Teaching and learning generalizable abstractions

August 28, 2024

Huang Ham[1], Bonan Zhao[2], Thomas L. Griffiths[1,2], Natalia Vélez[1]

[1]Department of Psychology, Princeton University

[2]Department of Computer Science, Princeton University

Keywords: pedagogy; social learning; Bayesian computational models; cultural evolution

Address for correspondence:

Natalia Vélez

Department of Psychology

Princeton University

Princeton, NJ 08540

E-mail: nvelez@princeton.edu

Abstract

A hallmark of effective teaching is that it grants learners not just a collection of facts about the world, but also a toolkit of abstractions that can be applied to solve new problems. How do humans transmit and acquire generalizable abstractions from examples? Here, we applied Bayesian models of pedagogy to a necklace-building task where teachers create necklaces to teach a learner "motifs" that can be flexibly recombined to create new necklaces. In Experiment 1 (N = 151), we find that human teachers produce necklaces that are simpler (i.e., have lower algorithmic complexity) than would be expected by chance, as indexed by a model that samples uniformly from all necklaces that contain the target motifs. This tendency to select simpler examples is partially captured by a pedagogical sampling model that tries to maximize the learner's belief in the underlying motifs. In Experiment 2 (N = 295), we find that simplicity is beneficial. Human learners recover the underlying motifs better when teachers produce simpler sequences, and they learn best from human teachers rather than from model-generated examples. Our results suggest that the computational principles that underlie effective communication and teaching may also provide a first step towards understanding the transmission of culturally-specific abstractions.

1 Introduction

Teaching is a powerful means of transmitting culturally-specific abstractions, thereby laying the foundations to accumulate generalizable skills and knowledge across generations (Kline, 2015; Legare, 2019). One important kind of abstraction is a *motif*, a recurring pattern that can be composed into a larger work. For example, the basic stitches in knitting (knits and purls) are used to make recurring motifs or stitch patterns (e.g., stockinette, rib stitch) that can be flexibly combined to make many products (e.g., hats, scarves, sweaters). Motifs are found in many cultural products and often bear traces of the communities that produced them (Pesowski, Quy, Lee, & Schachner, 2020; Schachner, Brady, Oro, & Lee, 2018), such as the elaborate cable-knit patterns of Aran sweaters, the fundamental rhythmic pattern or *clave* of salsa music, and meander designs on the borders of Greek pottery.

Motifs pose a particular challenge for existing computational theories of social learning. Existing pedagogical sampling models characterize teaching and social learning as a series of recursive inferences: Teachers select examples that will maximize a learner's belief in a target concept, and learners work backwards from the examples provided to infer the concept that the teacher is trying to communicate to them (Gweon, 2021; Shafto, Goodman, & Griffiths, 2014; Shafto, Wang, & Wang, 2021). This basic principle can explain a wide variety of communicative behaviors, including how teaching through demonstration differs from goal-directed behavior (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016; Tominaga, Knoblich, & Sebanz, 2022), how parents tune their speech to teach phonetic structures to infants (Eaves, Feldman, Griffiths, & Shafto, 2016), and how teachers improve their teaching based on feedback from learners (Chen, Palacci, Vélez, Hawkins, & Gershman, 2024). In addition, these computations appear to be neurally instantiated in mentalizing regions when teachers make decisions about what information to communicate to a learner (Vélez, Chen, Burke, Cushman, & Gershman, 2023).

However, prior work has largely focused on capturing how learners acquire solutions to particular problems (such as how to operate a particular toy; Aboody, Velez-Ginorio, Santos, &

Jara-Ettinger, 2023; Bridgers, Jara-Ettinger, & Gweon, 2020; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011) or identify the extension of particular categories (such as inferring the extent of a hidden shape on a canvas; Shafto et al., 2014; Vélez et al., 2023). Most of these problems involve a teacher choosing examples that constitute a part of the target that they intend to teach, such as a single function of a toy with many functions or a single pixel inside a larger shape. Teaching motifs presents the converse problem. For example, suppose an expert knitter draws a novice's attention to the rib stitch pattern on the collar of a sweater. Her goal is not to help the novice identify sweaters based on this distinctive sub-component, as is the case in traditional teaching games (Avrahami et al., 1997). Rather, the sub-component *itself* is the target of teaching; the novice can later abstract away this sub-component to knit sock cuffs and hatbands. We do not yet know to what extent pedagogical sampling models capture human behavior in this kind of teaching problem, where examples of the whole are used to teach parts.

To approach this question, we studied how people teach and learn motifs within a simple necklace-building task inspired by prior studies of cultural transmission (Clegg & Legare, 2016a, 2016b; Kleiman-Weiner et al., 2020). In Experiment 1 (N = 151), we tested to what extent existing pedagogical sampling models capture how teachers transmit motifs. In this task, teachers demonstrated motifs—recurring patterns of beads—by providing a single sample necklace. Overall, our pedagogical sampling model provides a better quantitative fit to teachers' decisions, compared to a baseline model that samples uniformly from all necklaces that contain the target motifs. The pedagogical sampling model also captures an important qualitative pattern in teachers' behavior: Teachers produce simpler examples than would be expected by chance. Next, in Experiment 2 (N = 295), we tested the limits of existing models by giving human learners a single sample necklace and asking them to both infer the underlying motifs and produce two new necklaces that incorporate these motifs. Overall, learners perform better at both tasks when given examples that are generated by human teachers or sampled from the pedagogical sampling model, compared to necklaces generated by the baseline model. The effectiveness of these examples is largely explained by simplicity—across the board, learners are better able to recover underlying motifs when given simpler examples. However, learners performed best overall when given human-generated examples, which suggests that state-of-the-art pedagogy models still miss aspects of what makes human teaching effective. We close by discussing how pedagogical sampling models could be extended to better capture how abstractions are culturally transmitted. All experiment materials, data, and analysis code are publicly available at `https://osf.io/rnb9e/?view_only=099dadc807964263a8e1196ce3dd2311`.

2 Computational framework

2.1 Task setup

The experiments below use a necklace-building task as a simple case study of how people acquire and transmit culturally-specific motifs. In prior work, necklace-building tasks have been used to study basic mechanisms that drive cultural transmission, including how faithfully children imitate (Clegg & Legare, 2016a, 2016b) and how adults learn concepts from step-by-step demonstrations (Kleiman-Weiner et al., 2020). Experimental stimuli were adapted from Kleiman-Weiner et al. (2020).

