Unbiased Cultural Transmission in Time-Averaged Archaeological Assemblages

Mark E. Madsen*

Department of Anthropology, Box 353100, University of Washington, Seattle WA, 98195 USA
(Dated: April 9, 2012: Version 1.3)

Unbiased models are foundational in the archaeological study of cultural transmission. Applications have assumed that archaeological data represent synchronic samples, despite the accretional nature of the archaeological record. I document the circumstances under which time-averaging alters the distribution of model predictions. Richness is inflated in long-duration assemblages, and evenness is "flattened" compared to unaveraged samples. Tests of neutrality, employed to differentiate biased and unbiased models, suffer serious problems with Type I error under time-averaging. Finally, the time-scale over which time-averaging alters predictions is determined by the mean trait lifetime, providing a way to evaluate the impact of these effects upon archaeological samples.

I. INTRODUCTION

What makes the archaeological record unique as a way of studying the evolution of culture is time depth, creating the possibility of studying both the unique histories of prehistoric peoples, and the common evolutionary processes that shape those histories. But that time depth is not represented by an empirical record consisting of a series of moments stacked together, each comparable to a synchronic snapshot of a population. Instead, archaeological deposits are almost always accretional palimpsests, representing cumulative artifact discard over durations of varying length [2, 3, 10, 33, 37]. Thus, when archaeologists determine the number of classes present in an assemblage, or calculate their relative frequencies, such measures summarize the bulk properties of deposition and discard over a significant span of time.[1] We refer to assemblages which are accretional in this manner as "time-averaged."

At present our application of cultural transmission models to archaeological data rely upon evaluating the predictions such models make for the richness, diversity, or characteristic frequency distribution of cultural traits. To date, we have not taken into account the effects of time-averaging in comparing these model predictions with archaeological data. If the accumulation of transmission events together in a cumulative palimpsest were to leave the distribution of trait frequencies unchanged, we could compare model predictions such as heterogeneity (e.g., Wright's F_{st}) or the expected distribution of traits under neutrality directly to archaeological data. If, on the other hand, temporal aggregation of transmission events leads to altered distributions or inflated estimates for richness, then our current models may not provide testable explanations for archaeological data. Evaluating this possibility for the commonly employed neutral or unbiased transmission model is the goal of this paper.

In his dissertation, Fraser Neiman [26] considered a potential source of time-averaging effects in diachronic assemblages: variation in discard rates across traits. Neiman found that while individual variation in discard rates could strongly affect the measurement of variant frequencies, if samples were sufficiently time-averaged, individual discard rates would end up mirroring variant frequencies in the whole simulated population. With respect to this particular effect within accretional deposits, Neiman's results suggested that the predictions made by the neutral (or "unbiased") transmission model were directly applicable to the relative frequencies of traits as we would measure them in a time-averaged assemblage.

Nevertheless, there is good reason to think that some classes of measurements will not be directly comparable to the predicted behavior of cultural transmission models. Our colleagues in paleobiology, for example, have documented systematic differences between living and fossil assemblages, including increased species richness, reduced spatial and temporal variance in sample composition, and flatenned species abundance curves in time-averaged assemblages [28, 38, 39]. Lyman [23] extended these results to zooarchaeology, noting that time-averaging can be a significant problem when the process one is applying or studying occurs over a shorter time scale than the empirical record available to study its properties (see also Grayson and Delpech [17]).

These varied results suggest that Neiman's original analysis should be extended. In this paper, I report the results of a numerical simulation study of the sample properties of the unbiased, well-mixed model of cultural transmission, when observed through common statistical measures, and aggregated at a variety of assemblage durations. Archaeological applications of the unbiased transmission model increasingly employ statistical tests for neutrality, such as Slatkin's Exact test [34, 35], and diversity measures based upon trait frequencies [21, 27]. I focus upon both in simulating the effects of time-averaging upon unbiased cultural transmission.

Since cultural transmission models are not tied to an easily measurable time scale, as are genetic models given biological reproduction cycles, relating the simulation time scale to chronological time scales is important for archaeological analysis. I examine the role of the average trait lifetime in determining the onset of time-averaging effects, and employ this as a proposed time scale for determining the extent to which an archaeological assemblage will be comparable to unbiased

---

[1] As well as the action of various post-depositional and taphonomic processes, of course.

model predictions.

II. UNBIASED WELL-MIXED CULTURAL TRANSMISSION

Following Dunnell's [12] suggestion that the distribution of stylistic variation is explained by neutrality with respect to selection, Fraser Neiman [27] introduced the Wright-Fisher infinite-alleles (WF-IA) model to archaeology as a formal description of unbiased cultural transmission within prehistoric populations. Here I review the main characteristics of WF-IA as employed by archaeologists.

The well-mixed neutral Wright-Fisher infinite-alleles model [20] considers a single dimension ("locus") at which an unlimited number of variants ("alleles") can occur, in a population of N individuals.[2] The state of the population in any generation is given in several ways: a vector representing the trait possessed by each individual (census), a vector giving the abundance of each trait in the population (occupation numbers), or by the number of traits represented in a population by a specific count (spectrum).

In each generation, each of N individuals selects an individual at random in the population (without respect to spatial or social structure, hence "well-mixed"), and adopts the trait that individual possessed in the previous generation.[3] Equivalently, a new set of N individuals are formed by sampling the previous generation with replacement. At rate μ for each individual, a new variant is added to the population instead of copying a random individual, leading to a population rate of innovations $\theta = 2N\mu$ [16], with no "back-mutation" to existing traits.[4] An important consequence of this innovation model is that each variant is eventually lost from the population given enough time, and replaced with new variants. Thus, there is no strict stationary distribution for the Markov chain describing WF-IA, although there is a quasi-stationary equilibrium in which the population displays a characteristic number of variants, with a stable frequency distribution governed by the value of θ.

The distribution of variants in the full population is complex, and exact expressions are usually unknown for WF-IA, with most population properties obtained by diffusion approximation [16]. Archaeologists are fortunate, however, that the *sampling* properties of the model are well understood, since

we always deal with samples of past human activity, rather than population-level census information. The basic result, due to Ewens [15], is the joint distribution of variant counts in a sample of size n, which is given by the Ewens Sampling Distribution, given the population innovation rate (θ):

$$\mathbb{P}_{\theta,n}(a_i,\ldots,a_n) = \frac{n!}{\theta^{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!} \qquad (1)$$

where $\theta^{(n)}$ is the Pochhammer symbol or "rising factorial" $\theta(\theta+1)(\theta+2)\cdots(\theta+n-1)$. In most empirical cases, we cannot measure (or do not set through experiment) the value of θ, so a more useful relation is the distribution of individuals across variants (i.e., the occupation numbers), conditional upon the number of variants k_{obs} observed in a sample of size n:

$$\mathbb{P}(n_1,n_2,\ldots,n_k|k_{obs}) = \frac{n!}{|S_n^k|k!n_1n_2\cdots n_k} \qquad (2)$$

where $|S_n^k|$ denote the *Stirling numbers of the first kind*, which give the number of permutations of n elements into k non-empty subsets [1]. The latter serves here as the normalization factor, giving us a proper probability distribution.

From the core probability distributions which compose the WF-IA model, a number of measurable predictions can be derived, relevant to archaeological applications. In this study, I focus upon the most commonly used: statistical tests of neutrality, predictions for the amount of variation expected under neutrality, and the evenness with which variants are represented in the population (as revealed by several diversity measures).

