COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 37 (2013) 953–967

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12045

The Effects of Cultural Transmission Are Modulated by the Amount of Information Transmitted

Thomas L. Griffiths,[a,b] Stephan Lewandowsky,[b] Michael L. Kalish[c]

[a]*Department of Psychology, University of California*
[b]*School of Psychology, University of Western Australia*
[c]*Institute of Cognitive Science, University of Louisiana at Lafayette*

Received 5 March 2012; received in revised form 2 September 2012; accepted 1 October 2012

Abstract

Information changes as it is passed from person to person, with this process of cultural transmission allowing the minds of individuals to shape the information that they transmit. We present mathematical models of cultural transmission which predict that the amount of information passed from person to person should affect the rate at which that information changes. We tested this prediction using a function-learning task, in which people learn a functional relationship between two variables by observing the values of those variables. We varied the total number of observations and the number of those observations that take unique values. We found an effect of the number of observations, with functions transmitted using fewer observations changing form more quickly. We did not find an effect of the number of unique observations, suggesting that noise in perception or memory may have affected learning.

Keywords: Cultural transmission; Function learning; Bayesian modeling

An apocryphal story from World War I tells of a commander who conveyed an urgent message to his general by having each man speak it to his neighbor in the trench: "Send reinforcements. We are going to advance." The general was confused to receive the request that finally reached his ears: "Send three and sixpence. We are going to a dance." Information changes as it is passed from person to person, whether it is transmitted as a spoken message or by one person learning by observing the behavior of another. This process of cultural transmission provides the foundation for much of human knowledge: Most of the things we know we learn from other people, rather than by direct interaction

Correspondence should be sent to Tom Griffiths, Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, MC 1650, Berkeley, CA 94720-1650. E-mail: tom_griffiths@berkeley.edu

with our physical environment. As a consequence, understanding the factors that affect cultural transmission is important not just for preventing errors in the chain of command, but for understanding how the knowledge maintained by human societies changes over time.

Two basic questions about cultural transmission concern its effects and it speed: how it changes the information being transmitted and how quickly this process takes place. The first question is relevant to understanding how cultural objects such as languages, religious concepts, and social conventions are formed. Anthropologists have argued that since cultural transmission depends on cognitive processes such as learning and memory, we should expect these cultural objects to come to reflect the structure of the human minds that are involved in transmitting them (Atran, 2001; Boyer, 1998; Sperber, 1996). Support for this hypothesis comes from recent theoretical analyses showing that transmission of information along a sequence of Bayesian agents changes the information into a form that is consistent with the inductive biases of those agents (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007). Empirical results have borne out the predictions of this account, showing that as languages and concepts are transmitted along a sequence of human learners they take forms that are easier to learn (Griffiths, Christian, & Kalish, 2008; Kalish, Griffiths, & Lewandowsky, 2007; Kirby, Cornish, & Smith, 2008; Reali & Griffiths, 2009).

The second question, how quickly cultural transmission changes the information being transmitted, has been explored less extensively. This question has both practical and theoretical implications. On the practical side, identifying the factors that determine how quickly a message changes when it is passed from person to person has the potential to decrease misunderstandings of the kind experienced by the World War I general. On the theoretical side, knowing how quickly we expect languages and concepts to change over time would provide us with tools for answering questions such as whether enough time has passed for languages to have lost the influence of a common ancestor (Rafferty, Griffiths, & Klein, 2009) or how long ago two cultures diverged (Gray & Atkinson, 2003; Reali & Griffiths, 2010; Swadesh, 1952).

In this article, we analyze the impact of one factor that influences the rate at which cultural transmission has an effect: the amount of information transmitted between agents. We begin with a mathematical analysis of the simple case of transmission of a category defined on a single dimension. We then use a simulation to extend this analysis to the more complex case of transmission of a function, and we present an experiment exploring the predictions produced by this analysis.

