Supplementary material for:

The value of teaching increases with tool complexity in cumulative cultural evolution

Amanda J. Lucas, Michael Kings, Devi Whittle, Emma Davey, Francesca Happé, Christine A. Caldwell & Alex Thornton

Contents:

(1) Pilot study

Methods

To determine whether pipe-cleaner tools are more causally opaque than paper tools, we asked 16 participants (naïve to study hypotheses and predictions) to build (i) a paper tool from a single sheet of waterproof paper, capable of carrying as many marbles as possible while floating on water and (ii) a pipe-cleaner tool from 20 identical 30cm long pipe-cleaners to transport as many marbles as possible. We then tasked 48 naïve participants ("Replicators") with replicating these implements as accurately as possible. Replicators were randomly assigned to one of three conditions: (i) *emulation*, in which they could simply examine the model implement and try to copy it; (ii) *teaching*, in which the maker explained how to build the implement, but the model implement was not present or (iii) *both*, where the model implement was present during the teaching process. Each replicator made one paper tool and one pipe-cleaner tool within the same condition.

To quantify the fidelity of copying, we asked two independent raters (blind to conditions) to gauge the similarity of each implement to its model in terms of three metrics: design,

joins/folds and comparative robustness. Each metric was a 1-5 scale (1 = no similarities at all; no indication of copying; 5 = virtually identical), and we summed the metrics for each implement to give an overall similarity score out of a maximum of 15. Inter-rater reliability was high: $r = 0.70$, $p < 0.001$. As similarity scores are bounded by 15, for analysis we treated similarity as a proportional value, with logit transformation [1]. We used a linear mixed model with proportional similarity as the response variable and tool type (paper or pipe-cleaner), condition (emulation, teaching or both) and the interaction between them as explanatory variables. The identity of the model implement and replicator were fitted as random terms to account for repeated measures.

We also compared the efficacy of each implement to each model. We quantified the efficacy of every implement by asking the maker to test how many pennies a paper tool floating on the water could carry without sinking, or how many marbles a pipe-cleaner tool could carry for 5m across a room. We analysed the data in two separate LMMs with the number of pennies (for paper tools) or marbles (for pipe-cleaner tools) fitted as the response variable, condition as the explanatory variable and the identity of the model implement as a random term.

Results

Initial analysis showed that tool type and treatment interacted to determine the similarity between models and their copies (LRT: $F = 3.90$ $p = 0.026$). We therefore analysed each tool separately. Among paper tools, the degree of similarity was consistently high and did not vary between conditions (Fig S1a; LMM (logit-transformed); *emulation* vs *teaching*: β (s.e.) = -0.64 (0.41), $t = -1.529$, $p = 0.133$, CI (-1.44, 0.17); *emulation* vs *both*: β (s.e.) = -0.30 (0.42), $t = -0.718$, $p = 0.477$; CI (-1.10, 0.51). Among pipe-cleaner tools, however, emulated implements were considerably less similar to their models than those from the *teaching* (Fig. S1b; LMM (logit transformed), relative to *emulation*: β (s.e.) = 0.94 (0.41), $t = 2.31$, $p = 0.028$, CI (0.14, 1.74)) or *both* treatments (β (s.e.) = 0.83 (0.41), $t = 2.04$, $p = 0.05$, CI (0.03, 1.63)).

Analyses of implement performance showed similar results. There were no differences between the performance of model paper tools and their copies from any of the treatments (LMM: $p > 0.250$ in every case). In contrast, emulated pipe-cleaner tools performed worse than their models (β (s.e.) = -1.23(0.56), $t = -2.184$, $p = 0.034$, CI(-2.31,-0.14)), but there was no difference in the performance of models relative to their copies from teaching (β (s.e.) =

0.50(0.56), t = 0.896, p = 0.375, CI(-0.58,1.59)) or both treatments (β (s.e.) = -0.36(0.56), t = -0.643, p = 0.523, CI(-1.44,0.72)).