A. Statistical Tests for Neutrality

Equation (2) serves as the basis for tests measuring the fit between empirical samples and the neutral WF-IA. The two most important such tests are the Ewens-Watterson test using the sample homozygosity, and Slatkin's "exact" test [13, 16, 34, 35].[5] Both have been adopted for use by archaeologists, beginning with Neiman [27] and Lipo [22] who described Watterson's work in detail, Slatkin's exact test by Steele and colleagues [36] and Premo and Scholnick [32].

The Slatkin exact test is modeled on the Fisher exact test for contingency tables. Where the Fisher exact test determines the probability of an unordered configuration from the hypergeometric distribution, Slatkin's test determines the probability of a sample of traits (characterized by occupation numbers) with respect to Equation 2, using a recursive algorithm

---

[2] Conventionally, the model treats a diploid population, in which N individuals are composed of 2N chromosomes and thus there are always 2N genes tracked in the population. The haploid version is more appropriate for modeling cultural phenomena, and thus formulas given in this paper may differ from those given by Ewens [16] and other sources by a factor of two. For example, the key parameter θ is defined as $2N\mu$ rather than the common genetic definition $4N\mu$.

[3] An individual can select themselves at random since sampling is with replacement, and this would be equivalent to "keeping" one's existing trait for that generation.

[4] It is important to note that θ is not a measure of the "diversity" of traits in the population, as it has been employed in several archaeological studies, but is instead a *rate* parameter of the model.

---

[5] There are several other important tests of neutrality when dealing with DNA sequence data, including Tajima's D, the HKA test, and the McDonald-Kreitman test [13]. Because their assumptions are highly specific to the structure of sequence data, I omit consideration of them here.

to enumerate all possible unordered configurations.[6] If the observed configuration occurs with probability $\mathbb{P}(c_{\text{obs}})$, summing the probability of all unordered configurations which are *less probable* than than $\mathbb{P}(c_{\text{obs}})$ places the observations on the cumulative distribution curve of Equation 2 [34, 35], providing a *p* value suitable for a two-tailed hypothesis test.

The advantage of Slatkin's exact test arises in that is wholly independent of what the alternative hypothesis to neutrality might be in any given case, and thus is a "pure" test of whether the observed sample is likely to have been drawn from the Ewens Sampling Distribution [32]. Slatkin's exact test is employed in this study, in preference to the Ewens-Watterson test, which incorporates the diploid notions of "homozygosity," which has no analogue in our simple model of social learning and imitation.

B. Class or Trait Richness

The number of variants expected K_n in a sample of size n is another important observable quantity which is measurable for archaeological assemblages as the nunber of unique classes or attributes present. Class richness is often used as an indication of conformist cultural transmission, with conformism yielding fewer variants in archaeological samples than are expected given unbiased transmission [21]. Ewens [16, Eq. 3.84] gives the full distribution of K_n as:

$$\mathbb{P}(K_n = k) = \frac{|S_n^k|\theta^k}{\theta^{(n)}} \qquad (3)$$

This distribution is difficult to use directly, so an approximation is usually required.[7] Using the frequency spectrum of the WF-IA, Ewens [16, Eq. 3.94] gives the expected value of K_n for a specific θ value:

$$\mathbb{E}(K_n) = \int_0^1 (1 - (1-x)^n) \frac{\theta}{x} (1-x)^{\theta-1} dx \qquad (4)$$

The amount of variation present in samples, at a given population size, is a key observable difference between unbiased cultural transmission and "conformist" models, where individuals have a disproportionate tendency to adopt the most popular traits within the population. Simulation models of conformist transmission tend to be depleted of variation rapidly unless the conformist tendency is very weak, or the innovation rate very strong.

C. Diversity Measures

In archaeological contexts, we rarely have independent estimates of population size, which makes it difficult to compare K_n values between assemblages. Kohler and colleagues [21] attempted to estimate θ from observed samples, and use a comparison of θ estimates (t_e from Neiman [27]) and the evenness of observed trait frequences (Neiman's t_f). The latter measure is related to Sewall Wright's classic measures of heterozygosity ($\hat{F}$) in genetic models.

$$t_f = \frac{1}{\sum_{i=1}^k p_i^2} - 1 \qquad (5)$$

It is apparent that Equation 5 is a measure of the dispersion of counts across categories, and thus equivalent to Wilcox's "index of quantitative variation," which varies between 0 (when all cases belong to a single category), and 1 (when all cases are evenly divided across categories):

$$\text{IQV} = (\frac{k}{k-1})(1 - \sum_{i=1}^k p_i^2) \qquad (6)$$

In the current analysis, I focus upon t_f and its normalized variant, IQV, to determine the effect of time-averaging upon evenness. Paleobiologists have found that fossil assemblages have considerably "flatter" species diversity curves compared to living communities, and I expect that time-averaging will have the effect here of pushing IQV towards 1.0 compared to its value in unaveraged samples.

In this paper, I do not take up the accuracy of θ estimates from knowledge of K_n in archaeological samples.[8] Ewens [16] has shown that K_n is a sufficient statistic for θ. This means that having additional information (e.g., the trait occupation numbers themselves) does not increase the accuracy of the maximum likelihood estimate for θ. Since K_n and n are integers (and often relatively small ones), the precision with which θ can be estimated is limited, and the variance of those estimates is very high.

III. METHODS

The WF-IA model described in the previous section is most commonly implemented in theoretical population genetics as a "backward-time" model, generating random coalescent genealogies in order to study their properties [13, 16, 40]. The backward-time strategy is highly efficient when applicable, and allows the study of varied models of migration, mutation,

---

[6] With larger sample sizes and values of K_{obs}, it becomes impractical to enumerate all possible configurations and thus determine the likelihood of an observed sample. In such cases, Monte Carlo sampling of configurations from the Ewens Sampling Distribution is used.

[7] MathStatica 2.5 running in Mathematica 8.0 is able to work with the distribution numerically, but cannot provide symbolic solutions for moments. Given the rising factorial, it is computationally expensive to calculate values from Eq. (3) for large values of θ and n.

[8] Although the simulation results presented here suggest that θ estimates from samples of trait counts are highly biased, and time-averaging does not improve matters. After additional simulation studies to better understand the magnitude of the bias, I will describe the effects in a companion paper.

and population structure. Because coalescent-based methods only study the lineage of individuals leading from a current sample to the most recent common ancestor (MRCA), backward-time approaches are not well suited to studying the quantitative properties of the entire population as it evolves.

For archaeological research, we are frequently interested in characterizing an evolving population in our simulations, samples of which will become part of the depositional record whether or not individual cultural traits survive within the population. Thus, for the current research I employ a forward-time approach, evolving a simulation population of individuals generation by generation, subject to rules which specify transmission patterns and innovation rates. Several standard forward-time models for WF-IA exist, such as **simuPOP** [30, 31]. Genetic simulation frameworks usually have sequence-like representations of the heritable information, and feature statistical measures specific to sequence data. Thus, I use an individual-based cultural transmission simulation framework called **TransmissionFramework**, written by the author, which is designed specifically to support the types of models of interest to anthropologists.[9]

A. Model Validation

Simulation software can be complex, and it is important not merely that software is free of major bugs, but that a numerical simulation accurately reflects the underlying mathematical properties of the desired theoretical model. I validated the proper functioning of a **TransmissionFramework** implementation of the WF-IA model by comparing the analytical result of Equation (4) to simulated results, for a large sample of simulation runs. The expected number of traits is a good validation estimate because the number of variants present in a sample will be sensitive to the relative rates of copying and innovation events being handed correctly in the simulation code. Errors in handling these events in software will be magnified across many individuals over many simulation steps.