2. Mathematical analysis of convergence rates

The Bayesian framework that has previously been used to analyze the consequences of cultural transmission can also be used to analyze the rate at which it converges to equilibrium. Assume that the information being transmitted between agents concerns a category defined along a single perceptual dimension. Although we focus on this case for the sake

of simplicity, there are real instances of cultural transmission that take this form, such as estimating the value of a specific formant in a phoneme. Each agent sees m samples $x_1, x_2, \ldots, x_m$ from a Gaussian distribution generated by the previous agent and seeks to estimate μ, the mean of the distribution. This is done by computing a *posterior* distribution over values of μ based on the observations and inductive biases expressed in a *prior* distribution. We will assume that the variance of the distribution of the x_i, σ_X^2, is known, and that all agents have the same prior distribution on μ, with mean μ_0 and variance σ_0^2.

Under these assumptions, standard results from Bayesian statistics can be used to show that the posterior distribution on μ inferred by agent n after observing the sample produced by agent $n-1$ will be Gaussian with mean μ_n and variance σ_n^2 given by the following:

$$\mu_n = \frac{\bar{x}_{n-1}/\sigma_X^2 + \mu_0/\sigma_0^2}{m/\sigma_X^2 + 1/\sigma_0^2} \tag{1}$$

$$\sigma_n^2 = \frac{1}{m/\sigma_X^2 + 1/\sigma_0^2}, \tag{2}$$

where $\bar{x}_{n-1}$ is the mean of the sample produced by agent $n-1$ (for details, see Gelman, Carlin, Stern, & Rubin, 1995). To return to the example of estimating the value of a formant in a phoneme, this indicates that the estimate should linearly interpolate between the mean of the prior (μ_0) and the average of the observed values ($\bar{s}_{n-1}$), assigning each a weight inversely proportional to its variance. As m increases, the variance of the posterior decreases—the sample provides more information about the value of μ.[1]

We can turn this into a model of cultural transmission by indicating how each agent generates the data seen by the next agent. If agent n generates m observations by sampling a value of μ from this distibution and then sampling the observations from the resulting Gaussian, the mean of these observations $\bar{x}_n$ is drawn from a Gaussian with mean μ_n and variance $\sigma_X^2/m + \sigma_n^2$. Using the results presented in Griffiths and Kalish (2007), this process defines a Markov chain that will converge to a stationary distribution that is also a Gaussian, with the mean of the observations $\bar{x}_n$ approaching a distribution with mean μ_0 and variance $\sigma_X^2/m + \sigma_0^2$ as n approaches infinity. In the case of estimating the value of a formant, this indicates that over time the distribution of the formant values produced by the agents will converge to a form that reflects their prior, as indicated by μ_0 and σ_0^2.

These results, together with the properties of Gaussian distributions, can be used to evaluate the distribution of the mean of the observations generated by the agent n, $\bar{x}_n$, conditioned on the mean of the sample used to initialize the sequence, $\bar{x}_0$. In the Appendix, we show that this distribution is Gaussian with mean $\mu_0 + c^n\bar{x}_0$ and variance $(\sigma_X^2/m + \sigma_0^2)(1 - c^{2n})$, where $c = 1/(1 + \frac{\sigma_X^2}{m\sigma_0^2})$. The mean and variance of $\bar{x}_n$ thus converge geometrically to the mean and variance of the stationary distribution as n increases. The rate of convergence is set by the constant c, being faster for smaller values of c. The value of c is determined by the ratio of σ_X^2 to $m\sigma_0^2$, being small when this ratio is large.

Increasing the sample size, m, increases c, and thus slows the rate of convergence. In other words, as the amount of information transmitted between agents increases, the rate at which cultural transmission changes that information decreases.

3. Testing the predictions with a function-learning task

The analysis presented in the previous section provides clear mathematical results, but it assumes a situation that is simpler than the tasks that have previously been used to test predictions about cultural transmission. We chose to conduct an empirical test of the prediction that the amount of information passed between people should determine the rate at which cultural transmission converges to an equilibrium using a function-learning task, based on previous research that has established that this is a case in which people have strong inductive biases that influence iterated learning (Kalish et al., 2007).