In Experiment 1 (Teaching Abstractions), participants play the role of expert artisans; their task is to travel from one village to another, teaching an apprentice how to produce necklaces that will sell well in each village. In our setting, each necklace is a string of 10 orange and green beads that can be represented as a binary sequence. Each village has three favorite motifs, which are subsequences of 2 or 3 beads that can be recombined to make necklaces (Figure 1A). A necklace sells well in a village if and only if it includes all three of the motifs favored in that village (Figure 1B). In Experiment 2 (Learning Abstractions), participants played the role of the apprentice. In each village, participants saw one necklace generated by an expert teacher; their task was to infer the underlying motifs in the necklace and to use these motifs to create new necklaces of their own.

In both experiments, we modeled teachers' and learners' behavior using two models: a baseline model that assumes that the teacher samples uniformly among all necklaces that contain the correct motifs, and a Bayesian pedagogy model that selects necklaces to teach that will

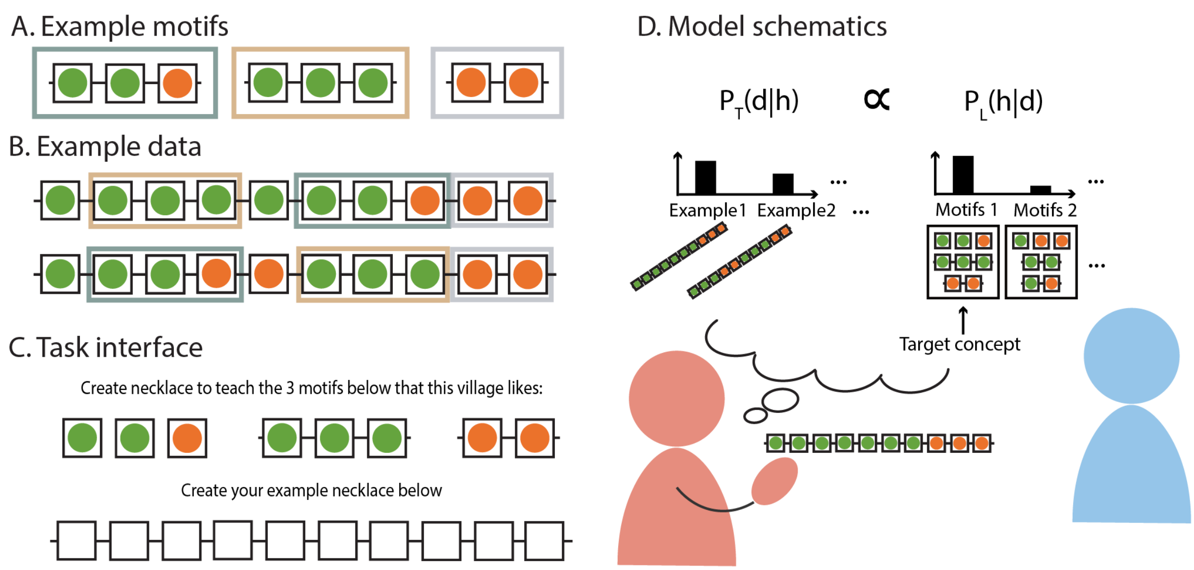

Figure 1: Teaching task. A-C. Experiment interface: A. On each trial, participants were shown the motifs favored by each village. B. These motifs can be recombined to create many new necklaces. C. Participants created a single sample necklace to teach learners these three motifs. D. Model schematic: The pedagogical sampling model actively selects necklaces d to teach $(P_T(d|h)))$ by anticipating how learners will recover the underlying motifs h from the sample necklace $(P_L(h|d))$.

maximize the learner's posterior beliefs in the correct motifs. To borrow terminology from prior work (Shafto et al., 2014), we will refer to these models as the "strong sampling" and "pedagogical sampling" models, respectively. We explain these models in more detail below.

2.2 Strong sampling model

Strong sampling refers to a process where examples are uniformly sampled from a target hypothesis (Shafto et al., 2014; Tenenbaum & Griffiths, 2001). In our task, each "example" is a sample necklace, and each "hypothesis" is a set of three motifs. In other words, the strong sampling model chooses uniformly among all necklaces that contain all three motifs favored by a particular village. We compared the strong sampling model to the behavior of both teachers and learners. In Experiment 1, comparing the fit of the pedagogical sampling model to that of the strong sampling model provides evidence about the extent to which teachers' decisions are guided by higher-order inferences about a hypothetical learner's beliefs (see *Teacher model*, below). In Experiment 2, we constructed a learner model that assumes that the teacher's examples were selected using strong sampling; we used this model as a baseline to measure how accurately human learners could be expected to recover the target hypothesis based on a single example (see *Learner baseline model*, below).

2.2.1 Teacher model

Given a hypothesis h, the strong sampling model predicts that the probability of selecting any sample necklace d is inversely proportional to the number of necklaces that contain the target motifs:

$$P_{\text{strong}}(d|h) = \begin{cases} \frac{1}{|h|} & \text{if } d \in h \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We generated the predictions of the strong sampling model by first creating a matrix $\boldsymbol{C}$ of hypotheses and data. The rows contain all possible 10-bead necklaces ($2^{10} = 1024$ possible necklaces) and the columns contain all possible triplets of motifs (with 2^2 possible 2-bead motifs and 2^3 3-bead motifs, there are $\binom{2^2+2^3}{3} = 220$ possible hypotheses). Each cell $c_{d,h}$ of this matrix indicates whether the necklace d could have been generated by combining the three motifs in h ($c_{d,h} = 1$) or not ($c_{d,h} = 0$). Thus, Equation 1 is equivalent to the following matrix operation:

$$P_{\text{strong}}(d|h) = \frac{c_{d,h}}{\sum_{d'} c_{d',h}}. \tag{2}$$

which corresponds to normalizing the entries of each column by its sum.

2.2.2 Baseline learner model

The baseline learner model assumes that the teacher selects a sample necklace d using strong sampling. We can use Bayes' rule to obtain the posterior distribution over motifs given the example provided:

$$P_{baseline}(h|d) = \frac{P_{strong}(d|h)P(h)}{\sum_{h'} P_{strong}(d|h')P(h')}. \tag{3}$$

where $P_{strong}(d|h)$ is the probability of selecting the sample necklace d under strong sampling when the true hypothesis is h (Equation 1) and $P(h)$ is the prior probability assigned to the hypothesis h.