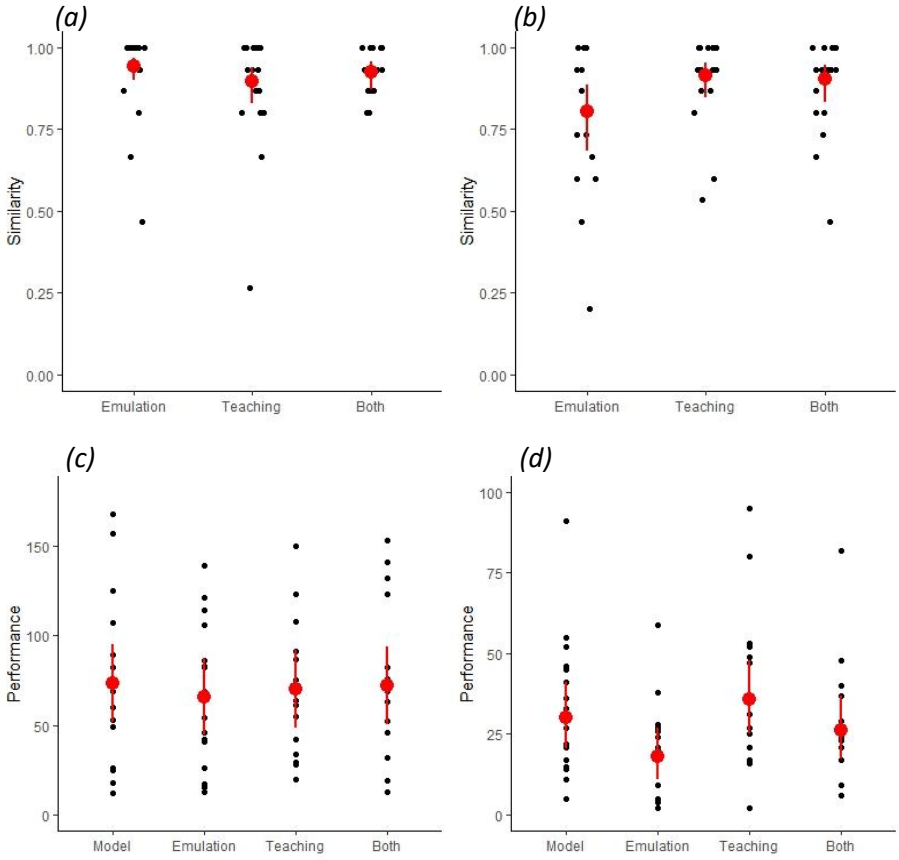


Figure S1. Copies of paper tools did not differ in (a) similarity or (c) performance from their models, but emulated copies of pipe-cleaner tools were (b) less similar and (d) performed worse than their models. Black dots are raw data; means and CIs are in red.

(2) Main experiment: supplementary methods:

(a) Transmission chain design

Our transmission chains were designed so as to ensure that building time and access to social information was standardised across conditions, while ensuring that participants had sufficient time to test their tools. We did not fix the time available for testing as pilot trials revealed that this was the most effective strategy for allowing tools to achieve their full scoring potential. (If we had set a long testing time for all tools this would have caused

unnecessary extension to the duration of experimental sessions, negatively impacting on the recruitment of participants.) Details of the procedure are given in Figure S2 below.

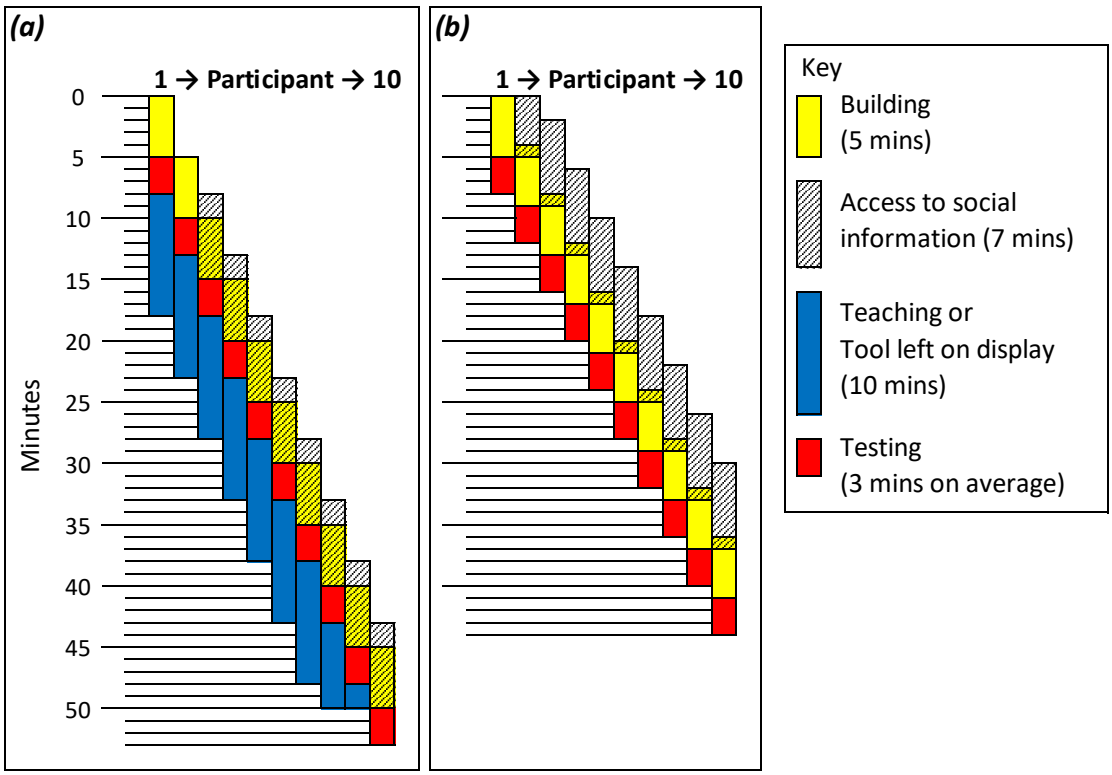


Figure S2. Design of transmission chains. Participants (1-10 within each chain) had five minutes to build their tools (yellow) and, from participant 3 onwards, access to social information for a standardised seven minutes (hatched areas). In (a) *Teaching* and *Emulation* chains, participants had access to social information for two minutes before starting building and throughout the building period. The 10 minutes during which participants acted as teachers or their tool was left in display are shown in blue. In (b) *Imitation* chains, participants could observe their predecessors for six minutes before building and an additional minute while building. Testing time (red) was unlimited, but for simplicity we depict it here as lasting three minutes (the mean time across all participants). Thus the actual chains were more fluid than these representations, within the parameters that learning (*Teaching* and *Emulation* conditions) and building (*Imitation* condition) always began when the participant two steps ahead in the chain finished testing their tool.