In function learning, each discrete trial involves presentation of a single magnitude of the stimulus variable (x), and the learner attempts to infer the underlying function relating y to x and produces an estimated magnitude $\hat{y}$ in response. Each response is followed by presentation of the correct value of y. Values of all variables are typically presented in graphical form. Tests of interpolation and extrapolation with novel x values after numerous (x, y) training trials confirm that people can infer continuous functions from these discrete trials. Previous experiments in function learning suggest that people have an inductive bias favoring linear functions with positive slope: Initial responses are consistent with such functions (Busemeyer, Byun, DeLosh, & McDaniel, 1997), and they require the least training to learn (Brehmer, 1971, 1974; Busemeyer et al., 1997). Accordingly, Kalish, Lewandowsky, and Kruschke, (2004) showed that a model that included such a bias could account for a variety of phenomena in human function learning. Finally, Kalish et al. (2007) showed that simulating cultural transmission of functions in the laboratory resulted in responses that converged on a linear function (with positive slope in 28 of 32 cases) irrespective of the information that was presented to the first generation.

The predictions for this case are similar to those seen in the mathematical analysis presented in the previous section. In general, as more information is passed from one person to another the rate of convergence decreases. Fig. 1 provides an example, produced from a simulation of a more complex Bayesian model described in detail in the Appendix. Increasing the amount of information each Bayesian agent provides to the next (again, expressed in terms of the variance of the posterior distribution) slows down convergence to the solution favored by the prior—in this case a linear function with a positive slope. However, learning functions introduces a factor that was not present in the simple one-dimensional case presented in the previous section: The amount of information a sample provides now depends both on the number of observations and the range of those observations.[2] As the range increases, the sample provides more information about the slope of the function. Intuitively, this is why it is a good idea to try to sample a wide range of values for the independent variable when conducting regression analyses. Since increasing the number of observations of x is likely to increase the range that those observations

