When extremists win: On the behavior of iterated learning chains when priors are heterogeneous

Daniel J. Navarro (d.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide

Arthur Kary (art.kary@gmail.com)
School of Psychology, University of New South Wales

Scott Brown (scott.brown@newcastle.edu.au)
School of Psychology, University of Newcastle

Chris Donkin (christopher.donkin@gmail.com)
School of Psychology, University of New South Wales

Abstract

How does the process of information transmission affect the cultural products that emerge from that process? This question is often studied experimentally and computationally via iterated learning, in which participants learn from previous participants in a chain. Much research in this area builds on mathematical analyses suggesting that iterated learning chains converge to people's priors. We present three simulation studies suggesting that when the population of learners is heterogeneous, the behavior of the chain is systematically distorted by the learners with the most extreme biases. We discuss implications for the use of iterated learning as a methodological tool and for the processes that might have shaped cultural products in the real world.

Keywords: Iterated learning; language evolution; cultural evolution; inductive biases; Bayesian cognition

Which aspects of our language or culture are shaped by the inductive biases possessed by people, and which aspects are shaped by the process of transmission from one learner to the next? A key framework for thinking about and disentangling these factors is known as *iterated learning*, shown schematically in Figure 1. Iterated learning is a particular kind of cultural transmission in which behavior arises in one individual (or generation) by learning from the observations of the previous person (generation), forming a chain of learners.

An appealing characteristic of iterated learning is that the behavior of iterated learning chains can be characterized mathematically: under certain assumptions, iterated learning chains with Bayesian learners will converge to a distribution that depends on the learners' priors and the size of the bottleneck (Griffiths & Kalish, 2007; Rafferty, Griffiths, & Klein, 2014). These results have allowed researchers to explore inductive biases in different tasks, including function learning (Kalish, Griffiths, & Lewandowsky, 2007), visual working memory (Lew & Vul, 2015), reasoning about everyday events (Lewandowsky, Griffiths, & Kalish, 2009), and category learning (Canini, Griffiths, Vanpaemel, & Kalish, 2014). They have been especially useful in studying language evolution (Kirby, Griffiths, & Smith, 2014).

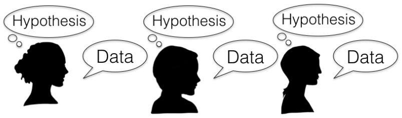


Figure 1: Schematic illustration of a typical iterated learning paradigm, which assumes that learner n learns on the basis of the data provided by learner $n-1$.

Importantly, the theoretical proofs about how iterated learning chains converge depend critically on the assumptions made. For example, if learners select the hypothesis with the highest posterior probability rather than sample from their posterior, an iterated learning chain will tend to exaggerate the prior (Kirby, Dowman, & Griffiths, 2007). Similarly, we use language to talk about things and events in the world. If one changes the mathematical assumptions to reflect this insight, then the stationary distribution of the chain more closely resembles the posterior distribution (Perfors & Navarro, 2014). In this paper we consider the role played by individual differences. Such differences are robustly observed in many areas of cognition, yet theoretical results typically assume that all learners share the same biases.

When individual differences exist, what should we expect to observe? One possibility is that the chain converges to a distribution that reflects the "average prior belief" in some sense. For instance, if 10% strongly believe in hypothesis A and 90% of people strongly believe in hypothesis B, one might hope that an iterated learning chain reflects 10% A and 90% B hypotheses. Alternatively, perhaps the chain will produce some other reasonable compromise between A and B that weights each learner in equal proportion. Our findings indicate that neither of these situations necessarily occurs: if people do not share the same priors, iterated learning is not guaranteed to converge to the prior in any meaningful sense. Instead, the distribution to which it does converge is disproportionately influenced by the most biased learners. We illustrate this using three simulation studies.

Case study 1: Language evolution

Do all learners have equal influence on the process of language evolution? Consider the pressures on a language to incorporate a particular grammatical rule or not. Some learners may have STRONG opinions about a particular rule or construction, whereas others might have WEAK opinions. Exactly who has which might might vary with the particular linguistic context and construction involved: for instance, children may to have a bias for regularization that adults do not share (Hudson Kam & Newport, 2005), but adult second-language learners may have biases based on transfer from their first language while children do not (Ellis, 2015). We are fairly agnostic at this point about what such biases might be; all that matters for the present purposes is that it is plausible that there are individual differences in at least some language learning biases. Our question is what effect this might have on the nature of the evolved language.

