ELSEVIER

Evolution and Human Behavior 25 (2004) 305–326

Evolution and Human Behavior

Cultural evolution in laboratory microsocieties including traditions of rule giving and rule following

William M. Baum*, Peter J. Richerson, Charles M. Efferson, Brian M. Paciotti

Department of Environmental Science and Policy, University of California, Davis, Davis, CA, USA

Received 12 June 2003; accepted 12 May 2004

Abstract

Experiments may contribute to understanding the basic processes of cultural evolution. We drew features from previous laboratory research with small groups in which traditions arose during several generations. Groups of four participants chose by consensus between solving anagrams printed on red cards and on blue cards. Payoffs for the choices differed. After 12 min, the participant who had been in the experiment the longest was removed and replaced with a naïve person. These replacements, each of which marked the end of a generation, continued for 10–15 generations, at which time the day's session ended. Time-out duration, which determined whether the group earned more by choosing red or blue, and which was fixed for a day's session, was varied across three conditions to equal 1, 2, or 3 min. The groups developed choice traditions that tended toward maximizing earnings. The stronger the dependence between choice and earnings, the stronger was the tradition. Once a choice tradition evolved, groups passed it on by instructing newcomers, using some combination of accurate information, mythology, and coercion. Among verbal traditions, frequency of mythology varied directly with strength of the choice tradition. These methods may be applied to a variety of research questions.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Experiments; Cultural evolution; Rules; Reinforcement; Outcomes; Evolutionary psychology

1. Introduction

In their book, *Culture and the Evolutionary Process*, Boyd and Richerson (1985) summarized much of the theoretical work and field research that has addressed cultural

* Corresponding author. 611 Mason #504, San Francisco, CA 94108, USA. Tel.: +1-415-345-0050.
E-mail address: wbaum@sbcglobal.net (W.M. Baum).

1090-5138/04/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.evolhumbehav.2004.05.003

evolution (see also Cavalli-Sforza & Feldman, 1981; Durham, 1991). Almost all of the laboratory research on social learning (e.g., Heyes & Galef, 1996; Rosenthal & Zimmerman, 1978) focuses on the individual-level mechanisms by which one organism acquires behavior from another. Remarkably little experimental research addresses the evolutionary question of how these individual-level mechanisms contribute to phenomena at the population level. Approaching cultural traditions and their change over time from a population-level perspective, anthropologists, historians, and other social scientists have sometimes discussed processes at the individual level but have often been unconcerned with or hostile to attempts to generalize about cultural evolutionary processes.

Perhaps the single most neglected field of empirical investigation in evolutionary social science is the study of the processes of cultural microevolution. How do individual-level processes, such as the choices individuals make when they imitate or take instruction from others, contribute to incremental changes in cultural traditions at the population level? Such microevolutionary studies are the bedrock of our understanding of organic evolution. Endler (1986) and Brandon (1990) provided excellent discussions of the centrality of studying microevolutionary processes in organic evolution. Some traditions of research in the social sciences approximate organic microevolutionary studies; examples include those of sociolinguists (Labov, 2001; Thomason, 2001), investigations of the diffusion of innovations (Rogers & Shoemaker, 1971), Martindale's (1975, 1990) dissections of aesthetic evolution, and certain studies of the sociology of religion (Roof & McKinney, 1987; Stark, 1997; Wilson, 2002). With the exception of Wilson's (2002) work, none of these studies derives from a sophisticated theory of cultural evolution. Only a handful of studies have so far connected the emerging theory of cultural evolution to empirical cases (Henrich, 2001; Hewlett & Cavalli-Sforza, 1986; McElreath, submitted).

Field investigations of cultural microevolution are limited by the complexity of field situations. Although cultural evolution is relatively rapid, it is often too slow to be observed during the period of one research grant. Key situations may be difficult to observe, as when, in language evolution, the presence of observers inhibits people from speaking their normal dialect. Deliberate control of critical variables is normally impossible. When more controlled studies are necessary to settle questions in evolutionary biology, experiments on caged populations of *Drosophila*, test tubes of *Escherichia coli*, and other laboratory systems are pressed into service. For cultural evolution, Jacobs and Campbell (1961) pioneered an analogous technique.