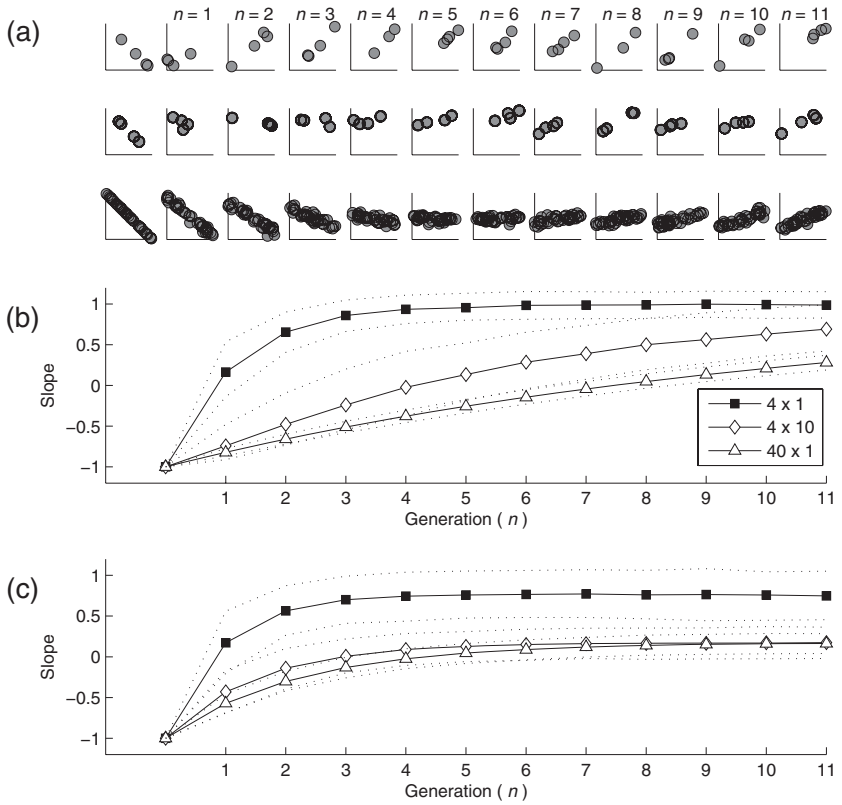

Fig. 1. Cultural transmission of linear functions by Bayesian agents. (a) Each panel shows the values of x and y for cultural transmission along one chain of agents. Each column shows the results of cultural transmission for one iteration, with the leftmost column showing the data provided to the first agent. Each agent observed data, computed a posterior distribution, sampled a hypothesized linear function from that distribution, and then generated the data provided to the next learner (see the Appendix for details). The rows show this process in three different conditions. The first row shows the 4 × 1 condition, where only four datapoints were observed or generated. The second row shows the 4 × 10 condition, where 40 datapoints were observed but these were generated by replicating four datapoints ten times (jitter is added to this plot to distinguish the points). The third row shows the 40 × 1 condition, where 40 unique datapoints were observed and generated. In each case, hypotheses converge on a linear function with positive slope, consistent with the inductive bias assumed for these Bayesian agents. However, the rate of convergence depends on the amount of information transmitted. (b) The mean slopes (solid line) and 68% confidence interval (dotted lines) for regression lines through the generated functions for 1,000 replications of the simulation presented in (a). The slope increases as a function of generation, with the rate determined by the amount information transmitted. (c) The difference between the 4 × 10 condition and the 40 × 1 condition disappears if observations of x and y include noise.

cover, the number of unique observations—that is, types rather than tokens—is also related to the rate of convergence.

Our experiment had three conditions, corresponding to the three situations illustrated in Fig. 1. In all conditions, participants in the first generation were trained on a negative

linear function, and subsequent generations of participants were trained on the responses of their immediate predecessors. The only difference between conditions was the training regime, which consisted of different combinations of the number of unique stimuli (types) and replications of each stimulus (tokens). In the 4 × 1 condition, training consisted of a single presentation of each of 4 unique stimuli. In the 4 × 10 condition, there were also 4 unique stimuli but each was presented 10 times. Finally, in the 40 × 1 condition, each of 40 unique stimuli was presented once. Training within each condition continued across generations until participants' responses had either flipped to a positive linear function (whereupon further changes are unlikely; Kalish et al., 2007) or a maximum of 11 generations had been trained within a condition.

As shown in the simulation results presented in Fig. 1, we expected the number of observations provided to learners to affect the rate of convergence of cultural transmission, with participants in the 4 × 1 condition converging faster than either the 4 × 10 condition or the 40 × 1 condition. Fig. 1 (b) also shows that it is possible for the number of unique observations to affect the rate of convergence, with a difference between the 4 × 10 and 40 × 1 conditions. However, this effect is weaker than the effect of the number of observations in two ways. First, the effect size is smaller, with the confidence intervals on the slopes overlapping. Second, the effect disappears if the observations are perceived with noise, since this effectively increases the range of the observations. Fig. 1 (c) shows that when x and y have noise associated with them (perhaps as a result of errors in perception or memory), there is little difference in the rate of convergence across conditions (see the Appendix for details). Consequently, whether we see an effect of the number of unique observations may depend on whether people can identify those observations as actually being unique.

3.1. Method

3.1.1. Participants

The participants were members of the campus community at the University of Western Australia ($N = 56$) and University of Louisiana at Lafayette ($N = 79$). Participants received remuneration ($10/h) or course credit for participation in the single experimental session.

Participants were randomly assigned to one of three experimental conditions and to a "family" within a condition (subject to the termination constraints below), with cultural transmission taking place across the generations of the family. There were five families in each condition, and participants were no longer added to a family after the 11th generation or after the responses of the latest participant clearly conformed to a positive linear function (assessed by the slope of the test responses), whichever came first.

3.1.2. Stimuli and apparatus

A Windows computer running a Matlab program designed using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) was used to present stimuli and to record responses.

On each trial, a gray filled horizontal bar (approximately 2 cm high) was presented at the top-left of the screen. The upper left corner of the bar was approximately 4 cm from the top and 4 cm from the left of the edge of the screen, and the horizontal extent of the bar indicated the magnitude (x) of the stimulus. No tick marks or scales were present.

The participant entered a response magnitude ($\hat{y}$) by clicking on a vertically oriented slider, which consisted of a thin (0.8 cm) unfilled rectangle that abutted the right side of the screen and was labeled "0" and "max," respectively, at the bottom and top ends. No other scale marks were present. The mouse was originally positioned to the left of the center of the slider, and people indicated their response by clicking within the slider. Upon clicking, a black horizontal bar appeared at that location within the slider and a confirmation button (labeled "OK") appeared to its left. Participants could adjust their response repeatedly and clicked the "OK" button to proceed.