The model further assumes a uniform prior over all sets of motifs (h), so $P(h)$ is the same for all h. Thus, Equation 3 can be simplified as:

$$P_{baseline}(h|d) = \frac{P_{strong}(d|h)}{\sum_{h'} P_{strong}(d|h')}. \tag{4}$$

2.3 Pedagogical sampling model

The pedagogical sampling model chooses necklaces to teach by anticipating how the learner will recover motifs from the example provided. Thus, rather than sampling uniformly from all necklaces that contain the target motifs, this model favors necklaces that are consistent with fewer alternative hypotheses. Intuitively, this strategy reduces the risk that learners will recover the wrong set of motifs from the sample necklace.

As an illustration, suppose that a village favors necklaces that contain the motifs “000”, “11”, and “001”. These motifs can be used to make many necklaces, including the following two examples:

$$\begin{aligned} 0000000111 &\rightarrow [000]00[001][11] \\ 0001100011 &\rightarrow 0[001]1[000][11] \end{aligned} \tag{5}$$

Here, the left side of each line shows the necklace as it would appear to a learner, and the right side breaks down each example into its component motifs. Note that both of these examples are ambiguous; there are several alternative sets of motifs that could have generated either of these necklaces. However, the model favors the necklace “0000000111” because there are fewer incorrect ways to parse it. Besides the 3 target motifs, this necklace is consistent with 4 incorrect motifs (i.e., “111”, “011”, “01”, “00”), which make up 13 consistent but incorrect alternative hypotheses (e.g., “00|000|11”, “11|01|000”). By contrast, “0001100011” can be parsed incorrectly in more ways. In addition to the 3 correct motifs, this necklace is consistent with 6 incorrect motifs (i.e., “00”,“01”,“10”,“100”,“011”,“110”), which make up 46 consistent but incorrect alternative hypotheses (e.g., “00|100|11”, “110|011|000”).

More formally, this model characterizes pedagogy as a form of cooperative communication between a teacher and a learner (Shafto et al., 2021). The model assumes that both the teacher and the learner have common knowledge about a space of hypotheses (i.e., all possible triplets of motifs) and a space of data (i.e., all possible necklaces; Shafto et al., 2014). The teacher selects necklaces to show to the learner that will maximize the learner's beliefs in the true set of motifs favored by a particular village, and the learner works backwards from the necklace provided to infer what motifs the teacher is trying to demonstrate. These recursive inferences between the teacher and the learner are captured using the following system of equations:

$$P_{\text{learner}}(h|d) = \frac{P_{\text{teacher}}(d|h)p(h)}{\sum_{h'} P_{\text{teacher}}(d|h')p(h')}, \tag{6}$$

$$P_{\text{teacher}}(d|h) \propto P_{\text{learner}}(h|d)^{\alpha}, \tag{7}$$

where α is a free parameter that controls how strongly the teacher favors examples that maximize the learner's posterior belief in the true motifs. In the results below, we fit α to individual participants' responses in Experiment 1.

Following the procedure in Shafto et al. (2014), we calculated a solution to this system of equations using fixed-point iteration. We began by normalizing each column of the matrix $\boldsymbol{C}$ by its sum, as in the strong sampling model (Equation 2). Next, we implemented the recursive inferences that distinguish pedagogical sampling from strong sampling by renormalizing this matrix. In the first iteration of this procedure, we normalized each row by its sum to generate $P_{\text{learner}}(h|d)$. Intuitively, each row of this matrix represents the posterior beliefs of a learner who assumes that the sample necklace d was generated using strong sampling. (Note that these first two iterations are equivalent to the strong sampling and learner models described in the previous section.) In the next iteration, we raised $\boldsymbol{C}$ to the power of α and normalized each column by its sum to generate $P_{\text{teacher}}(d|h)$. Each column of this matrix represents the choice probabilities of a teacher who selects examples by considering the beliefs of the learner in the prior iteration.

Each iteration of this procedure represents an additional recursive inference. For example, the next iteration yields the beliefs of a learner who tries to interpret what hypothesis the teacher is attempting to communicate, and the next iteration after that yields the choice probabilities of a teacher who tries to maximize the beliefs of a learner who actively interprets their examples, and so on. In principle, we could iterate this system of equations indefinitely; in practice, $P_{\text{learner}}(h|d)$ and $P_{\text{teacher}}(d|h)$ converge to a fixed point after a finite number of steps (Wang, Wang, Paranamana, & Shafto, 2020). We set the tolerance of convergence to 10^{-12}.

3 Experiment 1: Teaching abstractions

In Experiment 1, participants played the role of teachers. Their task was to provide a single sample necklace to teach a learner generalizable "motifs" that could be recombined to produce new necklaces that would sell well in a particular village. We compared the necklaces generated by human teachers to those selected by a model that randomly selects a necklace that is consistent with the target motifs (strong sampling) and a model that maximizes the learner's belief in the target motifs (pedagogical sampling).

3.1 Methods

3.1.1 Participants

We aimed to recruit 150 participants for the teaching task on Prolific (preregistration available at https://aspredicted.org/GPT_HNV). In both experiments, participants were recruited through the standard sample option; only participants who resided in the US, had an approval rate over 95%, were fluent in English, and completed 100 to 10000 studies were eligible to sign up. We obtained data from 151 participants (M(SD) age = 39.78(13.73), 94 female, 53 male, and 4 non-binary), potentially due to a server error at the time of submission. Participants earned $3 for their participation, and they were told that they could earn a performance bonus of up to $1 based on how well participants in Experiment 2 learned from the examples they selected. Thus, participants were incentivized to provide helpful examples. In all experiments, participants provided informed consent in accordance with the requirements of the Institutional Review Board.

3.1.2 Procedure

Participants played the role of master artisans; their task was to travel to 18 different villages and teach an apprentice how to produce necklaces that would sell well in each village. On each trial, participants were shown the motifs favored by a new village and were asked to create a single 10-bead necklace that contains all motifs favored by that village. Each village had a unique set of motifs, and villages were presented in a randomized order.

On each trial, participants saw the 3 motifs favored by the village displayed on the top of the screen, and they typed in a sample necklace by placing beads on an empty string in sequence. For example, if a village had the motifs 000, *10*, **111** teachers could pass on these motifs by producing the necklace 000*10***111***10* (emphasis added for demonstration). Participants could erase beads from the sequence to correct mistakes and press a "submit" button when they were finished with the sample necklace. To better align the task with our modeling assumptions, we constrained participants' responses so that they had to include all three motifs in their sample necklace; if the necklace they produced was not valid, participants were prompted to correct the necklace before proceeding.

