Inferring processes of cultural transmission: the critical role of rare variants in distinguishing neutrality from novelty biases (supplementary material)

James P. O'Dwyer[1] and Anne Kandler[2]

[1]Department of Plant Biology, University of Illinois, Urbana IL 61801 USA
[2]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

S1 Neutral Theory in Cultural Evolution

Neutral theory in cultural evolution has been mainly modelled using the Wright- Fisher infinitely many allele model (see e.g. [1] for a review of the mathematical properties, [2] for its introduction to cultural evolution as well as [e.g. 3, 4, 5, 6] for further applications to cultural case studies). The theory assumes that in finite populations cultural variants are chosen to be copied according to their relative frequency, and new variants not previously seen in the populations are introduced by a process resembling random mutation. Changes in frequency therefore occur only as a result of drift. While these inherent assumptions are likely to be violated in the cultural context (for detailed discussions see [e.g. 2, 3, 6]) population-level patterns of various observed episodes of cultural change nevertheless resemble the ones expected under neutrality [e.g. 2, 5, 7]. Importantly, theses studies do *not* conclude that neutral evolution is the underlying evolutionary force shaping the observed empirical patterns. They rather suggest that if each act of choosing one cultural variant rather than another has a different motivation, the emerging population-level patterns will be that there are no directional selective forces affecting what is copied [8]. However, it still has to be shown that neutral predictions are distinguishable from predictions generated by alternative cultural selection scenarios [see e.g. 9, for a discussion of this issue in the ecological context]. If a (potentially unknown) number of cultural scenarios result in very similar predictions, then the meaning of the rejection of the neutral hypothesis becomes hard to interpret.

In the following we provide a brief overview over the characteristics of the Wright- Fisher infinitely many allele model. This model assumes that the composition of the population of instances of cultural variants at time

t is derived by sampling with replacement from the population of instances at time $t-1$ resulting in non-overlapping generations. The (temporally constant) population size J and the variables m_i and n_i stand for the abundances of variant i in the population at times $t-1$ and t, respectively. Then

$$p_i = \frac{m_i}{J}(1-\mu),\ i = 1, 2, \ldots$$

describes the probability that a specific instance is of variant i. Further, μ denotes the innovation rate which describes the probability that a novel variant, not currently or previously seen in the population, is introduced. In general, the probability that the configuration of abundances $[m_1, m_2, \ldots]$ at time $t-1$ is transformed into $[n_0, n_1, n_2, \ldots]$ at time t is given by

$$P(X_0(t) = n_0, X_1(t) = n_1, \ldots | X_1(t-1) = m_1, \ldots) = \frac{J!}{\prod_i m_i!} \prod_i p_i^{n_i} \tag{S1}$$

with $p_0 = \mu$ and $\sum_i m_i = \sum_i n_i = J$. The state space of the Markov process defined by these transition probabilities is extremely large making the derivation of population-level properties of this stochastic process almost intractable. But Eq. (S1) implies that the extinction of any variant is inevitable over time and the time evolution of a single variant can be described by a two-variant formulation

$$P(X_i(t) = n_i | X_i(t-1) = m_i) = \binom{J}{n_i} p_i^{n_i} (1-p_i)^{J-n_i}. \tag{S2}$$

We note that under neutrality all variants are considered identical and therefore we can drop the index i from Eq. (S2).

It follows from the Eq. (S2) that the probability that a newly introduced variant with abundance 1 goes immediately extinct is given by

$$P(X(t) = 0 | X(t-1) = 1) = \left(1 - \frac{1}{J}(1-\mu)\right)^J \to e^{\mu - 1} \text{ for large } J.$$