During the training phase, a response was immediately followed by feedback, which consisted of the word "correct" printed within a frame connected to the slider with a line at the vertical location that corresponded to the correct target value y. The feedback remained visible for a minimum of 1.6 s, with the duration being extended in linear proportion to the response error (i.e., the difference between the response magnitude and the true magnitude) to encourage accurate responding.

3.1.3. Procedure

The experiment involved a training phase followed by a test phase. The stimuli used in the training phase varied by condition: There were 4 unique stimuli in the 4×1 and 4×10 conditions (repeated 10 times in the latter) and 40 unique stimuli in the 40×1 condition. For learners who formed the first generation of any family, all unique stimuli were randomly sampled from the set of stimulus magnitudes (x) in [1,100]. Target magnitudes (y) were assigned according to the negative linear function $y = 100 - x$, allowing us to monitor the rate at which responses moved away from this function and toward a positive linear function.

The test phase always involved 40 test trials, irrespective of condition. In the 4×1 and 4×10 conditions, the test phase involved all four unique training stimuli plus 36 new stimuli with x values sampled uniformly from [1,100]. In the 40×1 condition, 20 of the training stimuli were used as test stimuli, together with 20 new stimuli. Test stimuli were presented in random order.

For generations following the first, training stimuli were sampled from the test phase responses of the participant in the previous generation of that family. For the 40×1 condition, the training set for a given generation was simply the test set of the previous generation. For the other two conditions, the training set consisted of two of the test stimuli that had x values drawn from the training set of the previous participant and two of the test stimuli that had new x values. In all cases, the magnitude estimates $\hat{y}$ provided during the test phase by the previous participant became the new target values y.

The sequence of training trials consisted of a random permutation of all unique stimuli and their replications (i.e., $4 \times 1 = 4$, $4 \times 10 = 40$, and $40 \times 1 = 40$ training trials). Training trials were separated by a 1 s blank screen. Test trials were identical to training

trials, except that no feedback was presented after the response was entered. Participants were informed about this change at the outset.

The experiment was preceded by four practice trials during which feedback was presented. All practice trials involved the pairing ($x = 50$, $y = 50$) for all conditions. The constant values prevented possible biasing toward any particular function relating x and y during training.

4. Results and discussion

Owing to the brevity of training in the 4 × 1 condition, analysis focused on responses from the test phase. Fig. 2 shows the responses for each participant in all three conditions. For each condition, one row of panels corresponds to a family, whereas columns correspond to generations. Thus, the participants in the left-most column all received stimuli that were sampled—using the regime determined by the condition—from the same negative linear function. All remaining participants in each condition were trained on stimuli that were contingent upon the responses of the preceding generation.

Of greatest interest in Fig. 2 is the evolution of responses across intergenerational transmission, from left to right across columns in each row. It is immediately apparent that in the 4 × 10 condition (panel (b)) and in the 40 × 1 condition (panel (c)), there were three families who failed to converge across 11 generations; that is, the last descendant in each family continued to respond according to the negative linear function that was used for the first generation. The remaining two families in each condition converged after eight and four generations (40 × 1), and after seven and seven generations (4 × 10). In striking contrast, *all* families converged in the 4 × 1 condition, namely after ten, five, one, six, and three generations (panel (a)). The rapid switches in slope across successive generations are consistent with the results of previous iterated learning experiments using a function-learning task (Kalish et al., 2007) and are consistent with having a multimodal prior distributon on functions rather than the smooth prior assumed in the Bayesian linear regression model we used to motivate our experiment (Griffiths, Lucas, Williams, & Kalish, 2009).

A summary of the data is provided in Fig. 3, which shows the cross-generational evolution of average slopes (i.e., best-fitting slope estimates for each subject averaged across generational peers in all families) in the three conditions.[3] The figure makes two important points. First, it clarifies that in all conditions there was movement away from the initial function to the positive linear alternative, consistent with the results of Kalish et al. (2007). Second, the figure highlights that convergence was faster in the 4 × 1 condition than in the other two, which in turn did not differ much from each other.