3.1.3 Computational modeling

Model fitting and comparison: We compared models to participants' responses using random-effects Bayesian model selection (Rigoux, Stephan, Friston, & Daunizeau, 2014). First, we used maximum likelihood estimation to fit the α parameter of the pedagogical sampling model to each participant's responses. For each participant, we then evaluated the fit of the strong sampling and the pedagogical sampling models using the Bayesian information criterion ($\text{BIC} = klog(n) - 2log(L)$), where k is the number of free parameters in the model (0 for strong sampling and 1 for pedagogical sampling), n is the number of trials observed per participant (which was fixed at $n = 18$), and L is the maximized value of the probability of the data under the model. Finally, we used $-0.5 \times BIC$ as an estimate of log model evidence for each participant (Bishop & Nasrabadi, 2006) and used it to compute the protected exceedance probability (pxp) for each model. Protected exceedance probabilities treat models as random effects that can vary between participants; this measure can be interpreted as the probability that a given model occurs most frequently in the population.

Model simulation: In addition to the model comparison procedure described above, we also used fitted models to create simulated datasets. Intuitively, simulated datasets allow us to compare to what extent the necklaces produced by participants overlap with those that would be produced by each model. To create simulated datasets, we first used the fitted α parameters for each participant to create matrices of choice probabilities ($p_T(d|h)$) and then sampled from this matrix of choice probabilities to obtain simulated responses.

Measuring sequence complexity: We measured the algorithmic complexity of sample necklaces produced by human participants and in simulated datasets using the block decomposition method (Zenil, Toscano, & Gauvrit, 2022). We chose this measure of algorithmic complexity because of its theoretic connection with Kolmogorov complexity and algorithmic information theory (Chaitin, 1969; Kolmogorov, 1965; Solmonoff, 1964) and also because it captures well human subjective judgments of sequence randomness (Gauvrit, Singmann, Soler-Toscano, & Zenil, 2016; Gauvrit, Zenil, Delahaye, & Soler-Toscano, 2014; Planton et al., 2021). Intuitively, complexity scores provide an estimate of the length of the shortest program needed to recreate each necklace.

3.2 Results

Both human participants and the pedagogical sampling model tended to create sample necklaces that were simpler than those created by a strong sampling model (Figure 2A–B). We performed a linear mixed-effects regression that predicted the algorithmic complexity of sample necklaces based on fixed effects and random slopes of teacher type (human, pedagogical model, strong sampling model) and random intercepts by villages (18 different ground-truth sets of motifs). Both participants and the pedagogical sampling model created sample necklaces with lower algorithmic complexity than those produced by the strong sampling model (human-generated vs. strong-sampled necklaces: $b = -0.565, t(17.001) = -6.709, p < 0.001$; pedagogically-sampled

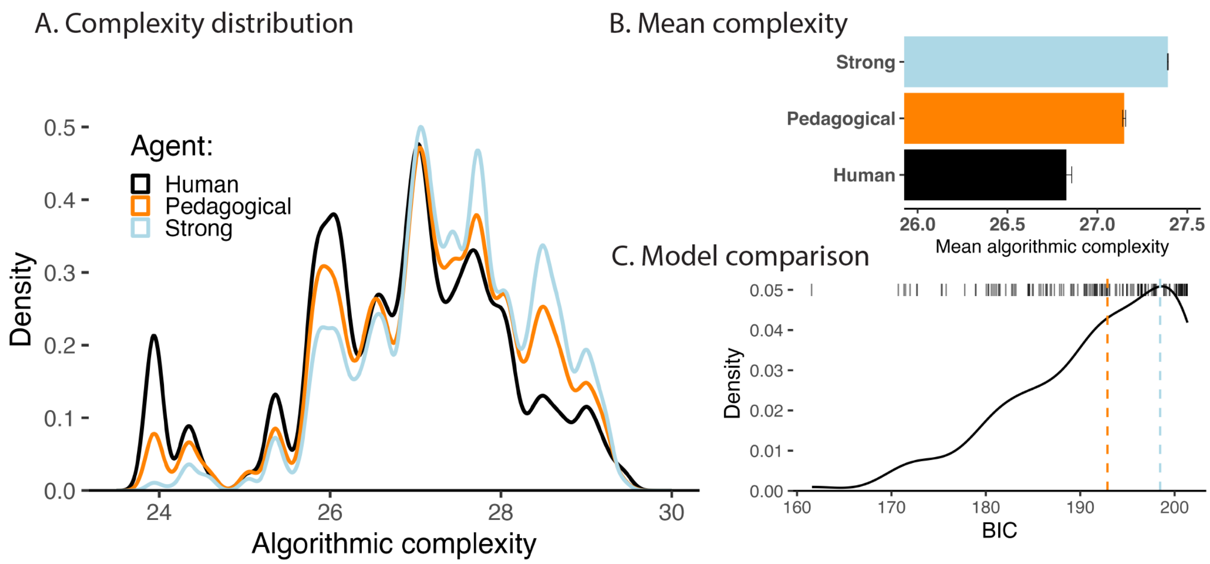

Figure 2: Experiment 1 results. A. Distribution of algorithmic complexity scores for human-generated necklaces (black line) and model-simulated necklaces (orange, blue lines). B. Average algorithmic complexity by teacher type. Human-created sample necklaces have the lowest mean complexity. Error bars denote standard error of the mean. C. Model comparison: Each tick-bar represents the fit between the pedagogical sampling model and a single participant's responses, as measured by the Bayesian information criterion (BIC). Orange dotted line indicates the mean model evidences for the pedagogical sampling model. The blue dotted line indicates the BIC of the strong-sampling model; because this model selects uniformly among all valid necklaces, the probability of each participant's responses under this model is a fixed value. Thus, points to the left of this blue line (lower BIC) are better fit by the pedagogical sampling model. Overall, participants' responses were better captured by the pedagogical sampling model, compared to the strong sampling model.

vs. strong-sampled necklaces: $b = -0.243, t(17.003) = -7.108, p < 0.001$). Accordingly, participants' choices were also better captured by the pedagogical sampling model (pxp = 1.000; see Figure 2C for model evidences). These results suggest that the pedagogical sampling model captures a specific pattern in how people teach generalizable abstractions: Namely, effective teaching favors simpler examples. Note that we did not explicitly instruct the model to favor simpler examples—instead, this preference stems from a more general communicative principle.