Jacobs and Campbell (1961) began a tradition of an exaggerated visual illusion in a small group of subjects by composing the initial group primarily of stooges who publicly reported exaggerated estimates. The naïve members went along with the stooges initially, but as the stooges were replaced periodically with new naïve subjects, and then initially naïve subjects with new naïve subjects, the magnitude of the illusion reported gradually decreased to normal levels. The exaggeration persisted, however, for several replacements (''generations'') beyond the elimination of all stooges, suggesting some tendency for the tradition, once established, to be transmitted.

Our main argument in this paper is that the laboratory microsociety experiment is a flexible tool for examining many aspects of cultural transmission under controlled conditions. We collected observations of behavior to test whether this experimental design does lead to real cultural evolution and examined a simple but important issue as an example.

The theoretical question we investigated here is the nature of the individual-level mechanisms that shape cultural evolution. Unlike genetic transmission, cultural transmission allows individuals selectively to acquire variant behavioral patterns that they observe in others and to impose their own innovations on the final patterns they exhibit. One school of evolutionary psychologists (Atran et al., 2002; Boyer, 1994; Sperber, 1996; see also Tooby & Cosmides, 1989) argues that culture is driven—these authors sometimes seem to imply almost entirely driven—by innate, information-rich, psychological structures that strongly bias behavior, so strongly that perhaps "cultural transmission" would be an unnecessary concept. Boyd and Richerson (1985) argued, in contrast, that cultural transmission not only occurs but also may be accurate and little affected by behavioral predispositions, at least in the short term. Furthermore, Boyd and Richerson proposed that culture is often constrained less by powerful, fixed predispositions than by the success or failure of cultural practices in particular environments; that is, cultural practices are often shaped by their consequences. These may take the form of natural selection acting on cultural variation but more often consist of reinforcement of individual behavior (cf. Skinner, 1953, 1981), which Boyd and Richerson referred to as bias and guided variation. Doubtless, predispositions affect the extent to which practices are strengthened or weakened by success or failure. These effects may be construed, however, as quantitative rather than qualitative. Cultural traditions, fixed behavioral predispositions, and environmental contingencies (acting via predispositions to find certain stimuli rewarding and others aversive) all play roles in determining individual behavior (Baum, 2003). Progress requires assessing the relative strengths of these effects and understanding how they interact. Evolutionary models show that even if natural selection of cultural practices and selection of individual behavior by its consequences are weak at the individual level, they may still be powerful at the population level. Individuals aggregate in populations and cultural traditions cumulate through time, much as gene frequencies are modified over the generations by generally weak forces like natural selection. Consequently, a need exists in the social sciences for empirical methods that simultaneously address both individuals and populations. The methods we describe below meet this need.

To demonstrate the feasibility of studying this question under controlled conditions, we presented our groups (i.e., microsocieties) with a simple choice task: deciding which of two types of anagrams to solve. Solving one type of anagram had a high immediate payoff, which we hypothesized would tap a behavioral predisposition to take the choice with the highest payoff. The other choice, under some experimental conditions, paid better in the long run. If the members of the microsociety can detect the environmental contingencies, they will make more money by picking the counterintuitive choice. Finally, new members of the microsociety may learn from experienced members. In this series of experiments, we made the choice task a collective one; the four members of the microsociety had to agree on the choice by consensus. This structure was designed to encourage socialization of new members by experienced members. We adjusted the procedures for the baseline experiment roughly to equalize the effects of tradition formation, behavioral predispositions, and environmental contingencies. The simplicity of the choice our participants made led to evolution rapid enough to be observed in a few laboratory generations. These experimental sessions were then used as a point of departure for experiments designed to test hypotheses about what factors affect the relative importance of the three factors and their interaction. Here we test two basic expectations: (1) that the effect of environmental contingencies depends on the amount of reinforcement and (2) even when differences in contingencies are too weak to be reliably detected by individuals, adaptive traditions can arise in populations.

We also recorded simple ethnographic information on each experimental microsociety. Mainly we coded the talk that went on among microsociety members. This allowed consideration of how decisions were made and traditions transmitted. It also allowed examination of the content of traditions, for example, whether they were based on a correct or mythical interpretation of the environmental contingencies. We also expected that in societies of only four people, random differences between groups would be important. If so, and if cultural transmission is important, cultural diversity should arise between groups, giving us another means to be certain that true traditions exist in the microsocieties. Accordingly, the baseline and experimental conditions were each replicated several times.