(b) Procedure for testing tools

Once the five minutes of building time had elapsed, participants moved into a screened-off area to test their tool. Here, they were presented with a large bowl filled with marbles of two sizes (total 3kg) and a scoop. Each participant was encouraged to incrementally load as many

marbles as possible into their tool. They were not allowed to make any adjustments to the tool prior to or during testing.

Builders of paper tools were instructed to float the tool in a water-filled tray and load as many marbles as possible into it without it sinking. They therefore had to use their judgement, gradually adding more marbles until they felt they could not risk adding any more. If the tool began to sink, the score was given as the total number of marbles added before the tool began taking on water. Across conditions, the percentage of tools that took on water was 25% in the asocial condition (out of a total 110 tools) and 37%, 36% and 41% in the emulation, imitation and teaching conditions respectively (out of 100 tools in each case). The differences between treatments were not statistically significant, either when examining across all conditions ($\chi 2 = 7.078$; $p = 0.069$) or within the social learning conditions ($\chi 2 = 0.594$; $p = 0.742$).

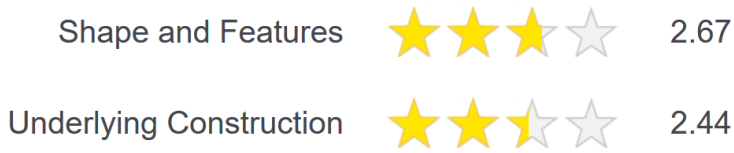
Builders of pipe-cleaner tools were instructed to fill their tool with as many marbles as they possibly could before carrying it by the handle or handles to a set of weighing scales 5m away. The loading of the implement took place on a tray and participants were permitted to lift their tool a few centimetres above the tray periodically during loading to check that the marbles were being held. When participants judged their tool could hold no more, they carried it the 5m distance to the scales. Once the implement left the tray area it was not possible to return or to add more marbles. Participants then deposited the marbles in a bowl on the scales. We then counted the total number of marbles successfully transported. Any marbles that fell out of the tool *en route* to the scales were not included in the totals. There were no cases in which the tool broke completely or the participant failed to transport any marbles.

(c) Similarity surveys

We used online Qualtrics surveys to generate estimates of the similarity between pairs of tools from the same experimental condition. One set of surveys gauged the similarity of tools and their successors within each transmission chain, while a second set of surveys compared pairs of tools from the same generation and condition between different chains (see main text). Survey participants were naïve to our research aims and hypotheses, and blind to experimental conditions.

Survey instructions explained that participants would be asked to rate how similar two tools were in terms of (a) **Shape and features** (whether the two tools look alike in terms of their overall shape and design features) and (b) **Underlying Construction.** For paper tools surveys, underlying construction refers to whether the implement has been made using the same types of folds, with the same precision. In the pipe-cleaner tool surveys, underlying construction refers to whether the pipe-cleaners are attached together in the same way, with similar-sized spaces between them.

Participants could either type their rating or click the cursor into an image of four stars and drag to highlight a score, ranging from 0.00 (no similarity) to 4.00 (identical). An example is shown below:



Before moving on in the survey, they had to demonstrate understanding of the rating method. They then went through a series of three quality control questions. In each question, they had to rate the similarity of two tools, which had been deliberately chosen to illustrate pairs of tools at the upper, lower and middle range of degrees of similarity. At the end of each question, participants were presented with a recommended similarity rating, which had been determined by the experimenter for that pair of tools. For instance: "In terms of Shape and Features the bottom implement looks to be an almost exact copy of the top. They are both rectangular boxes with deep sides, of identical proportions. We would suggest a score of 4. In terms of Underlying Construction the two implements have been constructed using exactly the same types of folds and with the same precision. We would suggest a score of 4." Participants whose answers deviated from the recommended scores by more than 0.5, or who went through the instructions and quality checks too quickly were excluded from the final dataset.

(c) Details of statistical analyses

To determine the factors influencing tool performance, we ran Linear Mixed Models with the total number of marbles carried fitted as the response term, with a random intercept and

slope, allowing the slope of change in performance across generations to differ between groups. In preliminary analyses using the entire dataset for both tool types, the best model included interactions between generation and both tool type and condition (Table S1;Table S2). For ease of interpretation, all subsequent analyses were therefore conducted on each tool type separately. We compared models including or excluding effects of condition, (Asocial, Emulation, Imitation, Teaching), generation and the interaction between condition and generation, as well as participants' craft experience, gender and the type of group (student or community group).

(i) Tool performance

To understand what determines the extent to which tools improve from one generation to the next within transmission chains, we built models where the response term was the difference in score between each tool and its main successor (defined as the tool two steps ahead in the chain, given that social learning from this tool was available across all three social learning conditions). As it is likely to be more difficult to improve on high-performing tools, we included the total number of marbles carried by each tool (hereafter referred to as "total") as a key predictor variable.

(ii) Improvements across the chain

To test whether the potential to improve upon high-performing tools depended on the type of social learning, we also examined the interaction between total and condition. Interactions between total and group type were examined as additional potential predictors, and group identity was fitted as a random term in all models.

(iii) Similarity between implements and their successors:

We then examined the factors influencing the degree of similarity between each tool and its successor in models with the mean similarity score from Survey 1 as the response variable. As ratings were continuously distributed, linear models with a Gaussian distribution provided a good model fit. We compared models including or excluding the effects of total (as above), condition and the interaction between them as well as generation and group type.