To study this, consider the following experimental design. Participants are presented with sentences in an artificial language that may incorporate a construction (e.g., pluralization rule, morphological marking, etc). After training, participants are asked to produce new sentences, which are presented as the input to the next learner in the chain. This is a relatively typical design, and a simple Bayesian model for this learning problem can be constructed as follows.

If θ denotes the probability that the grammatical rule should be followed, a Bayesian learner specifies a prior distribution $P(\theta)$. For simplicity we assume a Beta(a,b) distribution in which $P(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. In our simulations we assume that some learners enter with a STRONG bias about the grammatical rule, formalized via a Beta(1,10) prior. In contrast, a WEAK learner might have the opposite bias, but not a strong one, which can be formalized with a Beta(2,1) prior. Regardless of the biases the learner possesses, it is assumed that belief updating follows Bayes' rule. After a training session in which x of n sentences follow the rule, the posterior distribution $P(\theta \mid x)$ is

$$P(\theta \mid x) \propto P(x \mid \theta) P(\theta) \tag{1}$$

where $P(x \mid \theta) \propto \theta^x (1-\theta)^{n-x}$ is the probability of observing x out of n rule-consistent cases if the true probability is θ. Under these assumptions, the posterior over θ is a Beta$(a+x, b+n-x)$ distribution. When asked to generate a novel sentence, a Bayesian learner might sample a value of θ from their posterior, and their output satisfies the rule with probability θ. The number of rule-consistent sentences y generated by the learner is thus sampled from the posterior predictive distribution $P(y|x)$:

$$P(y|x) = \int_0^1 P(y|\theta)P(\theta|x)d\theta \tag{2}$$

This kind of model is often used to study regularization in iterated learning designs (Ferdinand, Thompson, Kirby, & Smith, 2013; Reali & Griffiths, 2009).

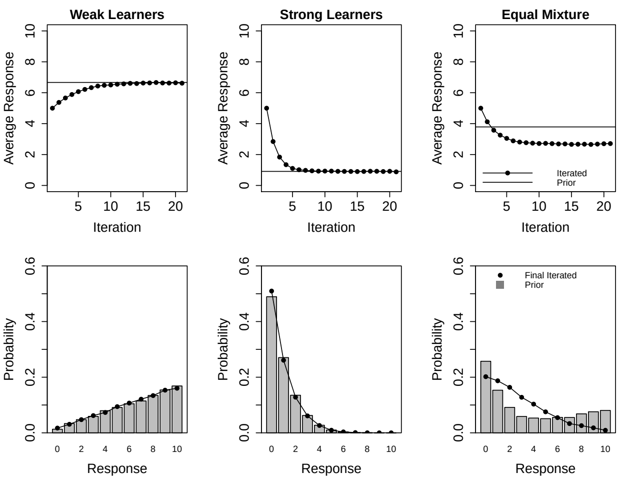
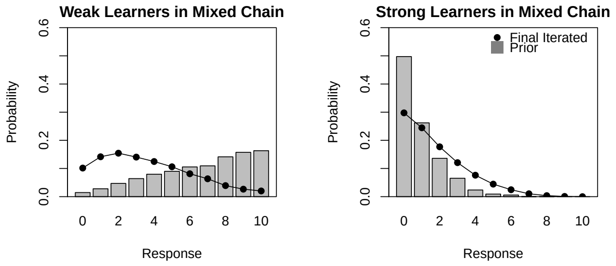


Figure 2: Simulating an iterated learning investigation of language evolution. When the learners all share the same bias (left and middle columns) the average proportion of responses converges to the prior mean (top row), and the distribution of responses converges to the prior distribution (bottom row). When the chain is a mixture of STRONG and WEAK learners, the average proportion of responses does *not* correspond to the average prior expectation, nor does the distribution converge to the average prior in the population.

Simulation

We simulate the results of three different kinds of iterated learning experiments. In all cases, the first person is taught ten sentences in an artificial language, five consistent with a grammatical rule; they then generate ten sentences used as input to the next learner. In the first experiment all learners have a STRONG bias about the rule, and in the second experiment all of them have a WEAK bias in the opposite direction. In the third experiment, half of the learners have STRONG biases and half have WEAK opposing ones. In each case results are aggregated across 100,000 simulated iterated learning chains.