The only other attempts to use laboratory microsocieties to study cultural evolution we know of were performed by Insko, Gilmore, Moehle, et al. (1982), Insko, Gilmore, Drenan, et al., (1983), Insko, Thibaut, et al. (1980), Schotter (2003), Zucker (1977), and Monestes (personal communication). Laland and Plotkin (1990, 1992) studied animal social learning using similar experimental procedures. For example, Insko et al. (1983) studied trios of groups (“villages”) making and trading origami products. Each group contained four persons, and about every 20 min a member in each group was replaced with someone naïve. Because the groups themselves interacted, and one group was more powerful than the other two, the focus of the experiment was largely on evolution of differences among the groups. The experimenters recorded the tendency of the three experienced subjects to instruct the newcomer, however, by coding “the number of complete statements made by each subject relating to strikes, slowdowns, and sabotage of already completed products” (p. 983) and by rating “each subject in each generation for the amount of task-related verbal activity” (p. 983). The verbal activity recorded included “directive statements relevant to production, strikes, negotiator selection, relations with other groups, and so on” (p. 983). Thus, they were able to document both increased production and reliable maintenance of verbal behavior across generations.

Directive statements of the sort coded by Insko et al. (1983) are known by behavior analysts as *rules*. A rule is a verbally generated stimulus that changes the listener’s behavior with respect to choice alternatives having different long-term and short-term consequences (Baum, 1995, 1994/2003). The changes in the listener’s behavior may be immediate or follow after some delay. One essential part of any culture is its rules, that is, the statements members of a cultural group make to one another that encourage behavior beneficial to the member or the group in the long run (Baum, 2000). Sociologists and anthropologists often call such rules norms or institutions. Correct behavior with respect to rules (or norms or institutions) is maintained by consequences (reinforcers for correct behavior, punishers for incorrect) delivered by members of the group. Rules may even be maintained in part because conformity to rules itself becomes rewarding. One important function of a rule is to facilitate avoidance of behavior that pays off better in the short run but is worse in the long run (e.g., choosing a diet including vegetables over one consisting of junk food or choosing to marry someone unrelated over a relative; Baum, 2000).

The present research was intended to extend the experiments of Jacobs and Campbell (1961) by incorporating some features of the experiment by Insko et al. (1983). We aimed to observe the evolution of traditions, both nonverbal and verbal, that persisted for many generations. We utilized a task of less complexity than origami folding, but we incorporated monetary payments for success and coded both the group's strategy with respect to the task and their verbal behavior with respect to the task directed toward one another and toward the newcomer (i.e., their rule giving) in each generation. The groups chose between solving anagrams printed on red cards and blue cards, each color associated with a different set of payoffs. Choosing a *red* anagram resulted in payment to each group member of 10 cents on solution, whereas choosing a *blue* anagram resulted in payment of 25 cents followed by a time-out during which no anagrams could be solved. The duration of the time-out determined whether more could be earned by choosing red or by choosing blue. Because anagrams could be solved about once a minute, if the time-out lasted 1 min, on average choosing blue would earn more, whereas if the time-out lasted 2 or 3 min, on average choosing red would earn more, even though the immediate payoff for blue was greater. Because anagrams were solved in variable times, these relations were usually difficult to detect and required both discussion among the group and explanation to the newcomer to the group.

2. Methods

2.1. Subjects

Participants were 278 students at the University of California, Davis (60 male, 218 female), all between the ages of 17 and 31, the majority being 19. Most of the participants were recruited from the psychology department's subject pool and received course credit for participating in addition to the money they earned in the experiment. Some were recruited by an advertisement in the campus newspaper and received 5 dollars for participating, in addition to the money they earned in the experiment.

2.2. Materials

Anagrams of five-letter words were prepared by choosing all five-letter words from the two highest frequency categories of the Thorndike-Lorge (1944) list of English words. Such anagrams require about a minute on average for an individual college student to solve (Mayzner & Tresselt, 1958). A computer program scrambled the letters in one of the orders: 14253, 25314, 31425, 42513, or 53142. These are estimated to be relatively difficult orders (Ireland-Galman, Padilla, & Michael, 1980). The program selected them approximately equally often. The final list contained 375 anagrams. These were printed in the center of 4- by 6-in. index cards in a 16-point bold font, one per card. Two complete sets were printed, one on red cards and the other on blue cards. Thus, on average red and blue anagrams were of the same difficulty, although some individual anagrams are harder to solve than others.