(iv) Convergence and diversification of designs: between-chain comparisons

Finally, we used the mean similarity scores from Survey 2 as the response term in glms to examine diversification and convergence of tool designs across chains. Models examined

whether tools became increasingly similar as chains progressed (generation 1, 5 or 10) as well as the effects of condition or the interaction between condition and generation.

(3) Statistical tables

model name	intercept	cond	gen	tool	cond* gen	cond* tool	gen* tool	cond* gen* tool	df	logLik	AICc	delta	weight
bothtools3	**5.682**	**+**	**0.5309**	**+**	**+**		**+**		**14**	**-1923.62**	**3875.8**	**0**	**0.787**
bothtools2	5.944	+	0.456	+	+				13	-1926.19	3878.8	3.06	0.17
bothtools4	5.79	+	0.5309	+	+	+	+		17	-1923.49	3881.7	5.97	0.04
bothtools1	5.976	+	0.477	+	+	+	+	+	20	-1922.82	3886.7	10.91	0.003

Table S1. Model selection summary for analyses of data including both tool types (cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model is shown in bold. The variance (s.d.) of the random effect was 1.29(1.14) and 0.038(0.19) for the intercept and slope respectively. The response variable was square root transformed.

variable	**estimate**	**S.E.**	**t value**	**p value**	**95% CI (lower)**	**95% CI (upper)**
(Intercept)	5.682	0.419	13.575	<0.001	4.852	6.512
generation	0.531	0.069	7.690	0.000	0.394	0.668
condition_e	-0.189	0.548	-0.346	0.730	-1.275	0.896
condition_i	0.196	0.548	0.358	0.722	-0.890	1.281
condition_t	-0.417	0.548	-0.762	0.448	-1.503	0.668
tool_pipe	-0.591	0.395	-1.498	0.138	-1.374	0.191
generation:condition_e	-0.192	0.090	-2.121	0.037	-0.371	-0.012
generation:condition_i	-0.312	0.090	-3.454	0.001	-0.491	-0.133
generation:condition_t	-0.110	0.090	-1.215	0.228	-0.289	0.069
generation:tool_pipe	-0.150	0.065	-2.300	0.024	-0.279	-0.021

Table S2. Summary for the best supported model in Table S1 (bothtools3). Experimental conditions are e (emulation); i (imitation); t (teaching), with the asocial condition (a) as the reference category; pipe = pipe-cleaner tools. Conditional R^2 for this model =0.482. The variance (s.d.) of the random effect was 1.29(1.13) and 0.04(0.19) for the intercept and slope respectively.

model name	intercept	gen	cond	cond* gen	craft	gender	grouptype	df	logLik	AICc	delta	weight
paperslope3	**4.890**	**0.389**			**0.453**			**7**	**-994.6**	**2003.40**	**0.00**	**0.574**
paperslope7	4.898	0.389			0.452		+	8	-994.6	2005.50	2.08	0.203
paperslope6	5.606	0.390	+		0.464			10	-992.6	2005.80	2.32	0.180
paperslope4	5.284	0.477	+	+	0.462			13	-991.2	2009.20	5.79	0.032
paperslope8	5.292	0.477	+	+	0.461		+	14	-991.2	2011.40	7.93	0.011
paperslopebasic	5.584	0.385						6	-1011.4	2034.90	31.50	0.000
paperslope2	6.292	0.385	+					9	-1009.7	2037.80	34.33	0.000
paperslope1	5.976	0.477	+	+				12	-1008.2	2041.20	37.78	0.000
paperslope5	5.999	0.477	+	+		+		13	-1008.2	2043.30	39.88	0.000
paperslopenull	6.825							5	-1027.9	2065.90	62.46	0.000

Table S3. Model selection summary for analyses of the performance of paper tools (cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model is shown in bold. The response variable was square root transformed.

variable	**estimate**	**S.E.**	**t value**	**p value**	**95% CI (lower)**	**95% CI (upper)**
(Intercept)	4.892	0.373	13.098	<0.001	4.152	5.623
generation	0.389	0.055	7.039	<0.001	0.280	0.499
craft	0.452	0.135	3.360	0.001	0.188	0.719

Table S4. Summary for the best supported model in Table S3 (paperslope3). Conditional R^2 for this model =0.418. The variance (s.d.) of the random effect was 1.27(1.13) and 0.056(0.24) for the intercept and slope respectively.

model name	intercept	gen	cond	cond* gen	craft	gender	grouptype	df	logLik	AICc	delta	weight
pipeslope4	**4.21**	**0.435**	**+**	**+**	**0.4695**			**13**	**-882.19**	**1791.3**	**0**	**0.473**
pipeslope8	4.375	0.435	+	+	0.4604		+	14	-881.8	1792.6	1.37	0.239
pipeslope3	4.304	0.247			0.4705			7	-889.5	1793.3	1.99	0.175
pipeslope7	4.452	0.247			0.4616		+	8	-889.16	1794.7	3.38	0.087
pipeslope6	4.855	0.247	+		0.4808			10	-888.25	1797	5.76	0.026
pipeslope1	4.797	0.435	+	+				12	-899.93	1824.6	33.34	0
pipeslope5	4.815	0.435	+	+		+		13	-899.84	1826.6	35.29	0
pipeslopebasic	4.993	0.235						6	-907.27	1826.7	35.46	0
pipeslope2	5.503	0.235	+					9	-906.42	1831.3	39.99	0
pipeslopenull	5.692							5	-918.22	1846.6	55.3	0