However, the pedagogical sampling model alone does not account for the full pattern of participants' responses. The examples generated by participants were still simpler than those simulated by the pedagogical sampling model (pedagogically-sampled vs. human examples:

$b = 0.322, t(16.997) = 5.648, p < 0.001$). Adding a penalty term for sequence complexity to the pedagogical sampling model did not close this gap (see Figure S1). These results suggest that participants' decisions may have been guided by additional inductive biases about what *kinds* of simple examples are useful; if this is the case, then learners might derive benefits from human-generated examples that are not fully captured by pedagogical sampling alone. In the following experiment, we tested whether learners can indeed recover motifs from a single sample necklace, and whether they learn best from examples generated by human teachers.

4 Experiment 2: Learning abstractions

Our results thus far suggest that reasoning about learners' mental states drives teachers to create simpler necklaces than would be expected by chance. However, it is an open question whether this simplicity aids learning—that is, whether the examples selected by teachers actually help others recover the underlying motifs. In our second experiment, we approached these questions by directly testing how well people learn triplets of motifs from observing a single sample necklace. Sample necklaces were selected from those generated by human teachers in Experiment 1 and from simulated datasets generated using the strong sampling and pedagogical sampling models.

4.1 Methods

4.1.1 Participants

We aimed to recruit 300 participants for the learning task on the Prolific platform using the standard sample option (`https://aspredicted.org/CK3_1NP`). We lost data from 5 participants due to server errors, leaving us with 295 participants for the learner task (M(SD) age = $40.02(12.31)$, 142 female, 149 male, and 4 non-binary). Participants were paid \$5 for completing the task, plus a bonus of up to \$1 contingent on performance.

4.1.2 Procedure

Participants were told that they were apprentices to a master artisan. Their task was to travel to 18 villages and learn how to produce necklaces that would sell well in each village. As in Experiment

1, participants were told that necklaces would only sell well in a particular village if and only if they contained all three motifs favored by that village. However, participants in Experiment 2 were not shown these motifs directly; instead, they saw a single sample necklace generated by a teacher and had to infer the motifs represented by the necklace.

On each trial, participants saw the sample necklace on the top of the screen and answered two questions about the village's motifs (Figure 3A). First, participants typed the motifs that they believed that the village favors. As described above (*Computational framework*), each motif was a sequence of 2 or 3 beads. Participants could erase the beads that they typed to correct mistakes, and submit the motifs when they were ready. Once they submitted the three motifs, participants could not change their responses. Next, participants were shown two empty necklaces and asked to create two new necklaces that would sell well in that village. Participants could only proceed if they produced two unique, length-10 necklaces that were distinct from the sample necklace. Villages were presented in a random order.

Participants completed three within-subjects conditions, which differed in how sample necklaces were generated. In the *Human* condition, participants saw sample necklaces created by participants in Experiment 1. By contrast, in the *Pedagogical* and *Strong* conditions, participants were shown sample necklaces that were simulated using the pedagogical- and strong-sampling models, respectively. (See Experiment 1 for more details on how model-simulated necklaces were generated.) Participants completed 18 trials total, comprising 6 trials for each teacher type. Participants were blind to condition; that is, they did not know what type of teacher generated each necklace. This experimental manipulation allows us to compare the effectiveness of human- and model-generated examples.

4.2 Results

We measured participants' performance using two outcome measures: The number of unique motifs correctly reported by the participant (*correct motifs*; range: 0–3, where higher scores indicate better performance) and the minimum number of changes that would have to be made to change participants' necklaces into a necklace that is consistent with the ground-truth motifs

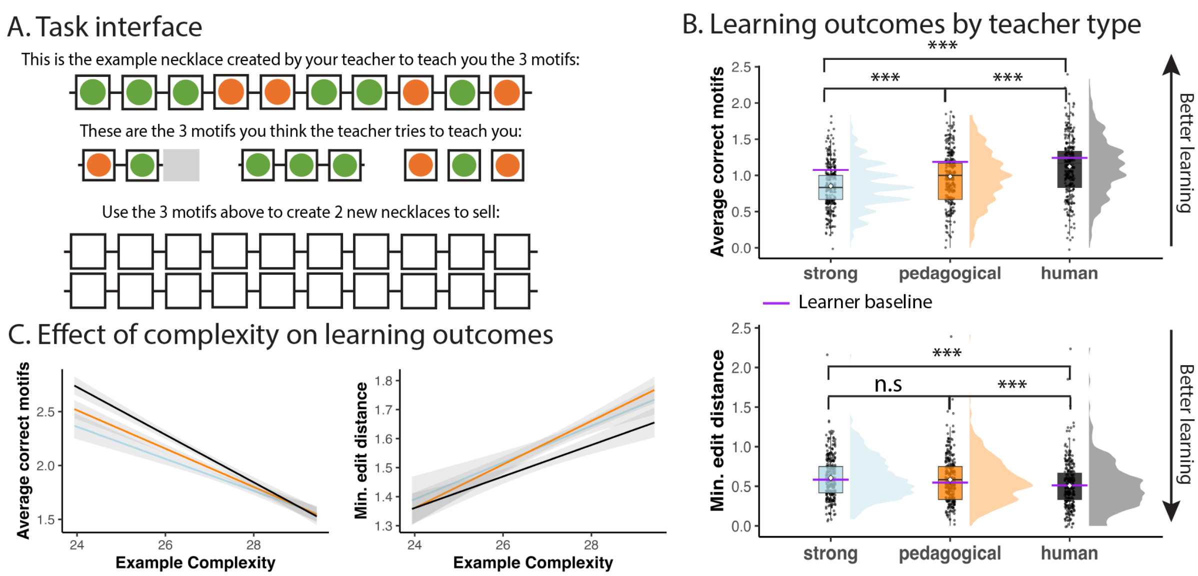

Figure 3: Experiment 2 results. A. Task interface: Participants saw a single sample necklace (top) that contained all three motifs favored by each village. They then explicitly reported the motifs (middle) and created two unique necklaces that were distinct from the sample necklace (bottom). B. Learning outcomes by teacher type: We measured learners' performance based on the number of motifs that they correctly reported (range 1–3; higher scores indicate better performance) and the minimum number of edits needed to transform learners' necklaces into a correct necklace (range 0–10; lower scores indicate better performance). Each point denotes a single participant's average performance; white diamonds denote average scores by teacher type. As a baseline, we compared these scores to the performance of a learner model that infers the underlying motifs by assuming that teachers select examples using strong sampling (purple line). C. Correlation between the algorithmic complexity of the sample necklace and learner performance, as indexed by the number of correct motifs (left) and edit distance (right). Shaded areas denote 95% confidence intervals. Participants performed better when they were shown simpler sample necklaces.