The experimenter, participants, and one or two coders sat around a table. The groups never showed any sign that they were uneasy about being observed and typically ignored the coders the whole time. The anagrams and baskets of coins (quarters and dimes) were in a cardboard box behind the experimenter, out of sight of the participants. Two standard interval timers sat on the table facing the experimenter and were unavailable to the participants. A separate office was used to brief and debrief the participants.

2.3. Procedure

After having given consent, a participant was given a set of instructions to read. These described the procedure in general terms, mentioning that money would be given for solving anagrams, that they would choose between anagrams on red and blue cards, that occasional time-outs would be called, that new people would occasionally enter the group, and that talking was permitted at all times. Questions before or during the experiment were answered by repeating phrases from the instructions. Having read the instructions, the participant received a recording sheet and waited to be called into the experiment.

The day's session began as soon as four participants arrived. As soon as they were settled at the table, the experimenter started an interval timer set for 12 min (the generation interval) and began the first trial by saying, "Choose, red or blue." This invariably caused consultation among the group. They would usually agree on a color, but the experimenter made sure on every trial by asking, "Does everyone agree?" and waiting to hear or see an indication from every participant. The experimenter then announced the color chosen, turned to the box behind him, attached an anagram of that color to a clipboard, and placed the clipboard on the table among the participants, who announced their solution as soon as they discovered it. If the group chose red, each participant was given a dime, and the experimenter again said, "Choose." If the group chose blue, each participant was given a quarter, and the experimenter said, "We need to take a time-out," and started an interval timer of the appropriate duration for the condition. Depending on the condition, time-outs were 1, 2, or 3 min in duration and fixed for the day's session. At the end of the time-out, the experimenter again said, "Choose." Participants recorded the group's choices and earnings on their recording sheets. These records were used afterwards as the data on choice. Recording errors were easily resolved by comparing among the four participants' sheets and by comparing recorded earnings with actual earnings.

At the end of the 12-min generation interval, the experimenter said, "It is time for the next person." If the group was in the middle of solving an anagram, they were allowed to complete it and were given the monetary payoff. If they had chosen blue, the next generation began with a time-out. If they were in the middle of a time-out, the time-out interval timer was halted and restarted at the beginning of the next generation. The participant with the lowest number in the group was asked to collect his or her earnings and recording sheet and was led back to the office for debriefing. The next participant was then led to the experimental room. As soon as the new person was settled, the experimenter started the

12-min generation interval and the time-out timer, if appropriate, either saying, "Choose" or "We need to take (or finish) our time-out." Generations continued until no more participants were available, the number of generations varying from 10 to 15, depending on how many participants showed up. When no more participants were available, the experimenter announced that the experiment was over, and the four remaining participants were debriefed.

Each participant filled out a questionnaire after leaving. Questions included what the participant thought the experiment was about, whether any leaders emerged within the group, whether the participant had heard anything about the experiment beforehand, and demographic information. The participant's money was counted and these actual earnings recorded.

During the experiment, the coders counted and classified all rules, that is, all utterances that aimed to affect participants' choices. These included statements summarizing any aspects of the experimental procedure ("informative" if accurate, "mythology" if inaccurate), statements summarizing the group's choices or reasons for their choices ("informative" if accurate, "mythology" if inaccurate), and any utterances telling any of the participants how to choose without explanation ("coercive"). Rules directed at the newcomer, usually occurring at the beginning of the generation, were coded separately from those directed at the group as a whole, which usually occurred during discussions of how to choose (i.e., group strategizing). Intercoder reliability checks correlating the number of rules counted per generation between pairs of coders (out of a total of three coders) invariably resulted in Pearson coefficients of .85 or higher. Although one might worry that coding during the experiment could have allowed the coders' biases to influence the recording of the rules, we judged that unlikely because the coders had no a priori hypotheses about rule giving. Idiosyncratic biases would have reduced intercoder reliability, and we saw no evidence of that. Except two sessions with 2-min time-outs, the coders also recorded trial duration, defined as the time from the experimenter's saying "Choose" to his saying "Choose" again or "We need to take a time-out."