The results are shown in Figure 2. As predicted by previous work, in both of the homogeneous cases iterated learning experiment transparently reveals the learner biases: the chain converges to the prior. However, when we consider the iterated learning experiment conducted with a mixed population (right panels of Figure 2) we observe a strikingly different result. In this situation – where half of the learners are STRONG and half are WEAK – the average bias in the population is to expect 38% of sentences to be rule-consistent. Yet, as the top right panel shows, the iterated learning chain converges to a smaller number, with only 27% of responses following the rule. More importantly, as the bottom right panel reveals, the distribution of responses bears very little resemblance to the underlying population biases. One might have hoped that, when learners bring different priors to an iterated learning experiment, the chain would converge to a weighted average of their priors. In this case, this weighted average would be a 50-50 mixture of the priors of STRONG learners and WEAK learners (plotted as a histogram). As the figure illustrates, the iterated learning chain (lines) does not converge to anything even remotely similar to this mixture distribution.


Weak Learners in Mixed Chain

Strong Learners in Mixed Chain

Figure 3: Distribution of responses in a mixed chain plotted as a function of the type of learner generating the response.

Discussion

Why does the iterated learning procedure behave this way when the population is heterogeneous? The answer can be found by separating the responses on the last iteration by learner type, shown in Figure 3. As is clear from inspection, the WEAK bias learners (left) are greatly influenced by the STRONG bias learners: their responses are rule-consistent 36% of the time, rather than 67% as one might expect given their Beta(2,1) prior, and the distribution of responses (lines) deviates markedly from their prior (histogram). The opposite effect occurs too (right panel), but it is much smaller: the STRONG bias learners increase the proportion of rule-consistent responses from the 9% rate implied by the Beta(1,10) prior to 17.5% in the iterated learning chain. Similarly, their distribution of responses is not markedly different from their prior.

As this example illustrates, when individual differences exist an iterated learning procedure is not guaranteed to reveal the inductive biases of the learner. The STRONG learners apply a strong inductive bias, and these learners require a lot of evidence before they are willing (or able) to apply the grammatical rule in question. As a consequence, data generated by a WEAK learner will have minimal ability to sway such a person. The reverse does not hold: the WEAK learners in this scenario are very responsive to external input. As a result, a WEAK bias participant makes a much larger adjustment from the prior than does a STRONG bias one, with the consequence that the overall behavior of the mixed chain is much more heavily driven by the group with the strongest bias.

Case study 2: Group decision making

Groups of people often arrive at beliefs that seem to lack any evidentiary basis, famously described by the "groupthink" phenomenon (Janis, 1982). How do these false beliefs arise? Do they necessarily reflect a bias shared by all reasoners, or can an entire community be misled by a small number of highly biased learners?

To examine this question, we consider a scenario in which a jury of 12 people begin their deliberations with a straw poll. A notepad is passed around the room, with each person writing down whether they would decide in favor of the plaintiff before removing their sheet of paper and passing the pad to the next juror. Unfortunately, each juror can read the indentations left by the previous one, forming an iterated learning chain. A Bayesian juror might reason about this by considering two hypotheses, namely that the evidence favors the plaintiff ($e = 1$) or the defendant ($e = 0$). The trial evidence sets the juror's prior belief that $P(e = 1) = \theta$, which is updated when the vote v of the preceding juror is revealed. The juror unconsciously assigns a reliability value r to this information, such that $P(v = 1|e = 1) = P(v = 0|e = 0) = r$. If the preceding juror voted for the plaintiff, the juror's posterior degree of belief that the verdict should favor the plaintiff becomes

$$P(e = 1 \,|\, v = 1) = \frac{r\theta}{r\theta + (1-r)(1-\theta)} \tag{3}$$

and the posteriors are calculated similarly when the earlier vote favored the defendant. For simplicity, we assume that jurors generate their vote probabilistically by sampling from the posterior.

As these equations illustrate, when $r = 0.5$ the current juror completely ignores the vote provided by the previous one and the posterior probability is identical to the prior. This arises naturally when the current juror is confident that their existing beliefs incorporate all relevant information about the case, and as such the opinions of other jurors can have no influence upon their own beliefs. We refer to such a juror as a GOAT – someone who forms their own view and is not led to conclusions by the opinions of others. In contrast, suppose the juror is underconfident or unsure about their beliefs, perhaps suspecting that other jurors have access to different information. Such a juror will set $r > 0.5$, because they attribute evidentiary value to the opinions of others. We refer to this kind of a juror as a SHEEP because they are more likely to adjust their vote to agree with the votes of others.