Further, the diffusion approximation of Eq.(S2) allows us to determine the transition probability density $f(x, p, \tau)$ as the solution of the diffusion equation

$$\frac{\partial f(x,p,\tau)}{\partial \tau} = a(p)\frac{\partial f(x,p,\tau)}{\partial p} + \frac{1}{2}b(p)\frac{\partial^2 f(x,p,\tau)}{\partial p^2}$$

with $a(p) = -J\mu p$, $b(p) = p(1-p)$ and appropriately scaled space and time dimensions $p = m/J$, $x = n/J$ and $\tau = t/J$ [e.g. 10]. In general, an explicit solutions of this equation can only be achieved under relatively restrictive assumptions, [e.g for $\mu = 0$, 11]. Nevertheless, it has been shown that some steady-state properties of the population of instances of cultural variants can be determined. The variant abundance distribution describing the expected number of variants with relative frequencies in the interval $(x, x + \delta x)$ at steady state can be approximated by

$$\phi(x) = \theta_c x^{-1} (1-x)^{\theta_c - 1} \tag{S3}$$

with $\theta_c = 2J\mu$ [12]. Additionally, the average number of different variants, S, in the populations can be described by

$$E\{S\} = \theta_c + \int_{1/J}^{1} \theta_c x^{-1}(1-x)^{\theta_c - 1} dx$$

(e.g. [1]).

We note that the variant abundance distributions given by Eq. (2.3) in the main text and Eq. (S3) generate similar results for sufficiently large J and sufficiently small ν.

S1.1 Simulation of the Wright-Fisher model

Simulations of the infinitely many allele Wright-Fisher model are relatively easily obtained through random sampling from previous generations. In detail, in each time step t a new set of J instances is generated through random copying from the population of instances of cultural variants at time step $t-1$ possessing the abundance configuration $[m_1, m_2, \ldots, m_{S(t-1)}]$ with $\sum_{i=1}^{S_{t-1}} m_i = J$. The variable S_{t-1} stands for the number of different variants at time step $t-1$ and m_i records their abundance. The probability that variant i is copied in each of the J production events is given by $p_i = \frac{m_i}{J}(1-\mu)$ where μ stands to the innovation rate. If an innovation occurs then a variant, not currently or previously seen in the population, is introduced.

After a burn in period which ensures that the system has reached an approximate steady state we determine the progeny distribution after $T = 200,000$ generation for $J = 1,000$ and various values of μ (see dotted lines in Fig. S1). We lack an analytical result for the cumulative Wright-Fisher progeny distribution, but drawing from our results for the overlapping generations neutral model we plot a power law with exponent $-1/2$ (red line) (we showed in the main text that for intermediate values of k the progeny distribution resembles a power law with exponent $-3/2$). As μ becomes small, we can see that this power law with fixed exponent becomes an increasingly accurate explanation of the first phase of this distribution, just as in the case of overlapping generations. It is likely that fitting a single power law to the whole distribution, including the exponential decline, would explain the apparent variation in power law exponent with μ and J identified in earlier studies.

S2 NZS Solutions for Species Dynamics and Species Abundance Distribution

The non-zero sum (NZS) formulation of neutral theory is an approximation to a neutral, overlapping generations model where all variants compete for a single resource, and the strength of competitive interactions is

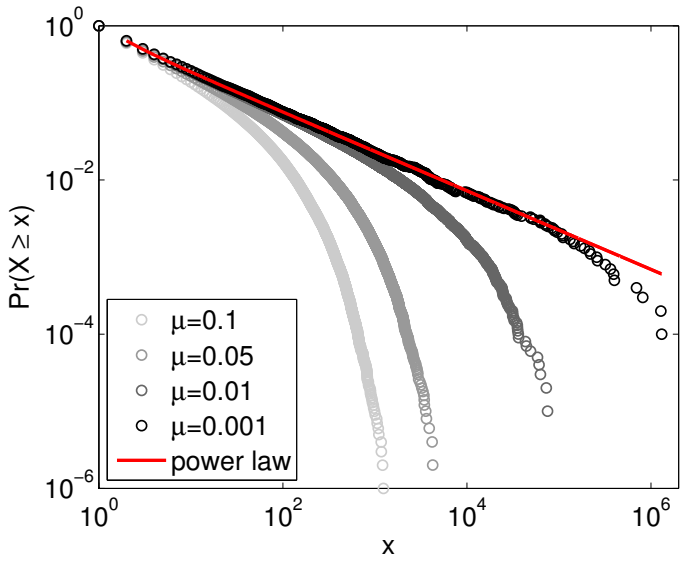

Figure S1: Progeny distribution determined by the Wright-Fisher infinitely-many alleles model for $J = 1000$, $T = 200,000$ and $\mu = 0.001$ (black circles), 0.01 (dark gray circles), 0.05 (gray circles), 0.1 (light gray circles).

