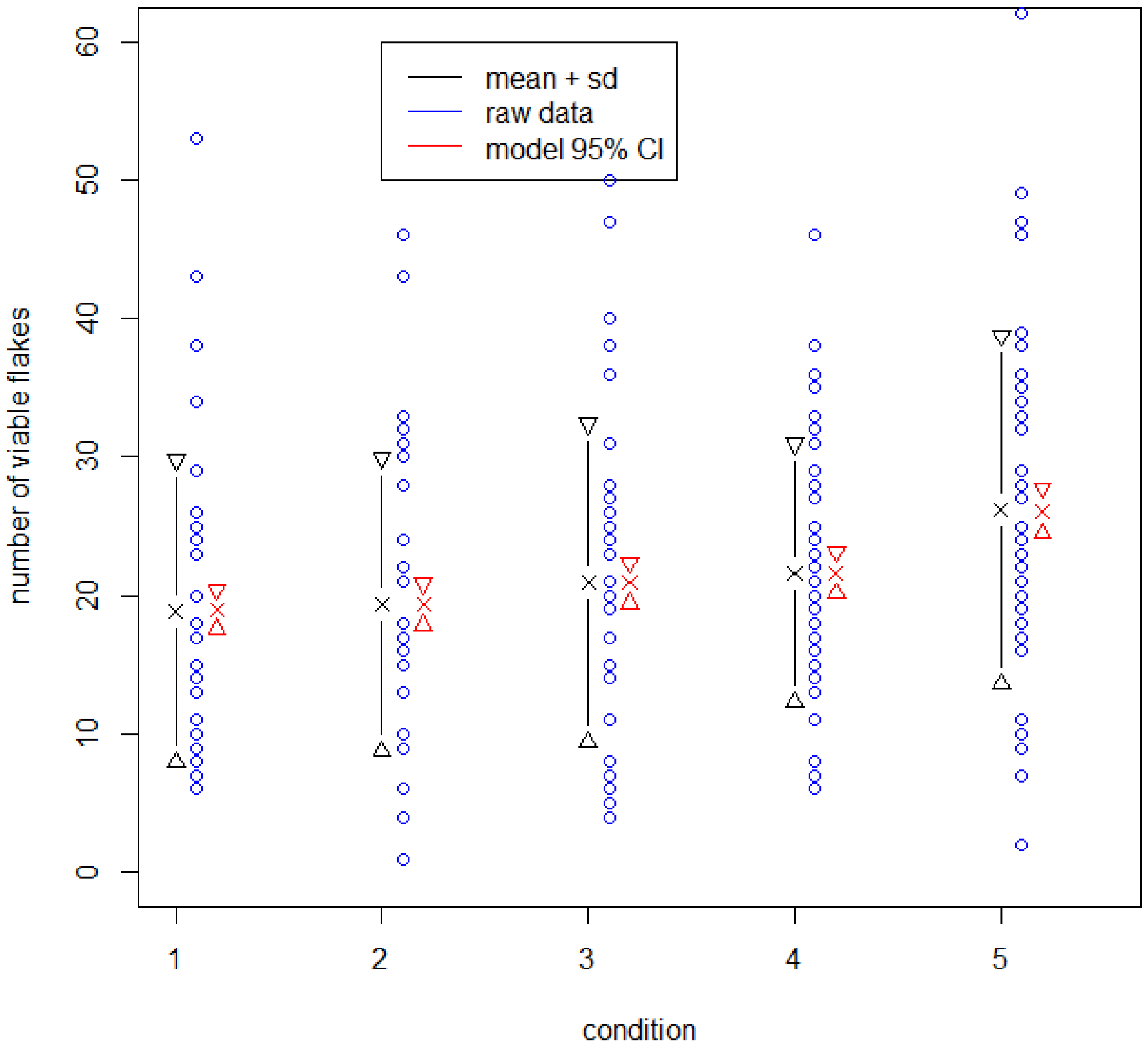
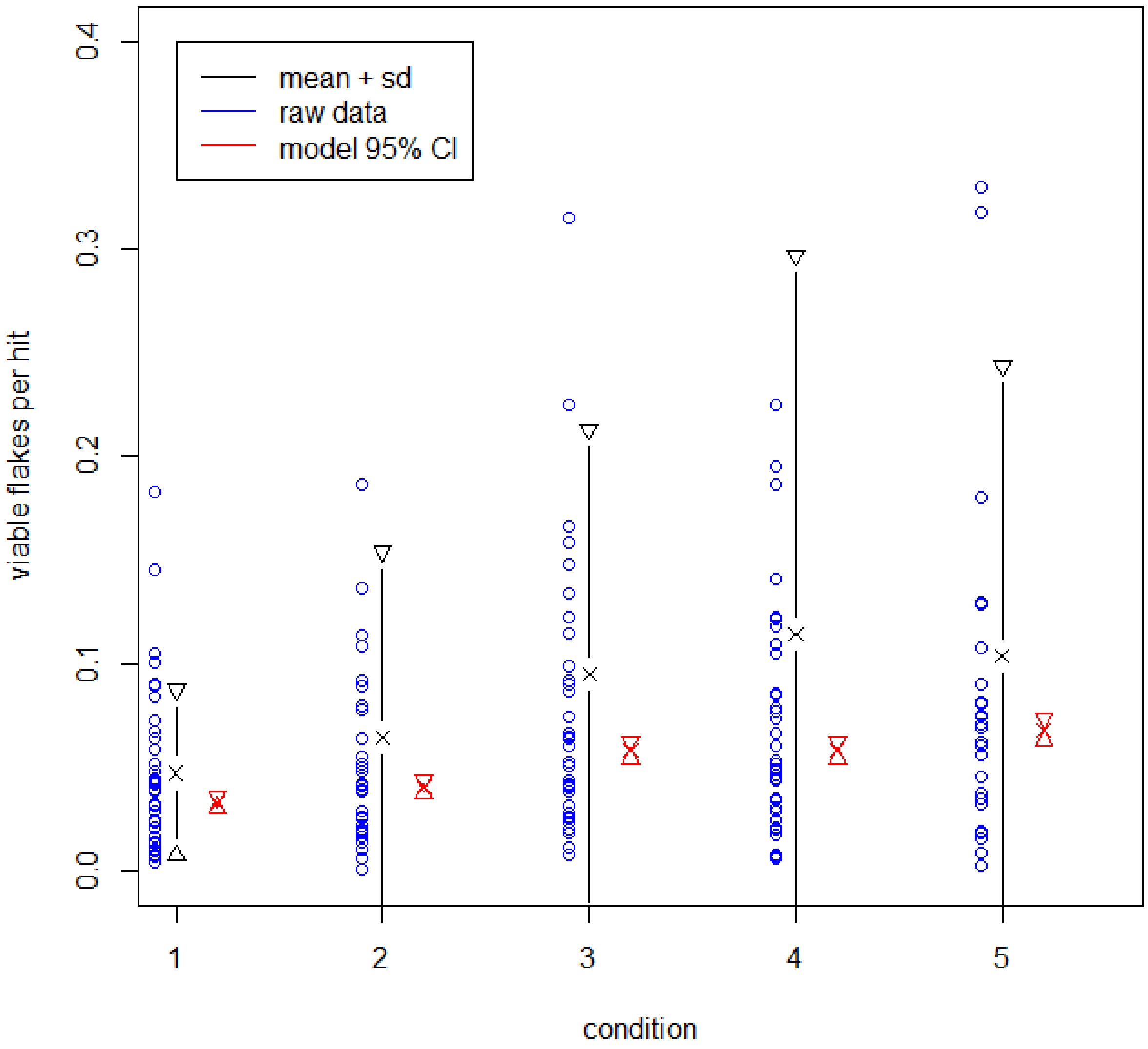