To ensure that behavior is correct across a range of useful θ values, I performed multiple simulation runs at θ values ranging from 2 to 40, for 5000 generations in a simulated population of 2000 individuals. Each parameter combination was represented by 3 simulation runs. The initial transient behavior of the model is discarded from data analysis by skipping the first 750 generations, given the mixing time analysis by Watkins [41]. At each time step in a simulation run, the simulator took a sample of 30 individuals and tabulated the traits held by those individuals, and recorded the value of K_n. This yielded 408,478 samples across across validation runs.

Using Mathematica 8.0 with MathStatica 2.5 installed, I calculated expected values for each θ level used in simulation, employing Equation (4). Table I compares the expected

Theta	$\mathbb{E}(K_n)$	Simulated $\bar{K}_n$	Sim. Stdev K_n
2	6.054	6.511	1.838
4	9.022	8.991	2.269
8	12.869	12.616	2.464
12	15.397	15.306	2.571
16	17.228	17.187	2.569
20	18.629	18.737	2.486
40	22.601	22.693	2.253

TABLE I. Comparison of expected K_n from (4) with simulated values from WF-IA model, for θ values from 2 to 40. Total sample size across θ values is 408,478 samples of size 30.

and observed values. In all cases, the analytical results are extremely close to the observed mean K_n values from simulation, and certainly well within 1 standard deviation. Thus, I conclude that the **TransmissionFramework** implementation of WF-IA employed in this study accurately represents the desired theoretical model.

B. Time-Averaging and Simulation Parameter Space

Time-averaging was implemented in **TransmissionFramework** by implementing a series of statistical "windows" within which trait counts were aggregated between time steps. At the end of each temporal window, a sample of appropriate size is taken, trait counts within that sample tabulated, and K_n values recorded. The simulator architecture allows an arbitrary number of temporal windows to be employed simultaneously (albeit with a small performance penalty for each window). As a consequence, during a single simulation run, the simulator tracks both unaveraged statistics and the same statistics averaged over any number of "assemblage durations." All trait samples taken in the simulator, whether unaveraged or for a specific assemblage duration, were also recorded to allow calculation of Slatkin's Exact test. Additionally, to facilitate analysis of time scales within the simulation model, for each trait the interval between entry and loss through drift was recorded. In the simulation results reported here, samples were of uniform size 100. Constant sample sizes reduces the effects of sample size on the reported results, although the interaction of the fixed sample size and the innovation rate will be a concern at high θ values.

All simulations reported here were performed with a population size (N) of 2000 individuals, and simulation runs were conducted for the following values of θ: 0.1, 0.25, 0.5, 1.0, 2.0, 5.0, and 10-100 at intervals of 10. This range encompasses innovation rates which are very small, through populations in which a full 5% of the population has a never-before seen variant each generation. Simulations were performed in several batches, with a core set of runs performed for 40,000 steps in order to determine the effects of long-duration time-averaging, yielding simulated assemblages at a variety of windows ranging from 3 steps to 8000 steps (the exact durations sampled are given in the first column of Table II). In order

---

[9] **TransmissionFramework** is available as open-source software at `http://github.com/mmadsen/TransmissionFramework`, and runs on any platform with a Java 1.6+ runtime, with optional scripts requiring Ruby 1.9. The version used for this study is tagged as Version 1.6 in the source repository.

TA Duration	Min Sample Size	Max Sample Size
1	130494	247491
3	4497	43494
7	1926	18639
15	897	8694
31	435	4209
62	216	2103
125	105	1038
250	516	981
500	255	486
1000	114	228
2000	57	114
4000	27	54
8000	12	16

TABLE II. Breakdown of sample sizes for analysis of trait richness (K_n), by size of time-averaging "window." Some values of θ required larger numbers of simulation runs to achieve stable result, thus the difference between samples sizes at the same TA duration.

to increase the sample size of long-duration assemblages, a second set of simulation runs using the same parameters were done with only the five largest windows recorded (the short duration window sizes were discarded to avoid a flood of raw data beyond that needed for stable statistical analysis). Finally, since the statistical behavior of the process at very small values of θ is highly variable, a third set of runs was performed to increase the number of samples for θ values between 0.1 and 1.0.

Trait samples were post-processed outside the simulator environment, since calculation of Slatkin Exact tests within the simulator itself would slow down the primary simulation model by a large factor. Montgomery Slatkin's original C language program was used in Monte Carlo mode to produce an estimate of $\mathbb{P}(E)$ for each sample of individuals. I modified Slatkin's original `montecarlo.c` program to not require the data to be embedded in the source code, instead taking data as a command line parameter, and outputting only the $\mathbb{P}(E)$ value and θ estimate, to allow easy post-processing of the simulation output.[10]

The simulation results reported here, once post-processed, comprise 3,024,085 sample values for K_n, across the θ values listed above, and broken down across assemblage durations as in Table II, and 1,113,134 Slatkin Exact test results for the same combinations of θ and assemblage duration.

---

[10] These modifications are available, along with all other analysis scripts, in the Github repositories http://github.com/mmadsen/saa2012, and the **TransmissionFramework** source code.

IV. RESULTS

Figure 1 gives the mean number of variants (K_n) in a samples of uniform size 100, for assemblage durations listed in the first column of Table II. The time axis represents simulation steps, each of which represents $N = 2000$ copying events within the population. Immediately apparent is the fact that time-averaging does inflate the number of traits seen in a sample, but to a much greater extent as the population innovation rate (θ) increases. At very low values of θ, the effect is very small and does not manifest except in very long duration assemblages. The "onset" of inflation due to time-averaging appears to vary systematically with θ, with inflation of K_n appearing in shorter duration assemblages, as θ increases.

This is a direct consequence of the equilibrium attained in the WF-IA model, and yields a "natural" time scale over which we should observe significant turnover in the trait population. At equilibrium, the population will fluctuate around a mean number of traits, as individual traits enter and leave the population constantly. This implies that at equilibrium, if we add traits at a higher rate due to migration or innovation, more traits must be lost to drift each generation. WF-IA thus satisfies a balance equation characterizing the average number of variants ($\bar{n}$)[14]:

$$\frac{\bar{n}}{\bar{t}} = \theta \tag{7}$$

where $\bar{t}$ represents the average number of generations that a new trait lasts in the population before its loss to drift (i.e., the mean trait lifetime).

An exact expression for mean trait lifetime has not been derived from the transition probabilities of the WF-IA Markov chain [14], but it can be approximated by summing the average amount of time that a trait within a population spends at specific frequencies (i.e., mean sojourn times). Ewens [16, Eq. 3.20] gives the following approximation:

$$\bar{t} \approx \mathbb{E}(t_i) = \sum_{j=1}^{\infty} \frac{2N}{j(j-1+\theta)}(1 - (1-p)^j) \tag{8}$$

Since θ is in the denominator of the summation, increasing the population rate of innovation introduces variants to the population more quickly, which increases the rate at which traits with low frequencies are lost from the population, thus reducing the mean trait lifetime. The mean trait lifetime is thus controlled by θ, but the latter is not directly observable in archaeological samples, whereas the duration of artifact classes is. The duration of an artifact class is not precisely equivalent to the WF-IA's trait lifetime, but the two concepts are commensurable given further development of model and classification methods (see Section V).