(*minimum edit distance*; range: 0–10, where lower scores indicate better performance). We modeled each outcome using a mixed-effects ordinal regression with fixed effects of teacher type (i.e., Human, Pedagogical, Strong, with Strong as the reference level) and random slopes of teacher type by village.

Overall, participants learned best when they received sample necklaces selected by human teachers, rather than necklaces generated by the pedagogical- and strong-sampling models (Figure 3B). Participants reported more correct motifs when the examples were generated by the pedagogical-sampling model (Pedagogical vs. Strong: $b = 0.349, z = 5.120, p < 0.001$) and by human participants from Experiment 1 (Human vs. Strong: $b = 0.702, z = 8.011, p < 0.001$), compared to the strong sampling model. However, participants recovered the most motifs overall when they received examples generated by a human (Human vs. Pedagogical: $b = 0.341, z = 4.230, p < 0.001$). Next, we found a similar pattern in the necklaces that learners generated. Participants generated necklaces that were closer to the target motifs (i.e., had lower minimum edit distances) when they received sample necklaces from a human teacher (Human vs. Strong: $b = -0.336, z = -4.249, p < 0.001$), but not necklaces generated by the pedagogical-sampling model (Pedagogical vs. Strong: $b = -0.091, z = -1.090, p = 0.276$). Overall, participants produced more accurate necklaces when they received examples from a human (Human vs. Pedagogical: $b = -0.244, z = -4.354, p < 0.001$).

On average, participants recovered approximately one of the three motifs specific to each village (mean(SE) number of motifs: 0.987(0.015)) and produced necklaces that were less than one bead away from an acceptable necklace (mean(SE) minimum edit distance: 0.566(0.013)) when provided a single sample necklace by a teacher. How well could participants be expected to do, given the sparse and ambiguous information given to them? We benchmarked participants' performance against the performance of the baseline learner model described above. We used this model in two ways. First, we sampled a triplet of motifs from $P_{baseline}(h|d)$ (Equation 4) to model how learners reported the motifs contained within each sample necklace. Second, the model samples uniformly from all necklaces that contain this triplet to create two new necklaces.

Overall, we find that the learner baseline captures qualitative patterns in learner performance. Like human learners, the baseline learner recovered approximately one of the three motifs specific to each village (mean(SE) number of motifs: 1.168(0.004)) and produced necklaces that were less than one bead away from an acceptable necklace (mean(SE) minimum edit distance: 0.547(0.006)). To compare quantitative fits, we compared participants' actual performance against this benchmark using paired Wilcoxon tests. Regardless of teacher type, participants reported slightly *fewer* correct motifs than the learner baseline (all $p < .05$ with the average difference of 0.178 motifs) and produced necklaces with *similar* minimum edit distances as the learner baseline (all $p > .2$). (We note one difference between the learner baseline and participants' behavior: In 30% of trials, participants reported motifs that were not consistent with the sample necklace they were provided. Thus, it is possible that participants may have been inattentive. In an exploratory analysis, we found that excluding these observations improved human learners' average performance slightly but did not affect our interpretation of the results; see SI.) Thus, while participants underperformed slightly relative to our baseline, they recovered about as much information as we could expect from a single example.

To understand what makes human-generated examples particularly effective, we next used mixed-effects ordinal regressions to model learners' performance based on an interaction between teacher type and the algorithmic complexity of the sample necklaces provided; we also included random effects of teacher type and algorithmic complexity by village. Participants who received more complex sample necklaces reported fewer correct motifs (effect on correct motifs: $b = -0.293, z = -3.699, p < 0.001$) and produced necklaces that were farther from correct necklaces (effect on minimum edit distance: $b = 0.269, z = 4.721, p < 0.001$).

Together, our results suggest that simplicity is beneficial: Participants indeed learned better when they were given simpler examples. However, participants still learned best when given examples by human teachers—even though they were not aware of our experimental manipulation—which suggests that existing models of teaching do not fully capture what makes human teaching so effective. In the Discussion, we will consider how to bridge this gap.

5 Discussion

Teaching useful and generalizable abstractions underlies cultural and technological achievements that require flexibility, innovation, and creativity. In this paper, we tested to what extent existing models of teaching capture how humans teach generalizable abstractions. Overall, we found that the general computational principles that underlie effective teaching and communication, as formalized by the pedagogical sampling model, also capture a specific pattern in how humans teach and acquire generalizable abstractions: Teachers favor simple examples, and learners learn best from simple examples, without explicitly building simplicity as an assumption into our models of either teachers or learners. However, our results also suggest that human teachers and learners are even more sensitive to simplicity than this model would predict, highlighting an exciting direction for future research.

Our results speak to prior theoretical debates on the importance of providing simpler data—or "starting small"—for effective learning. "Starting small" refers to a hypothesis that learning benefits from starting with simpler training data, both in humans and in artificial neural networks (Elman, 1993; Rafferty & Griffiths, 2010; Zhao, Lucas, & Bramley, 2024). Our findings reveal a similar phenomenon, where participants were better able to recover reusable abstractions when they received simpler examples. Moreover, our model reveals that one reason why simplicity is helpful is that it constrains the ways that learners can interpret the examples. However, the fact that participants also favored examples that were *even simpler* than pedagogical sampling alone would suggest that these models explain part but not all of what drives teachers to simplicity. We speculate that there are additional inductive biases favoring simple patterns that recursive Bayesian reasoning alone does not capture. For example, teachers may expect learners to parse necklaces directionally, as though they were reading a script; for example, if learners "read" necklaces from left to right, they may be more likely to interpret beads at the end of the necklace as overhang, rather than as part of a motif. In addition, our model assumes that learners can perfectly evaluate whether a necklace is consistent with a set of motifs. Human teachers may not share this assumption; instead, they may favor even simpler necklaces to guide fallible learners who may make mistakes when picking out motifs.