For statistical confirmation, we first fit a regression model to the data in Fig. 3 that had separate slopes and intercepts for each condition. This model fit very well, $R^2 = .937$, and the loss of fit was negligible when the slopes for the 4 × 10 and 40 × 1 conditions were constrained to be equal, $F(1,27) = .012$, $p > .10$. When the slope for the 4 × 1 condition was also constrained to be identical, the further loss of fit was considerable,

Fig. 2. Responses from each participant from the test phase in all three conditions. Each row of panels represents a family and each column a generation of participants. Intergenerational transmission ceased after 11 generations or once convergence to a positive function had occurred.

$F(1,28) = 10.43$, $p < .005$, confirming the obvious pattern in the figure and the faster convergence of the 4 × 1 condition (slope 0.145) than the other two (joint slope 0.089).

We also performed a Bayesian analysis in which the point at which the slope switched to a positive value was treated as a Poisson random variable with a different rate parameter for each condition.[4] We used a generic conjugate prior—an exponential distribution with unit mean—to obtain posterior distributions on the rate of the Poisson that were

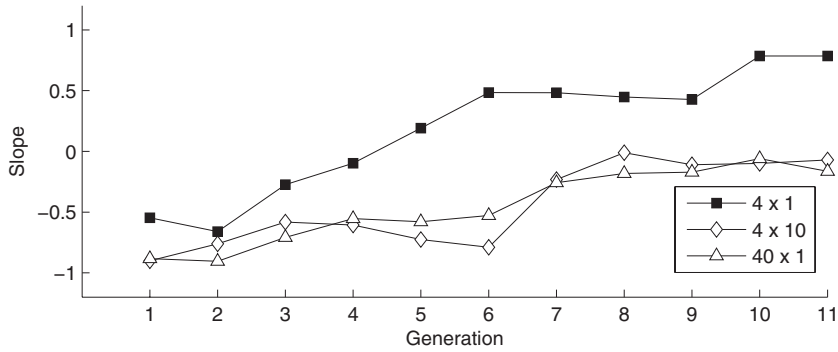

Fig. 3. Evolution of average slopes of linear regressions fit to the test phase responses of each participant.

Gamma(25,6), Gamma(48,6), Gamma(45,6), for the 4 × 1, 4 × 10, and 40 × 1 conditions, respectively. This results in posterior probabilities of .995 and .992 that the rate was higher in the 4 × 1 condition than the 4 × 10 and 40 × 1 conditions, and .582 that the rate was higher in 4 × 10 than 40 × 1. The Bayesian analysis thus supports the same conclusions as the regression model.

5. Conclusions

Mathematical analyses of cultural transmission by Bayesian agents predict that the rate at which information is changed by cultural transmission is inversely related to the amount of information that is transmitted. Our results partially bore out these predictions. In confirmation of predictions, convergence to a function that reflected people's inductive biases was faster when the function was transmitted using fewer observations (the 4 × 1 condition). The number of unique observations within the sample (40 × 1 vs. 4 × 10) did not have a statistically significant effect. These results are consistent with the conclusion that the amount of information transmitted between learners affects the rate of convergence of cultural transmission, but they reinforce the fact that the information provided by a sample depends on how people perceive it.

The lack of the predicted effect of the number of unique observations is an interesting finding that warrants further investigation. One possibility is that the sample size used in our study was simply not large enough to find this effect. If so, our results suggest that the effect of the number of unique observations must be weaker than the effect of sample size, which is consistent with the predictions produced by the simple model considered in the introduction (see Fig. 1 (b)). A second possibility is that human learners inserted sufficient noise into the observations to mask the fact that the number of unique observations was small. If perceptual or memory error was sufficient to "jitter" these observations into a sample that more closely resembled that seen in the 40 × 1 condition, we should not expect to see a difference between the conditions. As shown in Fig. 1 (c), such noise can remove the difference between the 4 × 10 and 40 × 1 conditions.

Our results have implications for both practical and theoretical questions related to cultural transmission. On the practical side, they illustrate how the amount of information passed between agents plays a crucial role in the ultimate fidelity of transmission. Our advice to the apocryphal World War I commander would be to tell his troops to repeat the message several times (i.e., instantiate a 4×10 rather than 4×1 condition), thereby increasing the probability that it would be successfully transmitted. On the theoretical side, the relationship between sample size and rate of convergence has the potential to deepen our understanding of which aspects of languages we might expect to change more rapidly as they are passed from generation to generation, providing a link to analyses that examine the relationship between the frequencies of linguistic constructions and the rate at which those constructions change over time (e.g., Reali & Griffiths, 2010).