Table III lists the observed mean lifetime of traits for each level of θ employed in this study, and the expected value as calculated using Equation 8. The observed values are systematically lower than the expected values, which reflects slightly faster loss of traits due to drift in a finite and small population compared to the large populations often studied in population genetics [14, 20].

Theta	Mean Trait Lifetime	$\mathbb{E}(t_i)$
0.10	36.54	36.89
0.25	25.61	24.05
0.50	21.10	19.97
1.00	17.31	17.21
2.00	14.57	15.21
5.00	12.43	13.05
10.00	10.83	11.57
20.00	9.50	10.16
30.00	8.68	9.36
40.00	8.12	8.79
50.00	7.72	8.36
60.00	7.36	8.01
70.00	7.08	7.72
80.00	6.83	7.46
90.00	6.60	7.42
100.00	6.40	7.05

TABLE III. Mean lifetime (in model generations) of traits, by θ, along with analytical approximation from Equation 8.

Examination of Figure 1 appears to show that the onset of time-averaging effects, however small, occurs around the time scale of the mean trait lifetime, for values of $\theta \geq 1.0$. This outcome is sensible given the enhanced probability of longer duration samples incorporating new variants in the sample due to innovation. In the analyses to follow, I scale the time variable by the mean trait lifetime, displaying assemblage duration as a multiple of this value. Thus, for the remainder of this paper, a scaled assemblage duration of 100 will indicate 100 times the mean trait lifetime at that specific θ value. For example, if we are examining results at $\theta = 5.0$, a scaled duration of 100 would indicate $12.43 * 100 = 1243$ simulation steps. In the following sections, I investigate the effects of temporal aggregation on Slatkin Exact test results, richness, and diversity, using assemblage durations scaled in this manner.

A. Slatkin Exact Tests

The Slatkin Exact test for neutrality, discussed in Section II A, determines the "tail" probability for a sample of size n, with observed number of traits k, to be derived from the Ewens Sampling Formula (Equation 2). The test employed in this study is Slatkin's Monte Carlo version, which allows the use of larger sample sizes, using random selection to create unlabeled configurations from the ESD to compare against the observed values. The resulting tail probability is converted into a standard hypothesis test by selecting an α value. For purposes of this study, I considered the upper and lower 5% of tail probabilities to indicate that a sample was probably not derived from a neutral transmission model, leading to $\alpha = 0.10$.

Since all of the samples generated by simulation in this study are verifiably generated by a neutral transmission process, what is of interest here is the fraction of samples which

fall into the upper and lower tail 5% tail regions, and thus would normally represent a rejection of the null hypothesis of neutrality. Since the ESD is a probability distribution, we should expect that up to 10% of the samples taken from simulation runs would be in the tail region, even though they were generated by the WF-IA.[11]

If time-averaging has no effect on the validity of the Slatkin Exact test employed against temporally aggregated samples, we would expect the fraction of samples in the two tails (upper and lower 5% in this case) to equal 10%. Anything over 10% would constitute evidence of extra Type I errors, since we know the samples to have been generated by a process meeting the definition of the null hypothesis. Therefore, after post-processing the simulation output to produce Slatkin tests as described in Section III B, I tabulated the fraction of Slatkin Exact tail probabilities that exceeded the expected 10% tail population. These are, in other words, "excess" failures of the Slatkin Exact test, beyond those expected by the probability distribution itself. For each θ level, and for each time-averaging duration, the mean "excess" failure rate was computed, from the 1,113,134 raw Slatkin Exact test results generated in the simulation study.

Figure 2 depicts the relationship between the excess failure rate, and time-averaging duration scaled by the mean trait lifetime (as previously described). The mean trait lifetime is indicated by a vertical red bar in each graph. Three major results are apparent. First, at values of $\theta \geq 1.0$, the excess failure rate in non-time-averaged data is zero, as one would expect, and then begins to increase (albeit slowly) as the time-averaging duration of samples exceeds the mean trait lifetime. In some cases, such as $\theta = 5.0$, the Slatkin Exact test continues to be accurate given the chosen α value through samples which are aggregated for 10 times the mean trait lifetime. But in all cases, with sufficient time-averaging, the Slatkin Exact test begins to suffer from increased Type I error, reporting an ever increasing fraction of samples as not derived from a neutral transmission process. The extremal situation is seen at very high rates of innovation, where nearly every test fails, at high levels of time-averaging. These failures are caused by saturation of a finite sample with singleton traits, causing the sample to display too much evenness in frequency to have been the result of sampling from the Ewens Sampling Formula. But unrelated to this saturation effect, there is considerable failure in employing the Slatkin Exact test to detect neutrality. For example, at $\theta = 5.0$, at 100 times the mean trait lifetime, approximately 70% of all samples appear in the tail region of the distribution, compared to the expected 10%. Clearly, the Slatkin Exact test is not robust for long-duration assemblages.

Second, at low θ values, the Slatkin Exact test does not appear to be robust, even without time-averaging. This does not seem to be widely noted in the literature, and I do not have

---

[11] This does point out the limited value of a very small number, or especially a single, Slatkin Exact test from a single set of class frequencies or counts, as has been archaeological practice (of necessity). Larger suites of Slatkin tests from assemblages within a region may represent a better means of applying this test in archaeological circumstances.

an explanation for this phenomenon in the current study. The likely cause is the very small innovation rate allowing existing traits to persist in the population and thus become overrepresented in a finite sample compared to the ESF. It may also be the case that given the rarity of innovations with very low values of θ, that stable results require much larger population sizes and larger sample sizes than used in the present study. I leave this as an open problem, for future research. If the result is correct, however, it would seem to suggest that archaeologists should think carefully about employing the Slatkin Exact test in situations where innovation or migration rates are very low. Finally, at intermediate θ values (e.g., 10.0 - 30.0), the failure rate peaks at intermediate levels of time-averaging, and then declines again). This, too, is an unexplained effect warranting future research.

In general, with long-duration assemblages, archaeologists should be careful interpreting the results of neutrality tests adopted from population genetics. The effect seen here can be summarized as: with significant time-averaging, trait frequencies generated by unbiased cultural transmission can falsely appear to be non-neutral and thus driven by bias or selection (Type 1 error). The longer the duration of an assemblage with respect to the mean trait lifetime, the larger the probability of a Type 1 error. With sufficient duration, in fact, the probability of a Type 1 error becomes virtually certain, and the Slatkin Exact test loses any discriminatory power.

In summary, if one were to employ Slatkin Exact tests to long-duration archaeological deposits, one would overwhelmingly come away with the impression that most cultural transmission was biased, either towards conformity or a pronovelty bias – regardless of the underlying process occurring during prehistory.

It is worth noting that the effect appears to work in reverse when the data are known to be generated by a biased transmission process, such as conformist transmission. Given the space constraints of a conference presentation, I leave the details of biased transmission models to another paper, but I performed a preliminary analysis of the conformist transmission model used by Mesoudi and Lycett [24].[12] At even reasonably low levels of conformity (only a 10% chance of a copying event being conformist rather than unbiased), Slatkin Exact tests upon unaveraged samples universally rejected the H_o of neutrality – virtually all samples concentrated themselves into the upper tail region. At very high levels of time-averaging, however, the null hypothesis began to be accepted, although at low frequency.