Simulations with homogeneous chains

We consider three scenarios. In the first scenario all jurors are GOATS who set $r = 0.5$ and have a modest opinion in favor of the defendant ($\theta = 0.4$). In the second scenario all jurors are SHEEP who set $r = 0.95$ and have a modest opinion favoring the plaintiff ($\theta = 0.6$). Finally we consider a situation where half of the jurors are SHEEP and the other half are GOATS. To illustrate what happens in these situations we simulated each scenario 100,000 times. The results are plotted in Figure 4. Not surprisingly, because the GOAT jurors ignore the input and generate responses directly from their own prior beliefs, the "chain" starts at their prior (on average, 40% of jurors vote for the plaintiff) and the total number of votes in favor of the plaintiff follows a binomial distribution.

What should we expect to see if all jurors are SHEEP? One reading of the literature suggests that, since iterated learning chains of Bayesian learners converge to the prior, and since the first SHEEP samples from their own prior, we should see a result not dissimilar to the one we see for GOATS. That is – while we might expect to see non-independence among successive jurors – we should find that on average a SHEEP juror should vote for the plaintiff 60% of the time, in accordance with their priors. However, as the middle column of

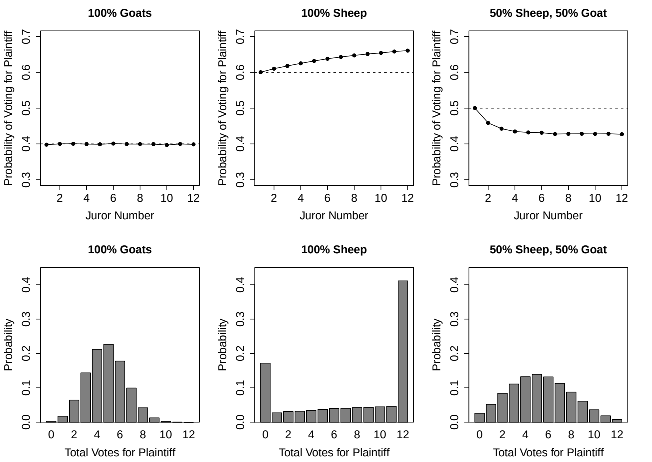


Figure 4: The jury straw poll. The top row plots the probability that each juror votes for the plaintiff, as a function of their position in the chain (the dashed line plots the population average prior), and the bottom row plots the distribution of votes for the plaintiff. The left and middle plots show juries composed entirely of GOATS and SHEEP respectively. The plots on the right depict a scenario when 50% of jurors are SHEEP and 50% are GOATS.

Figure 4 illustrates, this is not what happens. The first juror votes in accordance with their priors, but by the time the 12th juror is polled, the probability of voting for the plaintiff has risen to 67%. Moreover, it is simple to prove that this reflects the true stationary distribution of the chain. To see this, let $p = P(v_i = 1|v_{i-1} = 0)$ denote the probability that the i^{th} juror in the chain votes for the plaintiff given that the previous juror voted for the defendant, and similarly let $d = P(v_i = 0|v_{i-1} = 1)$ denote the probability that the i^{th} juror switches the other direction. The transition matrix for the strawpoll is thus

$$\boldsymbol{T} = \begin{bmatrix} 1-p & p \\ d & 1-d \end{bmatrix} \tag{4}$$

A chain with this transition matrix converges to a stationary distribution $\boldsymbol{\pi}$ in which the (marginal) probability of voting for the defendant and plaintiff is proportional to d and p respectively. To verify this, note that

$$\begin{aligned} \boldsymbol{\pi T} &\propto [d,p] \begin{bmatrix} 1-p & p \\ d & 1-d \end{bmatrix} \\ &= [d(1-p)+pd, dp+p(1-d)] \\ &= [d,p] \propto \boldsymbol{\pi} \end{aligned} \tag{5}$$

For a SHEEP juror, the probability of switching the vote from the plaintiff to the defendant is $d = (.1 \times .4)/(.1 \times .4 + .9 \times .6) = .069$, and similarly the probability of switching the vote towards the plaintiff is $p = (.1 \times .6)/(.1 \times .6 + .9 \times .4) = .142$. In the long run, a chain of SHEEP converges to a 67% probability of voting for the plaintiff even though each individual SHEEP only assigns a 60% prior probability to the plaintiff.