Two irregularities occurred that were judged unimportant. In one session with time-out of 2 min, a participant who had participated before got into the experiment. This was judged unimportant because the subject said almost nothing to any of the other participants and appeared afterwards to remember little about the earlier participation, which had been months before. In one session with time-outs of 3 min, one later generation contained only three participants and no newcomer. This appeared in no way to affect the group's choices. Other sessions with irregularities that might have affected the results were omitted. Sessions were run until each time-out condition had produced six usable sessions.

3. Results

3.1. Choice traditions

Fig. 1 shows the proportion of trials on which red was chosen, generation by generation, for all six sessions in the three time-out conditions. Comparing among the three graphs, one sees that the 3-min time-out produced the greatest uniformity across sessions, the groups all

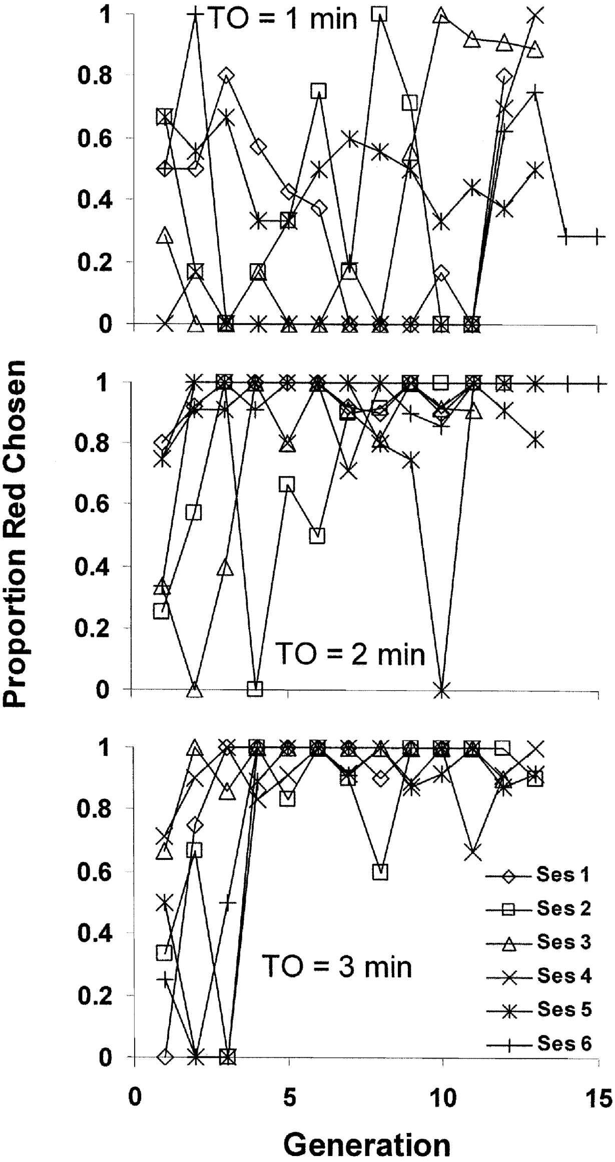

Fig. 1. Choice by generation for all six sessions at the three levels of time-out (TO) duration. Different symbols indicate the different sessions.

exhibiting a strong preference for red by the fourth generation (after three replacements). The 2-min time-out resulted in more variability across sessions, but still a general overall preference for red. Two sessions contained large shifts toward choosing blue past the fourth generation. The 1-min time-out produced an overall weak preference for blue, but little uniformity and much fluctuation in preference in most sessions.

Fig. 2 offers a possible explanation of the variation across time-out durations because it represents the groups' experiences with choosing and earning. For each generation in Fig. 1, the amount of money earned by each participant in cents is plotted against the proportion of trials on which the group chose red, the ordinate of the corresponding point in Fig. 1. The