Acknowledgments

Preparation of this article was facilitated by grant number BCS-0704034 from the National Science Foundation to TLG, a Discovery Grant from the Australian Research Council and an Australian Professorial Fellowship to SL, and grant number BCS-0544705 from the National Science Foundation to MLK. We thank Leo Roberts for assistance during data collection and Charles Hanich for assistance with data analysis.

Notes

1. The variance of the posterior is a direct proxy for the amount of information provided by the observations in the sense of Shannon (1948) since the differential entropy of a Gaussian is $\log(\sigma\sqrt{2\pi e})$, a monotonic function of the variance σ^2.
2. More precisely, the variance of the posterior distribution on the slope depends on the norm of the vector of values of x, as indicated in the Appendix.
3. For construction of this figure and the associated regression analysis, the final set of responses for any family that converged on the positive linear function was used in the computation of the mean for all subsequent generations up to the maximum of 11.
4. Chains that never switched were treated as switching at the last iteration, which is a conservative estimate for the purpose of this analysis.

References

Atran, S. (2001). The trouble with memes: Inferences versus imitation incultural creation. *Human Nature*, *12*, 351–381.

Boyer, P. (1998). Cognitive tracks of cultural inheritance: How evolved intuitive ontology governs cultural transmission. *American Anthropologist*, *100*, 876–889.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*, 259–260.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Decision Processes*, *11*, 1–27.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Gray, R. D., & Atkinson, Q. D. (2003). Language–tree divergence times support the Anatolian theory of Indo–European origin. *Nature*, *426*, 435–439.

Griffiths, T. L., & Kalish, M. L. (2007). A Bayesian view of language evolution by iterated learning. *Cognitive Science*, *31*, 441–480.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*, 68–107.

Griffiths, T. L., Lucas, C., Williams, J. J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, vol. *21* (pp. 553–560). Red Hook, NY: Curran Associates.

Kalish, M., Lewandowsky, S., & Kruschke, J. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*, 1072–1099.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, *14*, 288–294.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*, 5241–5245.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences USA*, *105*, 10681–10686.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Rafferty, A., Griffiths, T. L., & Klein, D. (2009). Convergence bounds for language evolution by iterated learning. In Niels Taatgen and Hedderik van Rijn (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society, (pp. 2451–2456). Cognitive Science Society: Austin, TX.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society, Series B*, *277*, 429–436.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Oxford, England: Blackwell.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, *96*, 452–463.

Appendix: Mathematical details

Convergence rate for transmission of the mean of a Gaussian

We can derive the distribution of $\bar{x}_n$ given $\bar{x}_0$ by observing that $\bar{x}_n$ can be seen as a linear function of $\bar{x}_{n-1}$, plus a new random variable ϵ_n drawn from a Gaussian. Without loss of generality, we can assume that the mean of the prior on μ is $\mu_0 = 0$ (if necessary, we can subtract μ_0 from all $\bar{x}_n$ so that this assumption holds), so that μ_n is simply a multiple of $\bar{x}_{n-1}$. This means that we can write

$$\bar{x}_n = c\bar{x}_{n-1} + \epsilon_n \tag{3}$$

where $c = 1/(1 + \frac{\sigma_X^2}{m\sigma_0^2})$ and $\epsilon_n \sim \text{Gaussian}(0, \sigma_X^2/m + \sigma_n^2)$. If $c = 1$, this is just a Gaussian random walk. If $c < 1$, it is a random walk with a tendency to shrink toward 0. We can recursively apply Equation 1 to obtain a specification of $\bar{x}_n$ in terms of $\bar{x}_0$ and a series of random variables

$$\bar{x}_n = c^n\bar{x}_0 + \sum_{j=1}^{n} c^{(n-j)}\epsilon_j. \tag{4}$$

This makes it easy to evaluate the expectation of x_n,

$$E[\bar{x}_n] = c^n\bar{x}_0 + \sum_{j=1}^{n} c^{(n-j)}E[\epsilon_j] \tag{5}$$

$$= c^n\bar{x}_0 \tag{6}$$

and its variance,

$$\text{var}[\bar{x}_n] = \sum_{j=1}^{n} c^{2(n-j)}\text{var}[\epsilon_j] \tag{7}$$