On the surface, the SHEEP result seems at odds with the convergence proof in Kalish et al. (2007) - Bayesian learners sampling from their posterior do not (in this instance) converge to the prior. To that end, it is useful to note that the SHEEP chain violates the assumptions of the original proof, because the SHEEP jurors use the wrong likelihood function for the learning problem. The SHEEP juror assigns evidentiary value to the opinions of other jurors when they should not, because all jurors have seen the same facts at trial. This miscalibration creates the "groupthink" behavior: the SHEEP jurors "double count" the evidence, and the iterated learning chain exaggerates their prior bias.

Simulations with mixed chains

Now consider what happens when SHEEP and GOATS are mixed together in equal proportions (Figure 4, right). The SHEEP assign prior probability of 0.6 to the plaintiff, whereas the GOATS assign prior 0.4, so the population average prior is 0.5. Alternatively, if we consider the behavior of the two homogeneous iterated learning chains, the SHEEP on their own would be expected to converge to 0.67 and the GOATS would converge to 0.4, so the average of these two long run probabilities is 0.54. If one did not know the detail of the models, it would be reasonable to expect a mixed chain to produce an average probability of voting for the plaintiff somewhere between 50% and 54%. Unsurprisingly, it does nothing of the sort. Because GOATS are insensitive to the opinions of others and SHEEP are highly sensitive, the GOATS dominate the mixed chain, and the long run behavior converges to a 43% probability of voting for the plaintiff. That is, the SHEEP "learn" to mimic GOATS but the GOATS make no such accommodation.

Discussion

The implications of the jury scenario are twofold. First, the SHEEP-only chain illustrates that it is possible for an iterated learning chain to exaggerate biases even when Bayesian learners sample hypotheses from the posterior. The result complements an earlier result by Perfors and Navarro (2014), which showed that the convergence of iterated learning chains is affected when there is an additional input to the chain (i.e., the world passes new information to learners). In the SHEEP chain we find that convergence is even influenced when learners mistakenly *believe* there is additional information being passed into the chain. This miscalibration drives a kind of groupthink, in which a collection of individually underconfident learners becomes overconfident as a group.

Second, the behavior of a heterogenous chain is not easily predicted by considering the behavior of the corresponding homogeneous chains, or the priors of individual learners. The mixed chain of SHEEP and GOATS is mostly driven by the GOATS, even though a homogenous chain of GOATS produces a much less extreme outcome than the a chain of pure SHEEP. The reason for this is obvious when we consider the decision making strategies used by the two learner types, but we rarely have access to such information in real life.

Case study 3: Categorization

Our third case study considers a categorization problem with non-Bayesian learners. We consider stimuli that vary along

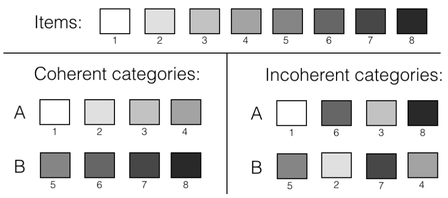


Figure 5: Categorization with eight items that vary along one dimension (top panel). Items can be organized into categories that are coherent (left panel) or incoherent (right panel).

a single dimension, with 8 exemplars spaced evenly across the range (i.e., at $x = 1, \ldots, 8$): an example is shown at the top of Figure 5. Each stimulus can be assigned to one of two categories (A or B), and we are interested in the inductive biases that people bring to this categorization problem.

An iterated learning design can be used to explore these biases. During category learning, each learner is shown training items that consist of four exemplars and their category labels, selected randomly subject to the constraint that there must be one exemplar of each category in the training set. During the test phase the learner must classify the remaining four exemplars. An iterated learning chain is constructed by using a random subset of responses from one learner as the training data for the next, again subject to the constraint that the learner must be shown at least one example of each category.

In our simulations we assume each participant applies the Generalized Context Model (GCM: Nosofsky, 1986). In the GCM, the probability of assigning a test item located at y to category A, given training items $\boldsymbol{x} = (x_1, \ldots, x_n)$ with labels $\boldsymbol{l} = (l_1, \ldots, l_n)$ is proportional to the summed similarities between y and the category A exemplars:

$$P(y \in A \,|\, \boldsymbol{x}, \boldsymbol{l}) = \frac{\sum_{i|l_i=A} S(x_i, y)}{\sum_{i|l_i=A} S(x_i, y) + \sum_{i|l_i=B} S(x_i, y)} \qquad (6)$$

where similarity decays exponentially with distance, $S(x, y) = \exp(-\lambda|x - y|)$. This model has one free parameter: the *specificity* parameter λ that describes how rapidly similarity decays. When λ is large, similarity falls away very quickly with distance, and when λ is small it diminishes more slowly.