equal across all pairs of variants. The defining master equation focuses on the dynamics of one focal variant, and characterizes its change in abundance through time, from an initial condition (usually taken to be $n = 1$, and known as point speciation in the ecology literature)

$$\frac{dP}{dt} = b(n-1)P(n-1|t) - bnP(n|t) - dnP(n|t) + d(n+1)P(n+1|t). \tag{S4}$$

This master equation is linear because the interactions between the focal variant and the rest of the population are treated in a mean field approximation. In effect, this equation assumes that the remainder of the population is of constant size, and then the pairwise competitive interactions are approximated by just adding to the mortality rate for this variant.

To solve Eq. (S4) for $P(n|t)$, we use the generating function $G(z,t)$ defined by

$$G(z,t) = \sum_k P(n|t)z^k$$

which in turn is the solution of

$$\frac{\partial G}{\partial t} = (z-1)\left(b(z-1) - (d-b)\right)\frac{\partial G}{\partial z}.$$

Using the method of characteristics, it can be shown that the equation above is solved by

$$G(z,t) = 1 + \frac{e^{-\nu t}(z-1)}{1 - \frac{b}{\nu}(1-e^{-\nu t})(z-1)}. \tag{S5}$$

Consistent with the main text, the speciation rate is defined by $\nu = d - b$. To obtain the solution (S5) we imposed $G(1,t) = 1$ ensuring the normalization of the probability distribution $P(n|t)$ (i.e. the sum over all values of n is equal to one), and $G(z,0) = z$ corresponding to the point speciation initial condition $n = 1$.

Eq. (S5) is the generating function of an exponential distribution with time-varying coefficients and the explicit solution of Eq. (S4) is therefore obtained by transforming back from this generating function to the exponential $P(n|t)$. For $n \geq 1$, it holds

$$P(n|t) = \frac{e^{-\nu t}}{(\nu + b(1-e^{-\nu t}))^2}\left[\frac{b(1-e^{-\nu t})}{\nu + b(1-e^{-\nu t})}\right]^{n-1}, \tag{S6}$$

while for $n = 0$

$$P(0|t) = 1 - \frac{e^{-\nu t}}{1 + \frac{b}{\nu}(1-e^{-\nu t})}.$$

The expected species richness in this model is given by

$$\begin{aligned} S &= \nu J \sum_{n=1}^{\infty} \int_0^{\infty} dt P(n|t) \\ &= \nu J \int dt \frac{e^{-\nu t}}{1 + \frac{b}{\nu}(1-e^{-\nu t})} \\ &= \frac{\nu J}{b} \log\left(\frac{b}{\nu}\right), \end{aligned}$$

i.e. we sum over all speciation events in the history of the population (of total size J), and compute the probability of those variants being in the population in the present time. Similarly, the expected distribution of variant abundances (known as the species abundance distribution in the ecology literature) in this model is given by

$$\begin{aligned} S(n) &= \nu J \int_0^{\infty} dt P(n|t) \\ &= \nu J \int_0^{\infty} dt \, \frac{e^{-\nu t}}{(\nu + b(1-e^{-\nu t}))^2}\left[\frac{b(1-e^{-\nu t})}{\nu + b(1-e^{-\nu t})}\right]^{n-1} \\ &= \frac{\nu J}{b}\left[\frac{b}{b+\nu}\right]^n. \end{aligned}$$

S3 NZS Solution for the Progeny Distribution

We now derive the joint probability distribution $Q(n,k|T,n_0)$ that after time T, a variant has n extant individuals, and has had a total of k birth events during the time interval from 0 to T, conditioned on the initial abundance n_0 at time 0. Marginalizing $Q(n,k|T,n_0)$ will lead to a prediction for the neutral progeny distribution, a quantity rarely considered in ecological contexts, but used as a test of neutrality in cultural evolution. Note that we are

not necessarily starting this time interval at the speciation time, and so the variant could have some arbitrary abundance n_0 at the start of our time interval. Initially, though, we will drop the n_0-dependence and work with initial condition $n_0 = 1$.