I conjecture that at low levels of conformist bias, and in relatively long duration assemblages, there can be equifinality between unbiased and conformist models, with respect to tests which fit trait counts to Equation 2. This equifinality could be resolved, however, with a sufficient number of independent samples Slatkin Exact tests, since even a time-averaged unbiased process generates $\mathbb{P}(E)$ in both halves of the ESF distribution, whereas a pure conformist process under time-averaging may generate $\mathbb{P}(E)$ values into the non-tail region of the ESF, but only from above, and not symmetrically around the center of mass of the ESF.

B. Trait Richness

The relationship between the mean number of traits in samples (of size 100), and assemblage duration is shown in (K_n) Figure 3. This is simply a time-rescaled version of Figure 1, of course.

At very low values of the population innovation rate (e.g., 0.1 - 1.0), there is often no time-averaging effect on the number of traits we expect to see in a sample, except at the largest durations. At intermediate (but quite high, compared to the usual values in genetic applications) levels of innovation, trait richness is elevated almost immediately once time-averaging reaches the mean trait lifetime, and increases at an accelerating rate thereafter. As θ values rise, the interaction between innovation rate and the finite size of samples taken dominates the picture, such that with assemblage durations between 100 and 1000 times the mean trait lifetime, the sample is "saturated" with variants: almost every element of a sample represents a different variant. This implies another of archaeology's repeated "sample size" effects: sample sizes must increase with increased assemblage duration and with the anticipated innovation rate of a prehistoric population, in order to accurately measure K_n values, for comparison across assemblages and sites.

C. Diversity Measures

Much of the current effort in distinguishing biased and unbiased transmission models rely upon trait evenness and the shape of frequency distributions, given Alex Bentley's application of power-law distributions to both ancient and contemporary data sets [4–9, 18, 19]. And one of the ways that unbiased and "conformist" models of cultural transmission differ is in the expected amount of variation. Compared to unbiased transmission, conformism of even a mild degree tends to strongly concentrate adoption onto a very small number of traits. It is difficult, however, to interpret the absolute number of traits (K_n) without knowledge of the population size, so Kohler [21] employed diversity measures instead in his classic examination of conformist transmission in Southwest ceramics.

The most commonly used measure in the archaeological literature on cultural transmission is t_f (Equation 5), since it is related to Wright's original measures of heterozygosity and thus associated directly with the historical development of the Wright-Fisher model. But it is useful to normalize the results of t_f between 0 and 1 so that we can compare different levels of theta and assemblage durations easily, in the same way that statisticians occasionally employ coefficients of variation or normalize covariances into correlation coefficients. Equation 6 does exactly this, and is called the "index of qualitative

---

[12] Alex Mesoudi graciously provided the source code for simulations used in their paper, and I constructed a **TransmissionFramework** module which replicated their exact algorithm for conformity.

variation" [42].

Figure 4 displays the relationship between the IQV for samples of size 100, and time-averaging duration scaled by mean trait lifetime, as before. IQV values range from 0.0, if only a single trait occurred within a sample (which happens in simulations with very low innovation rates), through 1.0, which indicates that traits are perfectly evenly distributed within a sample. Even at the highest innovation rate studied, values of 1.0 were not seen in *unaveraged* samples from the simulation runs. It is apparent that time-averaging can yield greater evenness among trait frequencies, although the plateau in IQV values seen at high θ and high assemblage duration is a function of the saturation of K_n in a finite sample seen above. At very low innovation rates ($\theta < 1.0$), time-averaging in contrast seems to have little effect on the dispersion of trait frequencies, with one or a very few traits always dominating a sample.

In between, when innovation rates are sufficient to guarantee at least one innovation on average per model generation ($\theta = 1.0$) but fewer than 10, there is non-monotonic behavior apparent in the IQV index. For example, at $\theta = 2.0$, time-averaging has no effect on IQV until duration is 10 times the mean trait lifetime ($\bar{t}$), at which point assemblages begin to appear *less even* in frequency distribution, until about 100 times the mean trait lifetime, when evenness begins to steadily increase. This effect is interesting, since it suggests that we cannot easily compare diversity indices between assemblages unless we both control for duration, and understand something about innovation rates. This suggests that we need better methods of estimating those rates, since θ estimates from trait count samples have extremely large variance, as well as bias [16].

V. DISCUSSION AND CONCLUSIONS

When we examine the effects of time-averaging on the sample properties of unbiased transmission, using the mean lifetime of traits as our fundamental time scale, several lessons for practical applications seem to emerge.

First, it appears that assemblages with modest amounts of temporal aggregation display few of the distributional alterations that characterize long-duration assemblages. Statistical tests of neutrality, diversity measures, and thus arguments based upon them, can probably be used with little adaptation. One caveat to this occurs in populations with very low innovation rates, where sampling in finite populations can give highly variable results given the slow rate at which innovations accumulate (and thus the rate at which any variation in traits can be measured).

Second, for assemblages that have a duration longer than the mean trait lifetime, it is important to measure and control for the relative duration of assemblages when comparing statistical results across samples. Without doing so, we cannot interpret relative differences of diversity indices or trait richness values as indicative of different modes of transmission, or differences in innovation rate.

The rough distinction drawn here between "short" and "long" duration assemblages, scaled with respect to mean trait lifetime, needs additional research to fully operationalize. This choice of time scale requires us to link the abstract "traits" of our models, to the units of measurement employed by archaeologists. This is not a trivial task, and is one that is rarely even attempted in archaeological applications of cultural transmission models. Instead, we frequently act as if "traits equal types," despite occasional acknowledgement of the difference. The traits modeled in WF-IA are clearly not the equivalent of archaeological classes, but there is no reason why our transmission models could not incorporate more complex "traits," and I am certainly not the first to suggest such additions to cultural transmission models [25].

But there is a metatheoretical issue here as well. We do not study DNA sequences, and we have no direct empirical access to the information prehistoric populations were learning, teaching, and imitating. We will never find "units of transmission" in any empirical sense for archaeological applications of cultural transmission models. Long ago, Osgood [29] recognized that when anthropologists study the ideas held within a social group under study, what is actually being studied are the ideas we *construct* about the ideas individuals in other cultures may have had. Dunnell [11] systematized this distinction, pointing out that we always operate with analytic classes whose construction is done by archaeologists, for archaeological purposes. These classes serve as a "filter" by which we detect patterns in artifact assemblages, which reflect patterns in the information which flowed within past populations. There is no "natural" set of classes to employ in studying cultural transmission, but we often forget to incorporate this fact into our analyses.

With respect to the application of cultural transmission models in archaeology, very few studies have examined the role of changing the level of analytic classification on the measurement of cultural transmission (although see [22, 36]). Because we need to link model "traits" and their quantitative characteristics to archaeological units in order to create empirically sufficient explanations, the direct representation of analytic classification into our cultural transmission models seems essential. Doing so will allow us the means to connect the abstract quantities we model, including mean trait lifetimes and innovation rates, to patterns in the archaeological record in ways that allow us to better fit our increasingly realistic models of social learning and transmission to diachronic and time-averaged archaeological data.

[1] Abramowitz, M. and L. Stegun
1965. *Handbook of Mathematical Functions*. Dover, New York.