Table S5. Model selection summary for analyses of the performance of pipe-cleaner tools (cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model is shown in bold. The top model set showed some support for an effect of group type, but this was not robust (student group scores < community groups: β (s.e.) = -0.372 (0.446), t = -0.834, p = 0.410; CI (-1.21; 0.47)). The response variable was square root transformed.

variable	estimate	S.E.	t value	p value	95% CI (lower)	95% CI (upper)
Intercept	4.215	0.498	8.458	0.000	3.264	5.164
generation	0.435	0.076	5.744	0.000	0.291	0.580
condition_e	0.251	0.713	0.352	0.727	-1.107	1.609
condition_i	0.456	0.712	0.641	0.526	-0.900	1.811
condition_t	-0.302	0.714	-0.423	0.675	-1.663	1.058
craft	0.465	0.108	4.315	0.000	0.257	0.682
generation:condition_e	-0.292	0.113	-2.585	0.014	-0.508	-0.077
generation:condition_i	-0.366	0.112	-3.260	0.002	-0.581	-0.152
generation:cond ition_t	-0.132	0.113	-1.168	0.250	-0.348	0.084

Table S6. Summary for the best supported model in Table S5 (pipeslope4). Experimental conditions are e (emulation); i (imitation);t (teaching), with the asocial condition(a) as the reference category. Conditional R^2 for this model =0.483. The variance (s.d.) of the random effect was 0.95(0.97) and 0.022(0.14) for the intercept and slope respectively.

variable	estimate	S.E.	t value	p value	95% CI (lower)	95% CI (upper)
Teaching vs imitation	0.234	0.102	2.307	0.025	0.031	0.437
Teaching vs emulation	0.155	0.120	1.288	0.208	-0.091	0.400

Table S7. Pairwise comparisons of slopes of improvement in teaching vs imitation and emulation chains.

model name	intercept	cond	total	cond* total	craft	grouptype	grouptype * total	cond* grouptype	cond* grouptype* total	gender	gen	df	logLik	AICc	delta	weight
papn2_9	**35.48**		**-0.6102**		**6.341**							**5**	**-1237.29**	**2484.8**	**0**	**0.939**
papn2_3	40.22	+	-0.6722	+	6.443							9	-1235.85	2490.5	5.66	0.055
papn2_8	32.96	+	-0.8499	+							5.909	9	-1238.19	2495.2	10.32	0.005
papn2_1	45.57		-0.6272									4	-1251.02	2510.2	25.36	0
papn2_2	46.96	+	-0.624									6	-1250.57	2513.5	28.65	0
papn2_5	54.01	+	-0.8412	+		+	+					10	-1246.63	2514.2	29.37	0
papn2_4	44.76	+	-0.6755	+		+						9	-1248.83	2516.4	31.61	0
papn2	50.75	+	-0.6921	+								8	-1250.02	2516.7	31.83	0
papn2_7	53.52	+	-0.7018	+						+		9	-1249.41	2517.6	32.77	0
papn2_6	59.19	+	-0.854	+		+	+	+	+			14	-1245.29	2520.4	35.61	0
papn2_null	10.07											3	-1283.37	2572.9	88.01	0

Table S8. Model selection summary for analyses of the relative difference between the performance of each paper tool and its successor in the chain (total = marbles carried by the model paper tool; cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model is shown in bold.

variable	estimate	S.E.	t value	p value	95% CI (lower)	95% CI (upper)
Intercept	35.7829	6.52899	5.481	< 0.001	22.392	49.195
total	-0.6145	0.06642	-9.251	< 0.001	-0.752	-0.466
craft	6.30033	2.65026	2.377	0.0182	1.127	11.547

Table S9. Summary for the best supported model in Table S8 (papn2_9). Conditional R^2 for this model =0.351. The variance (s.d.) of the random intercept was 212.80(14.59).

model name	intercept	cond	total	cond* total	craft	grouptype	grouptype * total	cond* grouptype	cond* grouptype * total	gender	gen	df	logLik	AICc	delta	weight
pipn2log_5	5.527	+	-0.006	+		+	+					10	131.6	-242.2	0.00	0.41
pipn2log_8	5.491	+	-0.006	+							0.006	9	129.6	-240.4	1.79	0.17
pipn2log	**5.515**	**+**	**-0.006**	**+**								**8**	**128.5**	**-240.4**	**1.81**	**0.16**
pipn2log_4	5.507	+	-0.006	+		+						9	129.0	-239.1	3.04	0.09
pipn2log_7	5.518	+	-0.006	+						+		9	128.7	-238.6	3.55	0.07
pipn2log_6	5.505	+	-0.006	+		+	+	+	+			14	134.1	-238.3	3.86	0.06
pipn2log_3	5.481	+	-0.006	+	0.024							9	128.3	-237.7	4.45	0.04
pipn2log_1	5.493		-0.005									4	114.8	-221.4	20.73	0.00
pipn2log_2	5.489	+	-0.005									6	116.8	-221.3	20.92	0.00
pipn2log_9	5.467		-0.005		0.019							5	113.3	-216.4	25.83	0.00
pipn2log_null	5.304											3	53.2	-100.2	141.98	0.01

Table S10. Model selection summary for analyses of the relative difference between the performance of each pipe-cleaner tool and its successor in the chain (total = marbles carried by the model pipe-cleaner tool; cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model (following application of the nesting rule) is shown in bold. Examination of Cook's distances revealed potentially influential datapoints (>3x mean Cook's distance, but still <1), but the results remained consistent when these were removed. The response variable was log transformed.