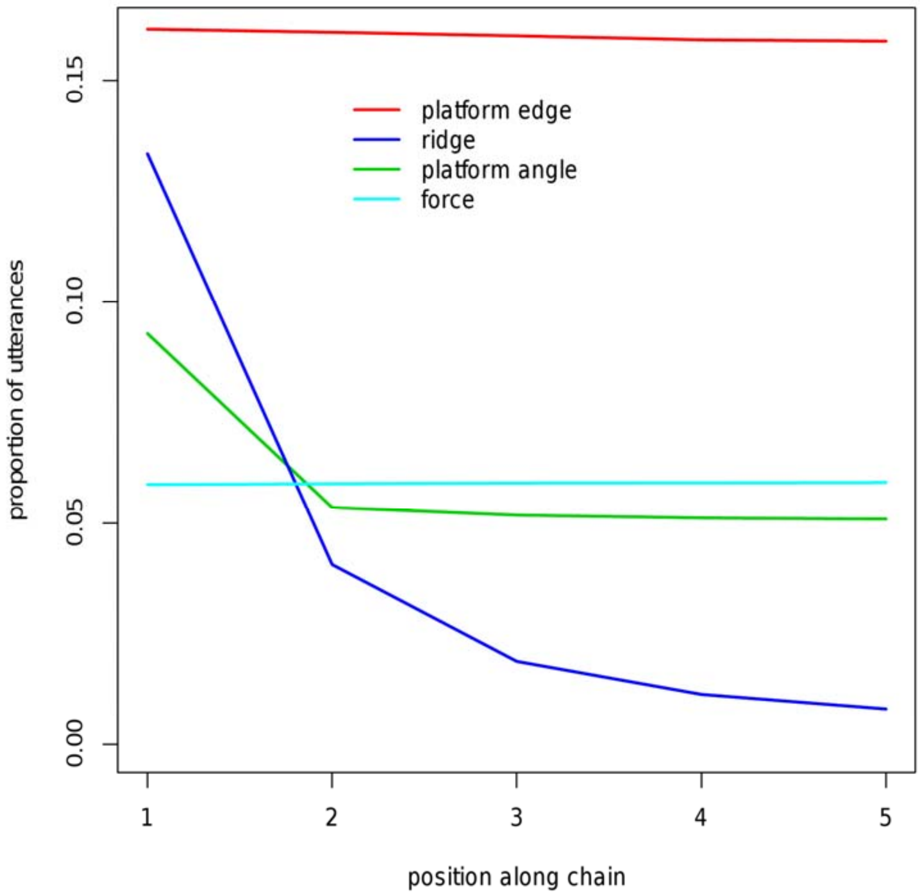