[2] Bailey, G.
1981. Concepts, time-scales and explanations in economic pre-history. *Economic Archaeology: towards an integration of ecological and social approaches*, Pp. 97–118.

[3] Bailey, G.
2007. Time perspectives, palimpsests and the archaeology of time. *Journal of Anthropological Archaeology*, 26(2):198–223.

[4] Bentley, R.
2007. Fashion versus reason-then and now. *ANTIQUITY-OXFORD-*, 81(314):1071.

[5] Bentley, R., M. Hahn, and S. Shennan
2004a. Random drift and culture change. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1547):1443.

[6] Bentley, R., M. Madsen, and P. Ormerod
2009. Physical space and long-tail markets. *Physica A: Statistical Mechanics and its Applications*, 388(5):691–696.

[7] Bentley, R. A., M. W. Hahn, and S. J. Shennan
2004b. Random drift and culture change. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 271(1547):1443–1450.

[8] Bentley, R. A., C. P. Lipo, H. A. Herzog, and M. W. Hahn
2007. Regular rates of popular culture change reflect random copying. *Evolution and Human Behavior*, 28:151–158.

[9] Bentley, R. A. and S. J. Shennan
2003. Cultural transmission and stochastic network growth. *American Antiquity*, 68(3):459–485.

[10] Binford, L.
1981. Behavioral archaeology and the "pompeii premise". *Journal of Anthropological Research*, Pp. 195–208.

[11] Dunnell, R. C.
1971. *Systematics in prehistory*. New York: Free Press.

[12] Dunnell, R. C.
1978. Style and function: A fundamental dichotomy. *American Antiquity*, 43(2):192–202.

[13] Durrett, R.
2008. *Probability models for DNA sequence evolution*, second edition edition. Springer.

[14] Ewens, W.
1964. The maintenance of alleles by mutation. *Genetics*, 50(5):891.

[15] Ewens, W.
1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112.

[16] Ewens, W. J.
2004. *Mathematical Population Genetics, Volume 1: Theoretical Introduction*, 2nd edition. Springer.

[17] Grayson, D. and F. Delpech
1998. Changing diet breadth in the early upper paleolithic of southwestern france. *Journal of Archaeological Science*, 25:1119–1129.

[18] Hahn, M. W. and R. A. Bentley
2003. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society Biology Letters B*, 270:S120–S123.

[19] Herzog, H. A., R. A. Bentley, and M. W. Hahn
2004. Random drift and large shifts in popularity of dog breeds. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 271:S353–S356.

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

[21] Kohler, T. A., S. VanBuskirk, and S. Ruscavage-Barz
2004. 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.

[22] Lipo, C. P.
2001. *Science, Style and the Study of Community Structure: An Example from the Central Mississippi River Valley*. Oxford: British Archaeological Reports, International Series, no. 918.

[23] Lyman, R.
2003. The influence of time averaging and space averaging on the application of foraging theory in zooarchaeology. *Journal of Archaeological Science*, 30(5):595–610.

[24] Mesoudi, A. and S. J. Lycett
2009. Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior*, 30:41–48.

[25] Mesoudi, A. and M. J. O'Brien
2008. The learning and transmission of hierarchical cultural recipes. *Biological Theory*, 3(1):63–72.

[26] Neiman, F. D.
1990. *An Evolutionary Approach to Archaeological Inference: Aspects of Architectural Variation in the 17th Century Chesapeake*. PhD thesis, Yale University.

[27] Neiman, F. D.
1995. Stylistic variation in evolutionary perspective: Inferences from decorative diversity and interassemblage distance in illinois woodland ceramic assemblages. *American Antiquity*, 60(1):7–36.

[28] Olszewski, T.
2011. Remembrance of things past: modelling the relationship between species' abundances in living communities and death assemblages. *Biology Letters*, 8(1):131–134.

[29] Osgood, C.
1951. Culture: Its empirical and non-empirical character. *Southwestern Journal of Anthropology*, 7:202–214.

[30] Peng, B. and M. Kimmel
2005. simupop: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687.

[31] Peng, B., M. Kimmel, and C. Amos
2012. *Forward-Time Population Genetics Simulations: Methods, Implementation, and Applications*. Wiley-Blackwell.

[32] Premo, L. and J. Scholnick
2011. The spatial scale of social learning affects cultural diversity. *American antiquity*, 76(1):163–176.

[33] Schiffer, M.
1987. *Formation Processes of the Archaeological Record*. Albuquerque: University of New Mexico Press.

[34] Slatkin, M.
1994. An exact test for neutrality based on the ewens sampling distribution. *Genetical Research*, 64(01):71–74.

[35] Slatkin, M.
1996. A correction to the exact test based on the ewens sampling distribution. *Genetical research*, 68(03):259–260.

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

[37] Stein, J.

1987. Deposits for archaeologists. *Advances in archaeological method and theory*, 11:337–395.

[38] Tomašových, A. and S. Kidwell
2010a. The effects of temporal resolution on species turnover and on testing metacommunity models. *The American Naturalist*, 175(5):587–606.

[39] Tomašových, A. and S. Kidwell
2010b. Predicting the effects of increasing temporal scale on species composition, diversity, and rank-abundance distributions. *Paleobiology*, 36(4):672.

[40] Wakeley, J.
2008. *Coalescent Theory*. Harvard University Press.

[41] Watkins, J.
2010. Convergence time to the ewens sampling formula. *Journal of Mathematical Biology*, 60:189–206.