variable	estimate	S.E.	t value	p value	95% CI (lower)	95% CI (upper)
Intercept	5.516	0.030	182.942	< 0.001	5.458	5.573
total	-0.006	0.001	-10.630	< 0.001	-0.007	-0.005
condi	0.004	0.044	0.089	0.930	-0.082	0.087
condt	-0.074	0.043	-1.717	0.091	-0.161	0.008
total:condi	-0.001	0.001	-1.317	0.189	-0.003	0.001
total:condt	0.003	0.001	3.721	0.000	0.001	0.004

Table S11. Summary for the best supported model in Table S10 (pipn2log). Conditional R^2 for this model =0.551. The variance (s.d.) of the random intercept was 0.002 (0.049).

variable	estimate	S.E.	t value	p value	95% CI (lower)	95% CI (upper)
Teaching vs imitation	0.004	0.001	4.684	<0.001	0.001	0.004
Teaching vs emulation	0.003	0.001	3.667	<0.001	0.002	0.005
Emulation vs imitation	-0.001	0.001	-1.110	0.269	-0.002	0.001

Table S12. Pairwise comparisons of slopes of degradation between focal pipe-cleaner tools and their successors.

model name	intercept	total	cond	total* cond	grouptype	gen	df	logLik	AICc	delta	weight
papsim_4	-1.692	0.010				0.105	5.000	-396.03	802.30	0.0	0.886
papsim_1	**-1.298**	**0.012**					**4.000**	**-400.05**	**808.30**	**6.0**	**0.045**
papsim_5	-1.804	0.011	+	+		0.106	9.000	-394.97	808.70	6.4	0.036
papsim_3	-1.270	0.012			+		5.000	-399.92	810.10	7.8	0.018
papsim_2	-1.374	0.012	+				6.000	-399.15	810.70	8.4	0.014
papsim_basic	-1.421	0.012	+	+			8.000	-399.08	814.80	12.5	0.002
papsim_null	-0.639						3.000	-417.93	842.00	39.7	0.000

Table S13. Model selection summary for analyses of the similarity between each paper tool and its successor in the chain (total = marbles carried by the model paper tool; cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model (following application of the nesting rule) is shown in bold.

variable	estimate	S.E.	t value	p value	95% CI (lower)	95% CI (upper)
Intercept	-1.302	0.137	-9.523	<0.001	-1.577	-1.031
total	0.012	0.002	6.225	<0.001	0.008	0.015

Table S14. Summary for the best supported model in Table S13 (paperaall_1). Conditional R^2 for this model =0.147. The variance (s.d.) of the random intercept was 0.002(0.04).

model name	intercept	total	cond	total* cond	grouptype	gen	df	logLik	AICc	delta	weight
pipsim_basic	**-1.208**	**0.012**	**+**	**+**			**8**	**-344.8**	**706.2**	**0**	**0.715**
pipsim_5	-1.151	0.013	+	+		-0.014	9	-344.7	708.1	1.94	0.272
pipsim_2	-1.221	0.013	+				6	-351.6	715.5	9.31	0.007
pipsim_1	-1.453	0.013					4	-354.3	716.7	10.59	0.004
pipsim_3	-1.442	0.013			+		5	-354.3	718.8	12.62	0.001
pipsim_4	-1.452	0.013				0.000	5	-354.3	718.8	12.68	0.001
pipsim_null	-0.958						3	-368.3	742.6	36.45	0.000

Table S15. Model selection summary for analyses of the similarity between each pipe-cleaner tool and its successor in the chain (total = marbles carried by the model paper tool; cond = condition; gen = generation). The top model set (ΔAICc<6) is highlighted in grey; the best supported model is shown in bold.

variable	**estimate**	**S.E.**	**t value**	**p value**	**95% CI (lower)**	**95% CI (upper)**
Intercept	-1.208	0.191	-6.324	<0.001	-1.579	-0.837
total	0.012	0.004	3.216	0.001	0.005	0.020
cond(i)	0.123	0.278	0.441	0.660	-0.418	0.664
Cond(t)	-0.676	0.268	-2.519	0.012	-1.198	-0.155
total:cond(i)	-0.014	0.006	-2.225	0.027	-0.026	-0.002
total:cond(t)	0.009	0.005	1.618	0.107	-0.002	0.019

Table S16. Summary for the best supported model in Table S15 (pipsim_basic). Conditional R^2 for this model =0.175. The variance (s.d.) of the random intercept was 0.00 (0.00).

variable	**estimate**	**S.E.**	**t value**	**p value**	**95% CI (lower)**	**95% CI (upper)**
Teaching vs imitation	0.022	0.006	3.54	<0.001	0.010	0.035
Imitation vs emulation	-0.014	0.006	-2.223	0.026	-0.026	-0.001
Teaching vs emulation	0.009	0.005	1.692	0.093	-0.001	0.018

Table S17. Pairwise comparisons of slopes the relationship between the performance of a pipe-cleaner tool and the similarity of its successor.