For the birth death process described in the last section it holds

$$\begin{aligned}\frac{dQ}{dT} = \quad & b(n-1)Q(n-1,k-1|T) - bnQ(n,k|T) \\ + \; & d(n+1)Q(n+1,k|T) - dnQ(n,k|T).\end{aligned}$$

Note that k does not affect any of the rates. We now consider a new generating function, $G(z,y,T)$, defined as

$$G(z,y,T) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} Q(n,k|T)z^n y^k$$

which then satisfies

$$\frac{\partial G}{\partial T} = [bz(yz-1) - d(z-1)]\frac{\partial G}{\partial z} \tag{S7}$$

with initial and boundary conditions

$$\begin{aligned}G(1,1,T) &= 1, \\ G(z,y,0) &= z\end{aligned} \tag{S8}$$

For a more general initial condition $n_0 \neq 1$ the latter condition changes to z^{n_0}.

Eq. (S7) has a solution of the form

$$G(z,y,T) = \frac{A(y) - C(y)\left(\frac{A(y)-B(y)z}{C(y)+B(y)z}\right)e^{T/F(y)}}{B(y)\left[\left(\frac{A(y)-B(y)z}{C(y)+B(y)z}\right)e^{T/F(y)} + 1\right]}$$

with

$$\begin{aligned}F(y) &= \left[(b+d)^2 - 4bdy\right]^{-\frac{1}{2}}, \\ A(y) &= 1 + F(y)(b+d), \\ B(y) &= 2byF(y), \\ C(y) &= 1 - F(y)(b+d).\end{aligned}$$

Due to the linear nature of the problem the solution for more general initial conditions, n_0, is given by

$$G(z,y,T,n_0) = G(z,y,T)^{n_0}.$$

Finally, we can marginalize over the unobserved n (assuming we have knowledge about the progeny, and not about total abundances/census counts) by setting $z = 1$

$$\begin{aligned}H(y,T,n_0) &= G(1,y,T)^{n_0} \\ &= \left(\frac{A(y) - C(y)\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right)e^{T/F(y)}}{B(y)\left[\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right)e^{T/F(y)} + 1\right]}\right)^{n_0}.\end{aligned}$$

Weighting $q(k|T,n_0)$ by the steady state species abundance distribution and taking the asymptotic limit of large T leads to

$$\begin{aligned}
H_{\text{extant}}(y,T) &= \sum_{n_0} S(n_0) H(y,T,n_0) \\
&= \sum_{n_0} S(n_0) \left(\frac{A(y) - C(y)\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right) e^{T/F(y)}}{B(y)\left[\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right) e^{T/F(y)} + 1\right]} \right)^{n_0} \\
&= -\frac{\nu J}{b} \log \left[1 - \frac{b}{d} \left(\frac{A(y) - C(y)\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right) e^{T/F(y)}}{B(y)\left[\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right) e^{T/F(y)} + 1\right]} \right) \right].
\end{aligned}$$

We do not yet account for new variants that can appear during the interval T, and themselves contribute to this birth event count. To include these instances we change the initial condition (S8) to $G(z,y,0) = y$, i.e. there is one instance in both, the variant population and its progeny distribution, immediately at speciation. Therefore, this contribution takes the form

$$\nu J \int_0^T d\tau y H(y,\tau,1)$$

with an extra factor of y relative to the results above. This means new variants arise at a rate νJ per unit time, they begin per definition with a single instance and single contribution to the progeny distribution, and persist from their innovation time up until T. So in total

$$\begin{aligned}
H_{\text{total}}(y,T) &= H_{\text{extant}}(y,T) + H_{\text{new}}(y,T) \qquad \text{(S9)} \\
&= -\frac{\nu J}{b} \log \left[1 - \frac{b}{d} \left(\frac{A(y) - C(y)\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right) e^{T/F(y)}}{B(y)\left[\left(\frac{A(y)-B(y)}{C(y)+B(y)}\right) e^{T/F(y)} + 1\right]} \right) \right] \\
&+ \nu J \left(y \frac{A(y)}{B(y)} T - \frac{2yF(y)}{B(y)} \log \left[\frac{C(y) + B(y) + (A(y) - B(y)) e^{T/F(y)}}{2} \right] \right).
\end{aligned}$$

This is the generating function of the neutral progeny distribution, under the non-zero sum formulation of the neutral theory.