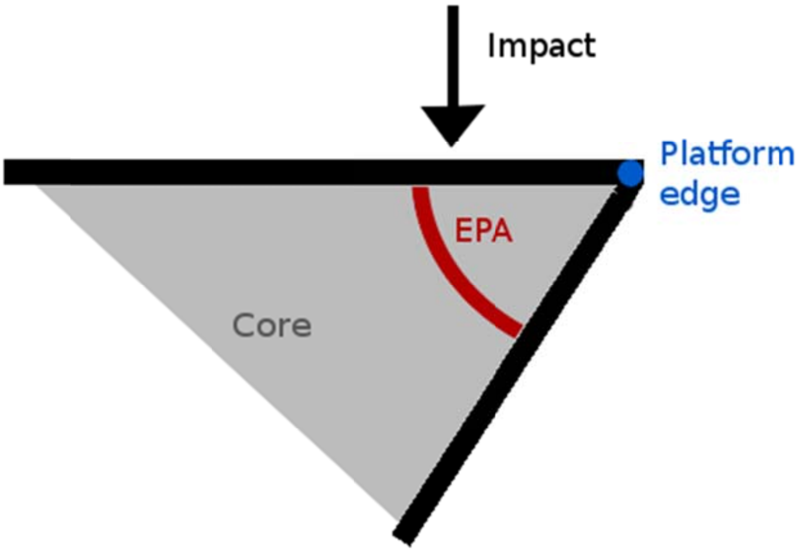
Supplementary Figure 1: A comparison of the raw data and model estimates. This figure shows the raw data (blue dots), raw data average +/- one standard deviation (black interval) and median model estimate with 95% central credible interval of the raw data average (red interval) for the total number of 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 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,	15.7,	15.4,	19.8,	23.6,

		[9.2, 17.9]	[11.1,21.4]	[11.1,20.7]	[14.6, 26.7]	[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 Table 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 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 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:

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

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

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

where *f(flake diameter)* was an unknown function, with the constraint that *f(x)* ≥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

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

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:

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

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 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.