Overall, our findings provide a first demonstration of the usefulness of Bayesian pedagogy for understanding how humans transmit generalizable, culturally-specific abstractions. However, there are still many aspects of this domain that our simple task and model do not capture. Most notably, our work examines how learners recover abstractions from a single data point. While our findings show that learners can obtain some information even from this very sparse data, our task stands in stark contrast to how skills are taught outside of the lab. Novices do not learn how to knit a moss stitch or play musical chords from just a single example, but instead through repeated interactions where an expert provides opportunities for learners to observe skills, corrects their work, and sometimes provides explicit instruction (Kline, 2015). This additional structure has been argued to be essential for the stable transmission of complex skills (Caldwell, Renner, & Atkinson, 2018; Tehrani & Riede, 2008). In addition, while the motifs in our task were simple sequences that could be recombined in arbitrary ways, real-world cultural motifs are often imbued with meaning (Cohn, 2012; Hawkins, Sano, Goodman, & Fan, 2023; Long, Fan, Huey, Chai, & Frank, 2024). For example, skull motifs can remind viewers of the inevitability of death, and peonies appear frequently in Chinese art as a symbol of prosperity and wealth. It is an open question how teachers and learners coordinate on the meaning of motifs, or how these meanings constrain how motifs are deployed. Thus, more work is needed to extend existing theories of pedagogy to fully embrace the complexity of teaching generalizable abstractions.

Our work provides a theoretical and empirical framework to understand how teaching enables learners to flexibly act, create, and innovate. We hope that revealing further components of this ability will provide a fuller picture of how human intelligence is augmented by social learning and culture.

References

Aboody, R., Velez-Ginorio, J., Santos, L. R., & Jara-Ettinger, J. (2023). When naïve pedagogy breaks down: Adults rationally decide how to teach, but misrepresent learners' beliefs. *Cognitive Science*, *47*(3), e13257.

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 586–606.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020, Feb). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, *4*(2), 144–152. doi: 10.1038/s41562-019-0748-6

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340.

Caldwell, C. A., Renner, E., & Atkinson, M. (2018). Human teaching and cumulative cultural evolution. *Review of Philosophy and Psychology*, *9*(4), 751–770. doi: 10.1007/s13164-017-0346-3

Chaitin, G. J. (1969). On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM (JACM)*, *16*(1), 145–159.

Chen, A. M., Palacci, A., Vélez, N., Hawkins, R. D., & Gershman, S. J. (2024). A hierarchical bayesian model of adaptive teaching. *Cognitive Science*, *48*(7), e13477.

Clegg, J. M., & Legare, C. H. (2016a). A cross-cultural comparison of children's imitative flexibility. *Developmental Psychology*, *52*(9), 1435–1444.

Clegg, J. M., & Legare, C. H. (2016b). Instrumental and conventional interpretations of behavior are associated with distinct outcomes in early childhood. *Child Development*, *87*(2), 527–542.

Cohn, N. (2012). Explaining "I can't draw": Parallels between the structure and development of language and drawing. *Human Development*, *55*(4), 167–192.

Eaves, J., B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review*, *123*(6), 758–771. doi: 10.1037/rev0000031

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Gauvrit, N., Singmann, H., Soler-Toscano, F., & Zenil, H. (2016). Algorithmic complexity for psychology: a user-friendly implementation of the coding theorem method. *Behavior Research Methods*, *48*, 314–329.

Gauvrit, N., Zenil, H., Delahaye, J.-P., & Soler-Toscano, F. (2014). Algorithmic complexity for short binary strings applied to psychology: a primer. *Behavior Research Methods*, *46*, 732–744.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, *14*(1),

2199.
Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.
Kleiman-Weiner, M., Sosa, F., Thompson, B., van Opheusden, B., Griffiths, T. L., Gershman, S., & Cushman, F. (2020). Downloading culture. zip: Social learning by program induction. In *Proceedings of the 42nd annual meeting of the cognitive science society.*
Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences*, *38*, e31. doi: 10.1017/S0140525X14000090
Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*(1), 1–7.
Legare, C. H. (2019). The development of cumulative cultural learning. *Annual Review of Developmental Psychology*, *1*, 119–147.
Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191.
Pesowski, M. L., Quy, A. D., Lee, M., & Schachner, A. (2020). Children use inverse planning to detect social transmission in design of artifacts. In *Proceedings of the 42nd annual conference of the cognitive science society.*
Planton, S., van Kerkoerle, T., Abbih, L., Maheu, M., Meyniel, F., Sigman, M., … Dehaene, S. (2021, Jan). A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLoS Computational Biology*, *17*(1), e1008598. doi: 10.1371/journal.pcbi.1008598
Rafferty, A., & Griffiths, T. (2010). Optimal language learning: The importance of starting representative. In *Proceedings of the 32nd annual meeting of the cognitive science society.*
Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, *84*, 971–985.
Schachner, A., Brady, T., Oro, K., & Lee, M. (2018). Intuitive archeology: Detecting social transmission in the design of artifacts. In *Proceedings of the 40th annual meeting of the cognitive science society.*
Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55-89. doi: https://doi.org/10.1016/j.cogpsych.2013.12.004
Shafto, P., Wang, J., & Wang, P. (2021). Cooperative communication as belief transport. *Trends in Cognitive Sciences*, *25*(10), 826–828. doi: 10.1016/j.tics.2021.07.012
Solmonoff, R. (1964). A formal theory of inductive inference. Part I. *Information and Control*, *7*, 224–254.
Tehrani, J. J., & Riede, F. (2008). Towards an archaeology of pedagogy: Learning, teaching and the generation of material culture traditions. *World Archaeology*, *40*(3), 316–331.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.
Tominaga, A., Knoblich, G., & Sebanz, N. (2022). Expert pianists make specific exaggerations for teaching. *Scientific Reports*, *12*, 21296. doi: 10.1038/s41598-022-25711-3

Vélez, N., Chen, A. M., Burke, T., Cushman, F. A., & Gershman, S. J. (2023). Teachers recruit mentalizing regions to represent learners' beliefs. *Proceedings of the National Academy of Sciences*, *120*(22), e2215015120. doi: 10.1073/pnas.2215015120

Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, *33*, 17582–17593.

Zenil, H., Toscano, F. S., & Gauvrit, N. (2022). The block decomposition method. In *Methods and applications of algorithmic complexity: Beyond statistical lossless compression* (pp. 125–149). Springer. doi: 10.1007/978-3-662-64985-5_6

Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*(1), 125–136.