(a) Shape and features									
model name	**intercept**	**gen**	**cond**	**cond* gen**	**df**	**logLik**	**AICc**	**delta**	**weight**
prop_glm_papera5	-1.587	+	+		6	-57.166	128.6	0.00	0.524
prop_glm_papera3	**-1.261**		**+**		**4**	**-60.445**	**129.9**	**1.31**	**0.271**
prop_glm_papera2	-1.511	+			4	-61.367	131.8	3.16	0.108
prop_glm_papera1	-1.209				2	-63.866	132.0	3.42	0.095
prop_glm_papera4	-1.320	+	+	+	10	-56.304	139.3	10.67	0.003
(b) Underlying construction									
model name	**intercept**	**gen**	**cond**	**cond* gen**	**df**	**logLik**	**AICc**	**delta**	**weight**
prop_glm_paperb5	-1.561	+	+		6	-50.144	114.6	0.00	0.361
prop_glm_paperb2	**-1.297**	**+**			**4**	**-53.028**	**115.1**	**0.52**	**0.278**
prop_glm_paperb1	-1.028				2	-55.785	115.9	1.30	0.188
prop_glm_paperb3	-1.273		+		4	-53.515	116.1	1.50	0.171
prop_glm_paperb4	-1.360	+	+	+	10	-49.568	125.8	11.24	0.001

Table S18. Model selection summary for between-chain comparisons of the similarity in paper tools. The top model set (ΔAICc<6) is highlighted in grey; the best supported model (after application of the nesting rule) is shown in bold. Condition appears in some models in the top set, but the effect is not robust (emulation vs teaching; shapes and features: $\beta = 0.576$, S.E. = 0.368, t = 1.57, p = 0.125, CI (-0.145, 1.298); underlying construction $\beta = 0.631$, S.E. = 0.314, t = 2.007, p = 0.051, CI (0.015, 1.247)).

(a) Shape and features									
model name	**intercept**	**gen**	**cond**	**cond* gen**	**df**	**logLik**	**AICc**	**delta**	**weight**
prop_glm_pipea1	**-1.439**				**2**	**-55.799**	**115.9**	**0.00**	**0.544**
prop_glm_pipea2	-1.690	+			4	-54.065	117.1	1.25	0.292
prop_glm_pipea3	-1.299		+		4	-55.015	119.0	3.15	0.113
prop_glm_pipea5	-1.550	+	+		6	-53.217	120.6	4.76	0.050
prop_glm_pipea4	-1.164	+	+	+	10	-51.102	128.7	12.79	0.001
(b) Underlying construction									
model name	**intercept**	**gen**	**cond**	**cond* gen**	**df**	**logLik**	**AICc**	**delta**	**weight**
prop_glm_pipeb1	**-1.601**				**2**	**-68.221**	**140.7**	**0.00**	**0.504**
prop_glm_pipeb2	-2.036	+			4	-66.099	141.2	0.47	0.398
prop_glm_pipeb3	-1.747		+		4	-67.997	145.0	4.27	0.060
prop_glm_pipeb5	-2.182	+	+		6	-65.853	145.9	5.19	0.038
prop_glm_pipeb4	-1.545	+	+	+	10	-63.129	152.7	12.00	0.001

Table S19. Model selection summary for between-chain comparisons of the similarity in pipe-cleaner tools. The top model set (ΔAICc<6) is highlighted in grey; the best supported model is shown in bold. Generation features in some of the models in the top sets (though not in the top models), but the effect is not robust: (shapes and features: $\beta = 0.551$, S.E. = 0.304, t = 1.81, p = 0.077, CI (-0.045, 1.147); underlying construction $\beta = 0.801$, S.E. = 0.397, t = 2.015, p = 0.051, CI (0.021, 1.580)).