S3.1 Approximations for large T

For large T, it holds

$$H(y,T,n_0) \simeq \left(-\frac{C(y)}{B(y)} \right)^{n_0} = \left(-\frac{1 - F(y)(b+d)}{2byF(y)} \right)^{n_0}.$$

Keeping only this leading term of this expansion, and considering the special case of $n_0 = 1$, $H(y, T, n_0)$ can be inverted analytically to give

$$q(k|T,1) \simeq \frac{2d}{b+d}\left(\frac{4bd}{(b+d)^2}\right)^k (-1)^k \binom{\frac{1}{2}}{1+k}. \tag{S10}$$

There is a power law phase $\propto k^{-3/2}$ resulting from the asymptotics of the binomial coefficient, and for sufficiently large k there is an exponential drop-off. Eq. (S10) can be written in terms of the per capita speciation rate ν, as

$$\begin{aligned} q(k|T,1) &\simeq \frac{2}{2-\frac{\nu}{d}}\left(\frac{4(1-\frac{\nu}{d})}{4(1-\frac{\nu}{d})+\left(\frac{\nu}{d}\right)^2}\right)^k (-1)^k \binom{\frac{1}{2}}{1+k} \\ &= \frac{2}{2-\frac{\nu}{d}}\left(1+\frac{\left(\frac{\nu}{d}\right)^2}{4(1-\nu/d)}\right)^{-k} (-1)^k \binom{\frac{1}{2}}{1+k}. \end{aligned}$$

For small enough ν the exponential phase kicks in only for relatively large cumulative abundances, i.e. for small ν, it holds

$$q(k|T,1) \simeq \left(1+\frac{\nu}{2d}\right) e^{-\left(\frac{\nu}{2d}\right)^2 k} (-1)^k \binom{\frac{1}{2}}{1+k}$$

which could be compared to the ν dependence of the species abundance distribution $S(n)$.

This concludes the consideration of a single variant, with $n_0 = 1$ instances initially. Because each variants is guaranteed to go extinct ($d > b$ in the NZS neutral model), there is a finite solution for the cumulative birth distribution at late times. If we now turn to the whole population, represented by $H_{\text{total}}(y, T)$, we encounter a problem. The first term $H_{\text{extant}}(y, T)$ is finite, as all of the variants summed over will go extinct and produce a finite number of birth counts. However, the second term $H_{\text{new}}(y, T)$ will produce an infinite number of birth counts, and eventually will dwarf the contribution from the steady-state variants contained in $h_{\text{total}}(y)$, i.e will dwarf contributions from variants that were already present at $T = 0$. Consequently, if the population persists indefinitely, all those initial variants will produce their contribution to the birth counts and eventually die out. The population, however, will continue to exist via new variants and the limit for the total number of births will tend to ∞.

We start with examining the limit of large T for $H_{\text{extant}}(y, T)$

$$\begin{aligned} \lim_{T\to\infty} H_{\text{extant}}(y,T) &= \lim_{T\to\infty} \sum_{n_0} S(n_0) H(y,T,n_0) \\ &= \sum_{n_0} S(n_0)\left(-\frac{1-F(y)(b+d)}{2byF(y)}\right)^{n_0} \\ &= -\frac{\nu J}{b}\log\left[\frac{-(b+d)+2dy+\sqrt{(b+d)^2-4bdy)}}{2dy}\right]. \end{aligned} \tag{S11}$$

There is no analytical expression for the distribution corresponding to this generating function, i.e. Eq (S11) cannot be inverted analytically. But using numerical techniques we confirm that the generating function produces a distribution characterized by a power law with exponent $-3/2$ followed by exponential decline.