Category coherence bias

Although not framed as a Bayesian model, the GCM imposes biases on how learners categorize, and these biases depend on λ. For instance, the GCM prefers "coherent" categories that assign similar items to the same category. A simple measure of "coherence" counts the number of times that adjacent items are assigned to the same category: the categories on the left of Figure 5 have maximal coherence of six, whereas the incoherent categories on the right have coherence zero. To investigate GCM biases, we simulated the iterated learning experiment described above 100,000 times using different values of λ, assuming that all learners in a chain have the same λ. The results (Figure 6, left) show that the GCM bias for coherent categories is strongest for large values of λ.

Given that individual differences in categorization exist, we ran a second simulation study (Figure 6, right). This time we mixed learners that varied in their λ values (sampling uniformly at random from 0.1, 1 and 10) into a single chain to investigate the effect heterogeneity has on each learner type. Unlike our previous simulations, the heterogeneity of the chain did not distort any of the three GCM learner types to a large extent: the right hand side of Figure 6 is not too dissimilar to the left. Based on this, one might conclude that the heterogeneity of the population has done very little to distort the categorization schemes produced by the various different learners. Unfortunately, this conclusion is unwarranted.

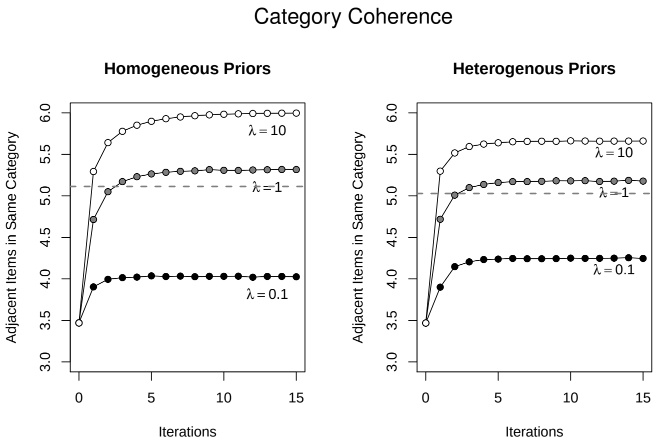


Figure 6: Exploring the "category coherence" bias using iterated learning. The y axis plots category coherence (defined in main text). **Left panel**: Category coherence assuming all participants share the same prior (λ). Here there are three chains each reflecting one of the three λ values. As λ grows higher, iterated learning produces more coherent categories. The grey dashed line reflects the average of the three chains on iteration 15. **Right panel**: When there are individual differences within participants, the learners all become somewhat more similar to one another but the effect is small.

Category size bias

Categorization is complex, and even this simple problem involves multiple biases. A preference for coherent categories is one kind of bias that a learner might express, but one might be just as interested in exploring the extent to which learners prefer categories to be of similar size. Does the GCM have a bias to split items evenly or unevenly? Does it depend on λ?

To that end, we counted the number of exemplars assigned to the smaller category in our previous simulations. Figure 7 plots this for the three homogeneous chains (left) and the single heterogeneous chain (right). The left panel shows that the GCM has a bias to prefer unevenly sized categories: this bias is weak when the learner generalizes narrowly ($\lambda = 10$), and strong when the learner generalizes widely ($\lambda = 0.1$). Unfortunately, almost none of this differentiation is evident when we look at the heterogeneous chains: the average response is substantially different from when the three learner types were taken separately, and there are almost no individual differences to be found, with all three learner types producing similar responses. With respect to the category size bias, mixing different learners into the iterated learning chain has almost completely erased their differences.