(4) Supplementary figures

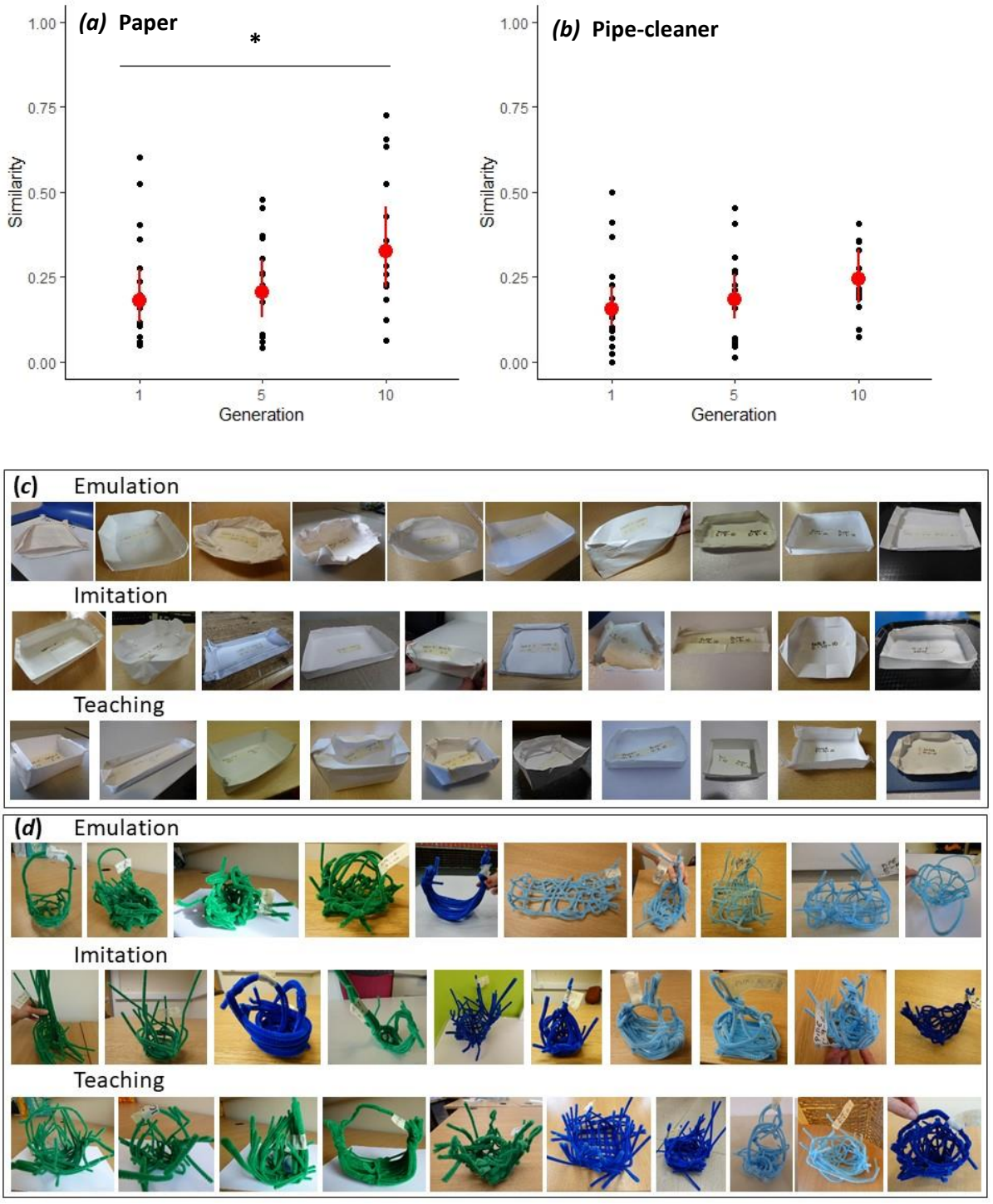


Figure S3. Paper tools from different chains become more similar to each other as chains progressed (a). Photographs of the final (10^{th}) generation illustrate that paper tools tended to converge on similar, flat-bottomed designs (c). In contrast, for pipe-cleaner tools there was no detectable increase in similarity as chains progressed (b), reflected in a greater diversity of final (10^{th} generation) designs (d). (a) and (b) show raw data (black dots) and means and CIs (red) for similarity in terms of shape and features. Analyses of similarity in terms of underlying construction produced qualitatively similar results.

(5) Participating Community Groups

We are hugely indebted to the following community groups from across Cornwall who took the time to take part in the study, either kindly welcoming us to their premises or visiting us at Exeter University (Penryn Campus). Special thanks to those who provided more than one group of 10 participants. The full list of participating groups is below:

Cober Valley Rotary Club, Cornwall Partnership Foundation NHS Trust, Cornwall College Staff, Lane and District Women's Institute, Hall For Cornwall, Devoran Pilot & Gig Club, Bude Women's Institute, Dracaena Centre, Tolvadden Community Fire Station, Falmouth TM Meditation, Knit & Natter Newquay, Mawnan Smith Women's Institute, Falmouth Cruise Ship Ambassadors, Nankersey Rowing Club, National Maritime Museum, Newquay Zoo, The Monkey Sanctuary, Potager Gardeners, Craft & Conviviality, Mylor Women's Institute, Cornwall Food Foundation, Camel Creek Adventure Park, Forms Plus Helston, Bodmin Community Organisers, Olive Branch Café, Falmouth and Exeter Student's Union (staff), Penpol Crafters, Falmouth and Exeter Library (staff), Falmouth Road Runners, Ludgvan Women's Institute, Bude Sewing Club, Active Plus, Red Wing Gallery, Mature Student Society, The Nute Family, University of Exeter Biosciences Staff, Falmouth Art Gallery.

(6) Potential Impacts of Group Composition

In all cases, members of participating groups knew each other. Although it was beyond the aims and scope of this study to examine the impacts of group composition, familiarity between individuals may well have facilitated social learning, as has been demonstrated in previous studies on humans and non-human animals [e.g. 4, 5]. Given that social network structure and the nature of relationships between group members are likely to have important influences on social transmission [5-8], future work on the origins of human cumulative culture would benefit from incorporating knowledge of the likely structure of ancestral groups.

References

1. Warton DI, Hui FKC. 2011 The arcsine is asinine: the analysis of proportions in ecology. *Ecology* **92**, 3–10.
2. Miton H, Charbonneau M. 2018 Cumulative culture in the laboratory: Methodological and theoretical challenges. *Proc. R. Soc. B* **285,** 20180677 (doi:10.1098/rspb.2018.0677)

3. Caldwell CA, Millen AE. 2009 Social learning mechanisms and cumulative cultural evolution: is imitation necessary? *Psychol. Sci.* **20**, 1478–1483.

4. Swaney W, Kendal J, Capon H, Brown C, Laland KN. 2001 Familiarity facilitates social learning of foraging behaviour in the guppy. *Anim. Behav*. **62**, 591-598.

5. Wood LA, Kendal RL, Flynn EG. 2013 Whom do children copy? Model-based biases in social learning. *Dev. Rev.* **33**, 341-356.

6. Coussi-Korbel, Fragaszy DM. 1995 On the relation between social dynamics and social learning. *Anim. Behav.* **50**, 1441–1453.

7. Kurvers, RH, Krause, J, Croft, DP, Wilson, AD, Wolf, M. 2014. The evolutionary and ecological consequences of animal social networks: emerging issues. *Trends Ecol. Evol.* **29**, 326-335.

8. Derex M, Mesoudi A. 2020 Cumulative cultural evolution within evolving population structures. *Trends Cogn. Sci.* **24**, 654-667.