Further, for large T, it holds for the new variants

$$H_{\text{new}}(y,T) \simeq -\nu JTy\frac{C(y)}{B(y)} = -\nu JT\frac{1-F(y)(b+d)}{2bF(y)}. \tag{S12}$$

As pointed out above, there are an unbounded number of birth events from new variants introduced during the interval T, and expression (S12) (valid for large T) will eventually dominate the finite numbers coming from the term $H_{\text{extant}}(y,T)$. The total number of births from new and extant variants are equal when roughly $T \sim \frac{1}{\nu}$. Beyond this point there are very few instances from the extant variants at $T = 0$, and an ever increasing number from novel variants introduced during the considered interval. Note also that this is not a normalized distribution yet and therefore it is not problematic that its coefficients diverge for large T: the coefficients of this generating function are the actual number of variants producing a given cumulative number of births, not the probability that a single variant will produce a given number of births. However, normalization leads to

$$\begin{aligned}\frac{H_{\text{total}}(y,T)}{H_{\text{total}}(1,T)} &\simeq \frac{yC(y)/B(y)}{C(1)/B(1)} = \frac{yC(y)}{B(y)}\frac{2b}{(d-b)\left(1-\frac{b+d}{d-b}\right)} \\ &= -\frac{yC(y)}{B(y)}\end{aligned}$$

for late times T. This normalized distribution at very late times is given exactly analytically by the same distribution we found above, but with $k \to k-1$ reflecting the fact that the initial single instance already counts as a birth event. So it always holds $k > 0$ and we obtain

$$q(k) = (-1)^{k-1}\binom{\frac{1}{2}}{k}\frac{2d}{b+d}\left(\frac{4bd}{(b+d)^2}\right)^{k-1}. \tag{S13}$$

This of course comes from the fact that at large enough T, we are just summing together the entire number of births for multiple variants starting with $n_0 = 1$. The corresponding cumulative distribution is straightforward to compute analytically in terms of a hypergeometric function for the cumulative distribution for (S13). Putting it together leads to

$$P(K \geq k) = (-1)^{k-1}\frac{b+d}{2b}\left(\frac{4bd}{(b+d)^2}\right)^k\binom{\frac{1}{2}}{k}{}_2F_1\left[\begin{matrix}1 & (-1/2+k) \\ 1+k & \end{matrix}; \frac{4bd}{(b+d)^2}\right].$$

S4 Maximum likelihood estimation

In this section we derive the maximum likelihood estimate of the ratio $\eta = d/b$.

The log likelihood of observing a given set of S cultural variants with abundances $\{k_i\}$ at late times is given by

$$L = \sum_{i=1}^{S} log(q(k_i)) = \sum_{i=1}^{S} \log \left[\frac{2d}{b+d} \left(\frac{4bd}{(b+d)^2} \right)^{k_i - 1} (-1)^{k_i - 1} \binom{\frac{1}{2}}{k_i} \right]$$

which can be rewritten as

$$L = \sum_{i=1}^{S} \log \left[\frac{2\eta}{1+\eta} \left(\frac{4}{\frac{1}{\eta} + 2 + \eta} \right)^{k_i - 1} (-1)^{k_i - 1} \binom{\frac{1}{2}}{k_i} \right]$$

by using the relation $\eta = \frac{d}{b} = \frac{\nu}{b} + 1$. It holds

$$\frac{\partial L}{\partial \eta} = \sum_{i=1}^{S} (k_i - 1) \frac{\eta - 1}{\eta(1+\eta)} + \frac{S}{\eta(1+\eta)}.$$

Setting $K_{\text{total}} = \sum_{i=1}^{S} k_i$ and solving $\frac{\partial L}{\partial \eta} = 0$ leads to

$$\eta = \frac{K_{\text{total}}}{K_{\text{total}} - S}.$$

S5 Simulation of the Overlapping Generations Model via Gillespie algorithm

The NZS approximation described in section 2(a) in the main text has been extensively compared with both, simulations and analytical results for ecological populations with symmetric, competitive interactions. In general, it has been demonstrated that the predictions of the NZS approximation for the distribution of variant abundances at a single point are valid when the innovation rate satisfies $\nu J >> 1$, and begin to break down when νJ is small. To test the validity of the approximation (3.1) given in the main text, we take the same approach and simulate a group of competing variants, but compute the resulting progeny distribution after a long time interval, rather than the species abundance distribution at a single point in time.