Supplemental Material

August 28, 2024

1 Building a complexity penalty into the pedagogical sampling model

In the main text, we report that teachers select examples that are even simpler than those selected by the pedagogical sampling model. As an exploratory analysis, we attempted to bridge this gap by extending the model with a penalty for more complex examples. We first obtained the learner's posterior beliefs from the pedagogical sampling model, as described in the main text:

$$P_{\text{learner}}(h|d) = \frac{P_{\text{teacher}}(d|h)p(h)}{\sum_{h'} P_{\text{teacher}}(d|h')p(h')}, \tag{1}$$

Next, we defined the utility of teaching hypothesis h (a triplet of motifs) with example d (a sample necklace) by combining the learner's posterior beliefs with the simplicity score:

$$U(d|h) = ln(P_{\text{learner}}(h|d)) - wC(d) \tag{2}$$

where the negative surprisal term, $ln(P_{\text{learner}}(h|d))$, captures the informational value of the example to the learner, and the complexity penalty, $C(d)$, is defined as the algorithmic complexity of d. w is the weight of the complexity penalty. Lastly, the utility score is converted into the probability of choosing d through a softmax function:

$$P_{\text{teacher}}(d|h) = \frac{e^{U(d|h)}}{\sum_{d'} e^{U(d'|h)}} \tag{3}$$

1.1 Results

We used the model estimation and model comparison procedures described in Experiment 1 of the main text. First, we used maximum likelihood estimation to fit the α and w parameters of the model. Next, we used Bayesian model selection to compare the fit of complexity penalty model (Figure S1, "with simplicity") to the original pedagogical sampling model (Figure S1, "without simplicity"). Even after directly penalizing complex examples, the original pedagogical sampling model best captures the behavior of human teachers (PXP = 1). These results suggest that people may not favor simpler examples for simplicity's sake;

instead, the decisions of human teachers may be guided by additional inductive biases that are not captured by existing theories. We return to this point in the General Discussion.

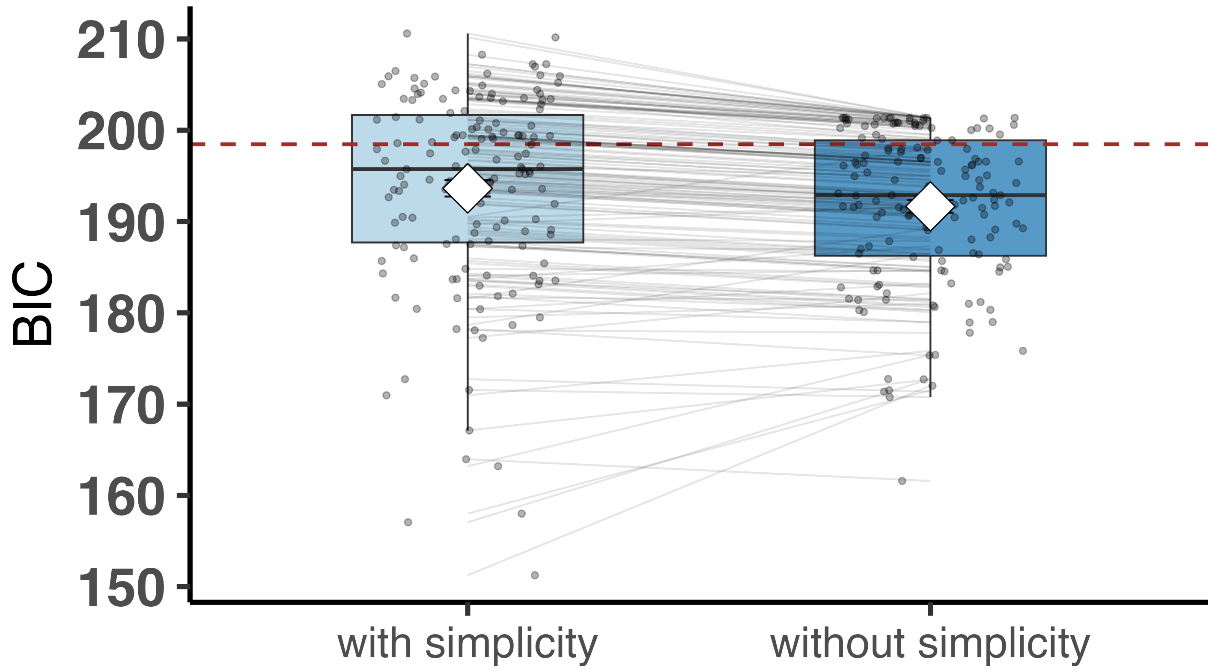

Figure S1: Bayesian information criterion (BIC): each dot represents the BIC of the model of each participant. The red dotted line indicates the BIC of the strong-sampling model. This shows overall, the original pedagogical model fit better (smaller BIC) than the pedagogical model that includes the simplicity score.

2 Stimuli

We chose 18 triplets of motifs for the 18 villages that participants visit in both the teacher task (Experiment 1) and learner task (Experiment 2). 9 villages favor triplets of motifs that contain 2 motifs with 2 beads and 1 motif with 3 beads, while the remaining villages favor triplets of motifs that contain 1 motif with 2 beads and 2 motifs with 3 beads. We avoided triplets of motifs that contain either all length-2 or length-3 motifs because the pedagogical sampling and strong sampling model do not make clearly distinguishable predictions for this subset of stimuli. The stimuli are: “10|010|011", “01|011|101", “01|10|010", “00|11|010", “00|01|110" ,“01|11|101", “10|11|110", “00|011|100", “11|010|101" ,“01|101|110",“00|10|011", “10|010|101", “11|001|011", “01|11|010", “01|10|101" , “00|100|101", “00|01|010", “01|010|101". “1” represents orange beads and “0” represents green beads.

3 Comparing learner performance to the baseline learner model after exclusions

In the main text, we note that human learners sometimes reported motifs that were inconsistent with the sample necklace they had received. It is possible that these trials reflect instances where participants were inattentive or made typing errors. Therefore, as an exploratory analysis, we also compared human learners' performance to the baseline learner model after excluding these trials. After excluding the 30% of the trials where the inferred motifs were not consistent with the example necklace, we overall saw a slight improvement in learners' performance compared to the learner baseline model. On average, participants recovered approximately one of the three motifs specific to each village (mean(SE) number of motifs: 1.091(0.015)) and produced necklaces that were less than one bead away from an acceptable necklace (mean(SE) minimum edit distance: 0.506(0.011)). The baseline learner recovered approximately one of the three motifs specific to each village (mean(SE) number of motifs: 1.163(0.007)) and produced necklaces that were less than one bead away from an acceptable necklace (mean(SE) minimum edit distance: 0.550(0.009)). After exclusion, participants still reported *fewer* correct motifs than the learner baseline if the examples were chosen by the models (all $p < .05$). However, if examples came from human teachers, participants reported a *similar amount* of correct motifs as the baseline learner model ($p = .80$). Participants also produced necklaces with *similar* minimum edit distances as the learner baseline if the examples came from the pedagogical sampling model ($p = .285$). However, if the examples came from the strong sampling model or human teachers, participants produced necklaces with *smaller* minimum edit distances than the learner baseline (all $p < .01$).