$$= (\sigma_X^2/m + \sigma_n^2)\sum_{j=0}^{(n-1)} c^{2j} \tag{8}$$

$$= (\sigma_X^2/m + \sigma_n^2)\frac{1 - c^{2n}}{1 - c^2} \tag{9}$$

$$= (\sigma_X^2/m + \sigma_0^2)(1 - c^{2n}), \tag{10}$$

where the fraction is the expression for an incomplete sum of a geometric series. Since all ε terms have Gaussian distributions, their sum has a Gaussian distribution. Thus, we have

$$\bar{x}_n|\bar{x}_0 \sim \text{Gaussian}(c^n x_0, (\sigma_X^2/m + \sigma_0^2)(1 - c^{2n})) \tag{11}$$

where $c = 1/(1 + \frac{\sigma_X^2}{m\sigma_0^2})$. Allowing μ_0 to take values other than 0, we obtain the result given in the article.

Function learning and Bayesian linear regression

The results shown in Fig. 1 were generated by simulating cultural transmission using the Bayesian linear regression model presented by Kalish et al. (2007). We briefly reproduce the information about that model here.

Assume that the agent is presented with data d consisting of a set of n pairs (x_i, y_i), and that the agent seeks to estimate a linear function of the form $y = \beta_1 x + \beta_0 + \epsilon$, where ε is Gaussian noise with variance σ_Y^2. We can summarize both the data and the estimated function using column vectors, $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$, and $\boldsymbol{\beta} = [\beta_1\ \beta_0]^T$. The prior used in Bayesian estimation is a distribution over the parameters $\boldsymbol{\beta}$, $p(\boldsymbol{\beta})$. We take $p(\boldsymbol{\beta})$ to be Gaussian with mean $\boldsymbol{\mu}_\beta$ and covariance matrix $\sigma_\beta^2 \mathbf{I}_2$.

The posterior distribution is a distribution over $\boldsymbol{\beta}$ given $\mathbf{x}$ and $\mathbf{y}$. Using our choice of prior and the assumption of Gaussian noise in $\mathbf{y}$, the posterior is Gaussian with covariance matrix

$$\boldsymbol{\Sigma}_{\text{post}} = \left(\frac{1}{\sigma_Y^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbf{I}_2\right)^{-1}, \tag{12}$$

and mean

$$\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}}^{-1}\left(\frac{1}{\sigma_Y^2}\mathbf{X}^T\mathbf{y} + \frac{1}{\sigma_\beta^2}\boldsymbol{\mu}_\beta\right). \tag{13}$$

This completes the specification of the model we used to simulate cultural transmission.

The analysis of cultural transmission by Bayesian agents given by Griffiths and Kalish (2007) indicates that if we repeatedly sample a hypothesis from this posterior distribution and then generate data by sampling from the corresponding likelihood function, over time the hypotheses considered by the learners will converge to the prior distribution. To explore how the amount of data seen by the learners influences the rate of convergence, we simulated this process for three conditions, corresponding to the conditions used in our experiment. In the 4 $\times$ 1 condition, four datapoints were sampled at each generation, while in the

4×10 four datapoints were sampled, but then replicated 10 times to make a total of 40 datapoints, and in the 40×1 condition 40 unique datapoints were sampled. In each case, x was chosen from a uniform distribution on [0,1]. For the first generation, y was set to $1 - x$. For subsequent generations, a value of $\boldsymbol{\beta}$ was sampled from the resulting posterior distribution and used to generate values of y for the new randomly drawn values of x, which were then supplied as data to the next learner. This process was continued for a total of eleven learners, producing the results shown in Fig. 1 (a) and (b). The likelihood and prior assumed by the learners had $\sigma_Y^2 = 0.0025$, $\sigma_\beta^2 = 0.01$, and $\boldsymbol{\mu}_\beta = [1\ 0]^T$, corresponding to a strong prior favoring functions with a slope of 1 and an intercept of 0. The results shown in Fig. 1 (c) were produced by repeating this simulation but adding Gaussian noise with a mean of zero and standard deviation of 0.2 to all x and y values.