The simulated populations are described by stochastic Lotka-Volterra systems, where variant i with current abundance n_i will increase abundance by one individual at a rate $b_0 n_i$, undergo intrinsic mortality and decrease abundance by one at a rate $d_0 n_i$. Further, competitive interactions involve the focal variant of abundance n_i in a population of current size J and occur at a rate $\alpha n_i J$. The strength of competition is controlled by the parameter α and its outcome is the loss of one instance either from the focal variant or from the rest of the population. New variants are introduced at a rate νJ with initial abundance 1, and are considered as an error in the birth process. Therefore the effective per capita birth rate (i.e. the rate of production of instances of the same variant) is $b_0 - \nu$. In summary, the rates of these processes for variant i

are as follows

process	description	rate
$n_i \to n_i + 1$	birth	$(b_0 - \nu)n_i$
$n_i \to n_i - 1$	intrinsic mortality	$d_0 n_i$
$n_i \to n_i - 1$	competition	$\alpha n_i \sum_{\forall j} n_j$
$0 \to 1$	speciation	$\nu \sum_{\forall j} n_j$

(S14)

where the labels i and j refer to the extant variants in the system at any given point in time, and the sums are taken over all variants, including variant type i.

The simulation of this population is based on the well-known Gillespie algorithm [13]. This approach involves a sequence of transitions drawn from the possibilities given in (S14), with a waiting time in between each of these events. For example, for a system with three variants, a birth event for one of the three types could be followed by a competitive interaction between the other two with the outcome that type three loses an instance, and so on. For a given configuration of instances the waiting time between two events is distributed according to an exponential distribution with a mean time equal to the inverse of the sum of all rates. When an event occurs, the kind of transition is randomly chosen with weights proportional to their rates. Consequently, events are more likely to involve an abundant variants, because all processes are weighted by total variant abundance (see (S14)).

Finally, the intrinsic rates given in (S14) represent the exact description of the population dynamics. In order to evaluate the accuracy of the NZS approximation we need to map those intrinsic rates onto the parameters of the NZS approximation. This mapping is such that the *effective* birth rate of each variant is given by $b = b_0 - \nu$, while the effective mortality rate (incorporating both intrinsic mortality and competition) is given by $d = b_0$. As d_0 does not directly enter the NZS prediction for the progeny distribution, we simulated these populations with $d_0 = 0$.

The NSZ expectation for the steady-state population size was derived in [14]

$$J_{\text{steady}} = \frac{b_0 - d_0}{\alpha} \tag{S15}$$

which simplifies to b_0/α when the intrinsic mortality vanishes. We therefore set an initial condition for the simulated population of b_0/α instances of only one cultural variant.

To ensure that the system has reached an approximate steady-state before we begin sampling the progeny distribution, we allow the system to burn in by waiting until the first monodominant variant has reached extinction. At this point *every extant variant has experienced entirely neutral dynamics, starting from a single instance*, and therefore no deviation from the average steady-state neutral population is expected. From this point onwards, we record all birth events, and begin accumulating the progeny distribution. In order to

provide a valid comparison with the late-time limit of the progeny distribution given by Eq. (3.1) in the main text, we stop sampling after $T = 100b_0/\nu$ time steps (see section S3 for a derivation of this stopping time). Additionally, we verified that the first two moments of the progeny distribution were asymptotic to constant values by this time, and therefore ensured that we indeed sampled the asymptotic progeny distribution for large T.

S6 Data set

The South Australian Attorney-General's department provides two data sets consisting of all boys' and girls' names registered from 1944 to 2013, respectively, in South Australia. These data sets can be found under (last accessed 27.02.2017)
https://data.sa.gov.au/data/dataset/popular-baby-names

Between 1944-2013, the total number of girls names registered each year varied from a low of 6748 (in 1944) to a peak of 11754 (in 1971), subsequently declining slightly to between $9000 - 10000$ in the last three decades. The total number of distinct names registered each year varied between 741 and 2923. For boys, the total number of names registered per year varied between 7069 and 12464, following a similar pattern to the girls' names, while the total number of distinct names registered each year varied between 477 and 2450. Clearly, there is systematic variation here in both numbers of names (reflecting a changing population size) and in the diversity of names (potentially reflecting a non-stationarity in the innovation rate). However, this variation may be as small as we can reasonably expect in cultural data.

