = 18 trials).

Supplementary Note 1

A Glossary of Knapping Terms

Successful knapping - the production of sharp flakes by striking a core with a hammerstone - is a somewhat complex procedure. Here we outline some key elements in order to explain some of the terms used throughout the main paper.

Platform edge

To reliably produce flakes the hammerstone should strike the core on a flat surface near an edge. This distance from the point of percussion to the edge is very important and has a large impact on the size of flakes produced. Generally, a distance to the edge of about 1cm is appropriate. See Supplementary Figure 1 for a helpful diagram.

Platform angle

The surface struck with the hammerstone needs to be slightly overhanging. The angle between the struck surface and the surface below (with its vertex at the nearest point where the two surfaces meet) is the exterior platform angle (EPA). For successful knapping this must be below 90 degrees, ideally around 70 degrees. See Supplementary Figure 1 for a helpful diagram.

Ridge

Ideally, the surface below the platform edge should have a ridge in the rock to direct the force. This helps control the size and shape of flakes produced.

Force

There is an appropriate amount of force with which to strike the core with the hammerstone. Too little and a flake will not be produced, but the core may be damaged. Too much and the core could crack into many pieces.

Cortex

Flint grows underground within chalk. When flint nodules are dug-up they have an outer layer of chalky cortex. This is not suitable for knapping and so needs to be removed for successful knapping.

Supplementary References

1. Braun, D. R., & Harris, J. W. K. (2003). Technological developments in the Oldowan of Koobi Fora: Innovative techniques of artifact analysis. *Oldowan: Rather More than Smashing Stones*, 117–144.

2. Key, A. J. M., & Lycett, S. J. (2014). Are bigger flakes always better? An experimental assessment of flake size variation on cutting efficiency and loading. *Journal of Archaeological Science*, *41*, 140–146.