[42] Wilcox, A.
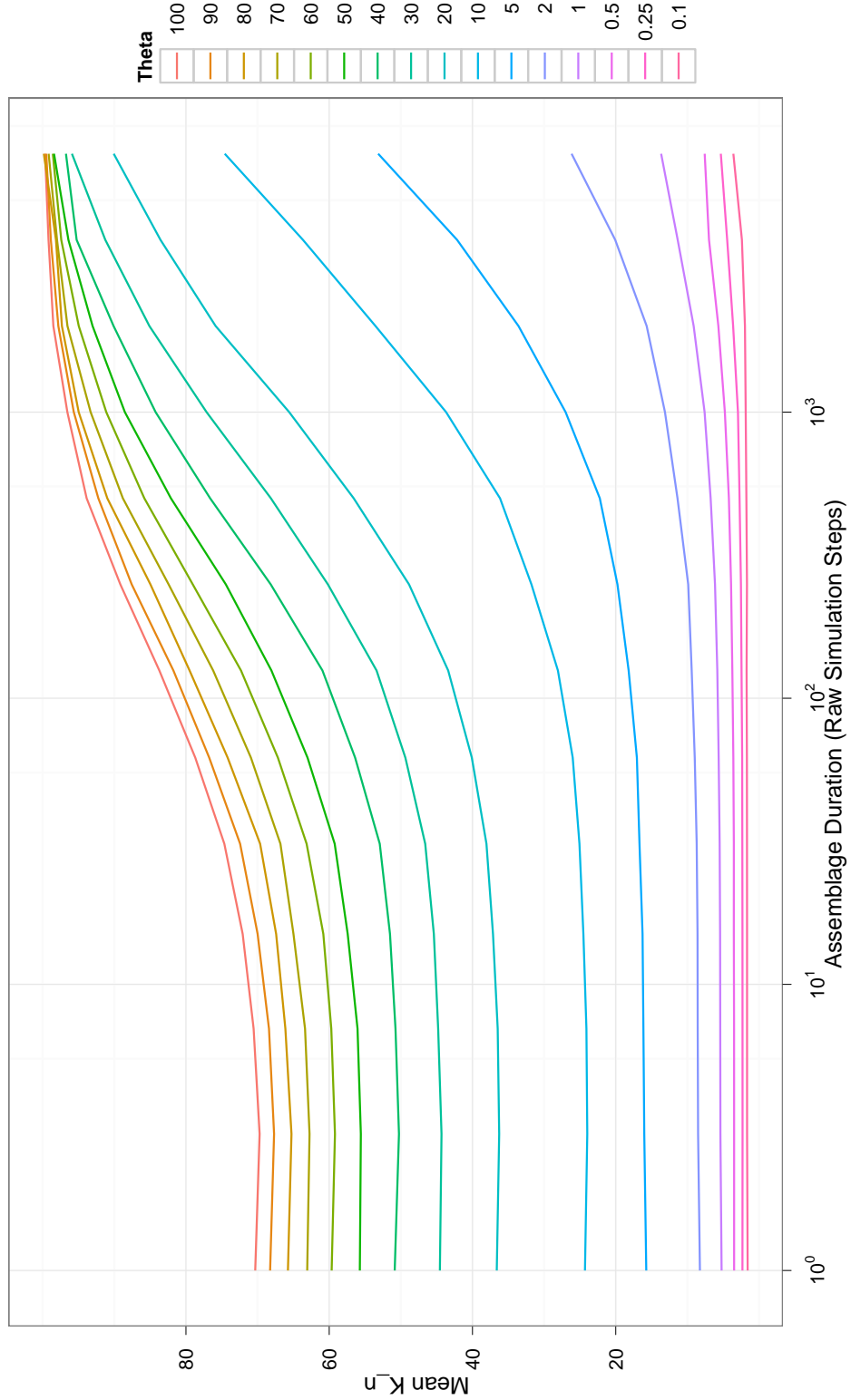1973. Indices of qualitative variation and political measurement. *The Western Political Quarterly*, 26(2):325–343.