We also note that this 70 year span is not a priori long enough to apply our asymptotic results. But we have also explored the change in the progeny distribution over time by considering the change in its first two moments as the time interval, T, over which the progeny distribution is computed is varied from one year up to 70 years. If these moments asymptote to a constant, this would indicate that the distribution is approaching its asymptotic form. We find that these moments are still changing in time as T approaches 70 years, but that this change is relatively slow, indicating that this value of T is close to the asymptotic regime. Therefore we propose that it is reasonable to apply our methodological approach, which assumes that the system is in a steady state with a temporally constant innovation rate ν, and compare the Australian baby name data to the asymptotic form of the progeny distribution for large time intervals that we derived in Eq. (3.2) in the main text.

We stress that in general we have to be careful in drawing conclusions from observed data too firmly. In part because the data likely does not reflect a population in steady-state, or with a constant innovation rate over time, and may only barely span a long enough time frame for our asymptotic results to be applicable. But at the very least, our approach might lead to ways to incorporate this variation, which is inevitably present in real data, and has been underexplored in studies of cultural evolution so far.

Additionally we note that different geographical regions will differ in their legislation towards the use of novel baby names (e.g. administrative approval processes might be more or less stringent) which naturally influences the rate of innovation. But our analysis is focused on the spread behavior of innovation, i.e. variants that have been introduced into the system with abundance one. Our results indicate that e.g. the ratio between singletons and variants with abundance two is sensitive to the underlying process of cultural transmission. External processes affecting the rate of innovation might not influence this ratio strongly. Further, the size of the 'name space' (meaning the space of all feasible names given the conventions of the particular language) is usually not known. This leads to the question whether the name space could become exhausted over time resulting in a decline of the innovation rate. While this is a valid concern we did not see a strong indication of such a phenomenon in the considered data set: the innovation rates did not show a strong decline over time.

References

[1] WJ Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*. Springer Science & Business Media, 2012.

[2] FD Neiman. Stylistic variation in evolutionary perspective: inferences from decorative diversity and interassemblage distance in illinois woodland ceramic assemblages. *American Antiquity*, 60:7–36, 1995.

[3] SJ Shennan and JR Wilkinson. Ceramic Style Change and Neutral Evolution : A Case Study from Neolithic Europe. *American Antiquity*, 66:577–593, 2001.

[4] TA Kohler, S VanBuskirk, and S Ruscavage-Barz. Vessels and villages: evidence for conformist transmission in early village aggregations on the pajarito plateau, new mexico. *Journal of Anthropological Archaeology*, 23(1):100–118, 2004.

[5] RA Bentley, MW Hahn, and SJ Shennan. Random drift and culture change. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1547):1443–1450, 2004.

[6] J Steele, C Glatz, and A Kandler. Ceramic diversity, random copying, and tests for selectivity in ceramic production. *Journal of Archaeological Science*, 37(6):1348–1358, 2010.

[7] RA Bentley, CP Lipo, HA Herzog, and MW Hahn. Regular rates of popular culture change reflect random copying. *Evolution and Human Behavior*, 28(3):151–158, 2007.

[8] S Shennan. Descent with modification and the archaeological record. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1567):1070–1079, 2011.

[9] J Rosindell, SP Hubbell, F He, LJ Harmon, and RS Etienne. The case for ecological neutral theory. *Trends in ecology & evolution*, 27(4):203–208, 2012.

[10] M Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

[11] M Kimura. Random genetic drift in multi-allelic locus. *Evolution*, pages 419–435, 1955.

[12] M Kimura and JF Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725–738, 1964.

[13] DT Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[14] JP O'Dwyer and RA Chisholm. A mean field model for competition: From neutral ecology to the red queen. *Ecology Letters*, 17:961–969, 2014.