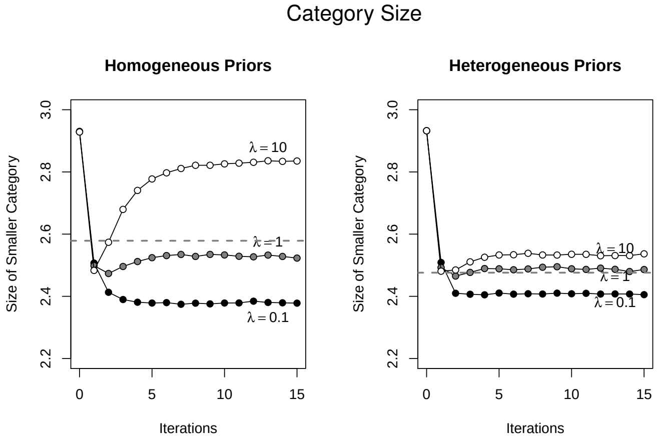


Figure 7: Exploring the "category size" bias using iterated learning. The *y* axis plots the number of items assigned to the smaller category. **Left panel**: Homogenous iterated learning chains when all learners use the same value of λ. The three plots in the figure are quite dissimilar: when λ is small the GCM strongly prefers an unequal allocation of items to categories, but when λ is large the preference is weak. The grey dashed line reflects the average of the three chains on iteration 15. **Right panel**: When the same GCM learners are mixed into a heterogenous iterated learning chain, most of this variation is suppressed (the curves are close to each other), and the average size of the smaller category (grey dashed line) has substantially decreased.

General discussion

The three case studies all display the same pattern. When all learners bring the same inductive bias to the problem, iterated learning behaves in the way that previous theoretical proofs suggest it should (Griffiths & Kalish, 2007). In particular, when learners are Bayesians with identical priors and correctly specified likelihoods, iterated learning reveals those priors. For a non-Bayesian learner an analogous inductive bias is uncovered. However, when learners bring different biases to the problem there is no guarantee that the responses of any one participant genuinely reflects their prior biases, nor is there any guarantee that the average responses reflect the average bias in the population. To the contrary, our case studies suggest that those learners with the most extreme biases exert a disproportionate influence on the chain. We briefly consider the implications if this pattern holds more generally.

Iterated learning leads a double life within the psychological literature. As a theoretical tool, the underlying dynamics of the chain provide valuable insights into how cultural and linguistic evolution works. From that perspective, our results open up new questions: for instance, does language evolution reflect the cognitive biases of all speakers, or do some subpopulations (e.g., children) exert stronger influences on the process? Similarly, learners with the most confidence in their own beliefs will exert a disproportionate influence on others, providing a justification for expressing overconfidence: if the goal is to have cultural influence rather than be correct, strong biases are better than weak ones. Regardless, the effect of heterogeneity in this context need not be a reason for concern so much as a reason to ask new questions.

On the methodological side, iterated learning has often been used as a tool for exploring the inductive biases of individuals. Based on formal results suggesting that the stationary distribution of an iterated learning chain is the prior, researchers in cognitive science have sometimes used these designs as a form of elicitation task, in which the (between-subject) distribution of responses is taken to be reflective of some (within-subject) latent mental representation of the world. In this context, our results suggest that some care is required. When people bring different priors to a task, there is no inherent reason to think that the stationary distribution of an iterated learning chain reveals those priors. The distortions are both systematic and difficult to predict. The latter point is especially troublesome from a methodological perspective. In our third case study, it was not obvious to us that heterogeneity among category learners would produce a large distortion of "category size" biases, but almost no distortion to the bias for "coherent" categories. In this context, we suggest that the interpretation of iterated learning experiments is difficult when individual differences exist.

References

Canini, K., Griffiths, T., Vanpaemel, W., & Kalish, M. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*.

Ellis, R. (2015). *Understanding second language acquisition* (2nd). Oxford University Press.

Ferdinand, V., Thompson, B., Kirby, S., & Smith, K. (2013). Regularization behavior in a non-linguistic domain. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Griffiths, T. & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Hudson Kam, C. & Newport, E. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Janis, I. L. (1982). *Groupthink: psychological studies of policy decisions and fiascoes*. Houghton Mifflin Boston.

Kalish, M., Griffiths, T., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kirby, S., Dowman, M., & Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28C*(108-114).

Lew, T. & Vul, E. (2015). Structured priors in visual working memory revealed through iterated learning. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society, Austin, TX. Cognitive Science Society*.

Lewandowsky, S., Griffiths, T., & Kalish, M. (2009). The wisdom of individuals: exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969–998.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Perfors, A. & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, *38*(4), 775–793.

Rafferty, A., Griffiths, T., & Klein, D. (2014). Analyzing the rate at which languages lose the influence of a common ancestor. *Cognitive Science*, *38*, 1406–1431.

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