FIG. 1. Mean value of K_n for time-averaged samples, plotted against assemblage duration in simulation steps, for each level of θ in the study. Note that the "onset" of time-averaging effects (as measured by increased K_n), is quite gradual at low θ, while high innovation rates display increased richness with very minor amounts of time-averaging.
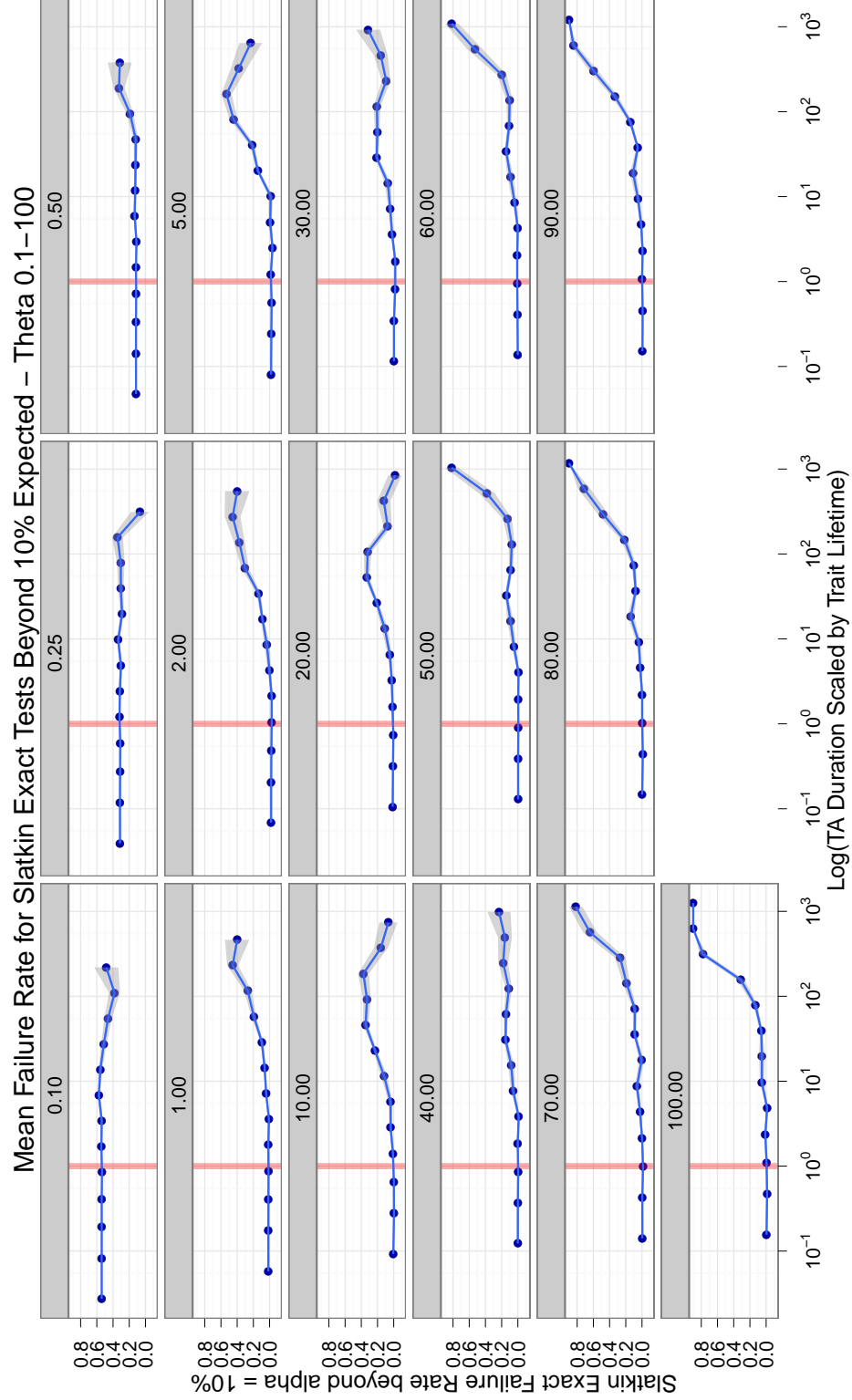
FIG. 2. Slatkin Exact test failure rate (above the expected 10% given two-tailed test with $\alpha = 0.10$, plotted against time-averaging duration scaled by mean trait lifetime, for each level of θ in the simulation study. The red vertical line indicates the mean trait lifetime for that θ value, and the shaded region encompasses the standard error of the estimates for mean failure rates at each duration.
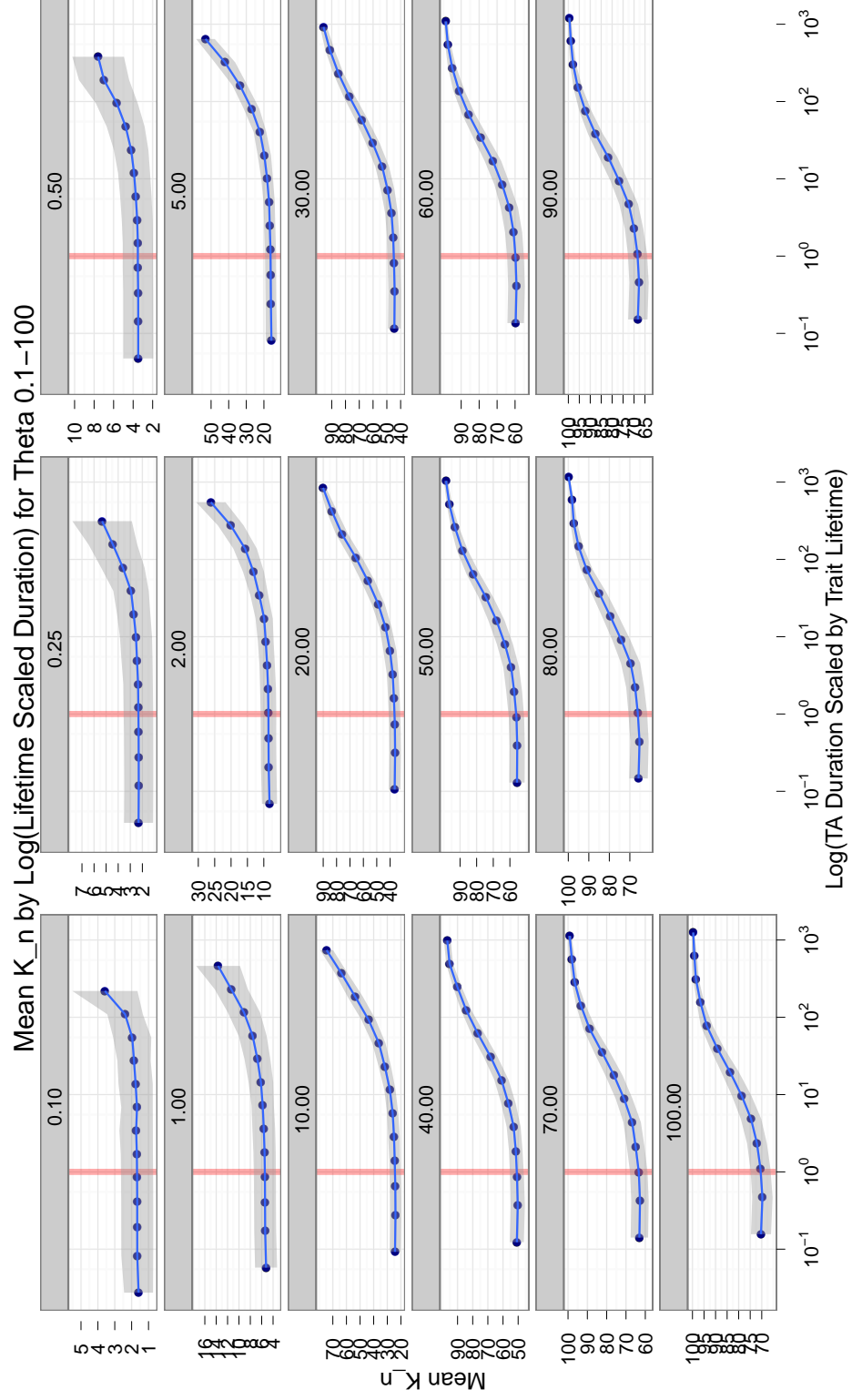
FIG. 3. Mean K_n for time-averaged samples, plotted against time-averaging duration scaled by mean trait lifetime, for each level of θ in the simulation study. The red vertical line indicates the mean trait lifetime for that θ value.
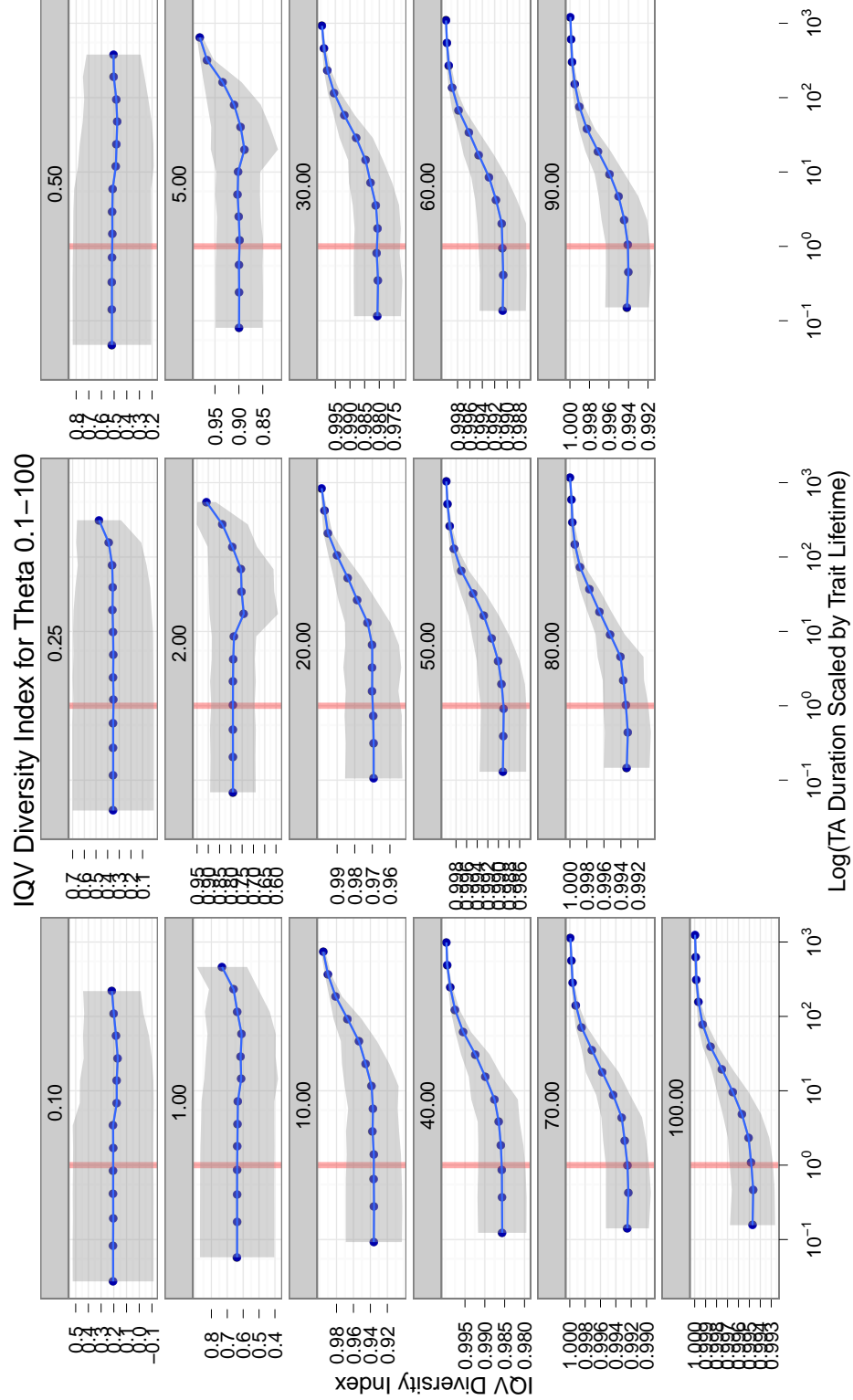
FIG. 4. IQV diversity index, derived from samples of size 100, plotted against time-averaging duration scaled by mean trait lifetime, for each level of θ in the simulation study. The red vertical line indicates the mean trait lifetime for that θ value.