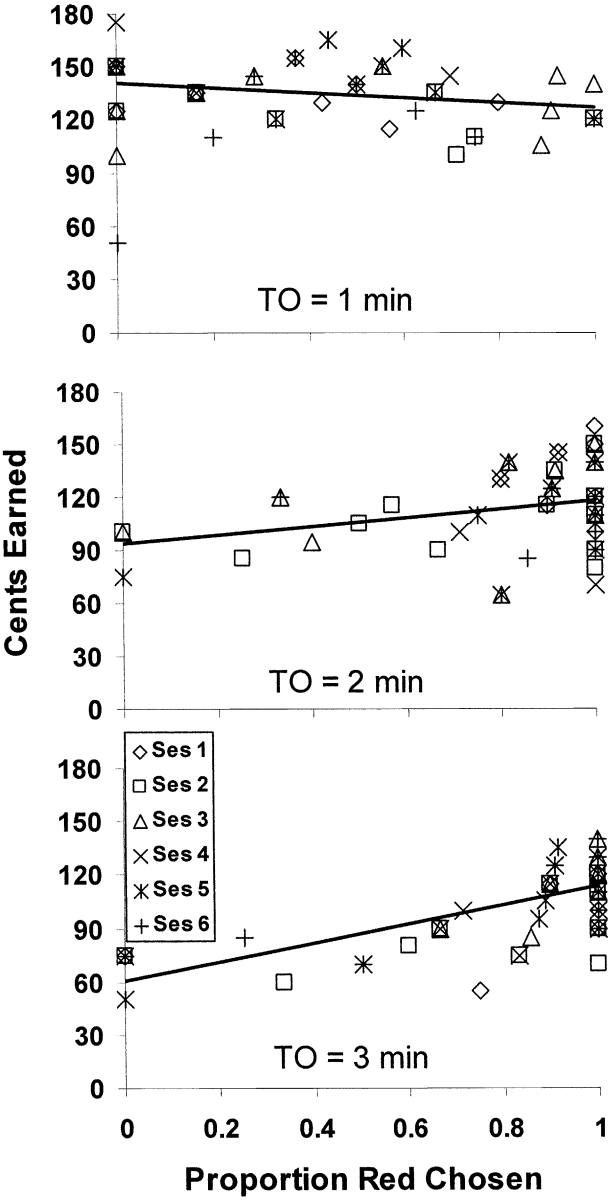

Fig. 2. Scatterplots of total earnings in cents per generation versus choice per generation in the different groups in the different sessions for the three levels of time-out duration. Each point represents one generation. The different symbols indicate the different sessions. The lines were fitted by the method of least squares.

lines were fitted to the points by the method of least squares. In all three conditions, earnings varied from generation to generation, both systematically and unsystematically. The unsystematic variation resulted from variation in the speed with which the groups solved the anagrams. Although they often solved an anagram in less than a minute, sometimes they took longer, and occasionally got "stuck" for a few minutes. This variation in speed remained about the same across the three time-out durations (means = 61, 60, and 61 s; Mdn = 55, 55, and 56 s; S.D. = 30, 27, and 31 s).

Fig. 2 shows that for the 3-min time-out, a substantial systematic relation occurred between earnings and choice, with more money earned for more choice of red, as indicated by the substantial positive slope (52 cents per choice unit) of the regression line ($r = .66$; $p < .000001$; two-tailed z test). For the 2-min time-out, the slope was still positive, but flatter (25 cents per unit of choice; $r = .28$; $p = .015$; two-tailed z test). Combined with the unsystematic variability in earnings, the flatness of the slope probably explains the lesser uniformity across sessions and the weaker preference for red shown in Fig. 1. Still, one might be surprised that an evolved preference for the red choice arose at all when reinforcement was so weak. For the 1-min time-out, although the regression line shows a slight negative slope (− 14 cents per unit of choice), favoring choice of blue, the unsystematic variability largely obscures the systematic relation ($r = -.25$; $p = .025$; two-tailed z test). The negative slope explains the overall weak preference for blue shown in Fig. 1, but the flatness of the relation explains the frequent deviations toward choosing red and the lack of uniformity across sessions.

3.2. Traditionality

If we say that the results in Fig. 1, particularly with the 3-min time-out, represent the evolution of traditions of choosing (choosing red or choosing blue), but with varying strengths of tradition (sometimes weak and sometimes strong), then we might try to quantify the strength of tradition (*traditionality*). For this purpose, we applied and compared two different criteria to the choice proportions: conservative and liberal. The conservative criterion counted as exhibiting a tradition any generation in which red or blue was chosen exclusively and *at least one participant* had never experienced choice of the other color. Each generation received a score according to the number of participants who had never seen the other color, thus varying between 0 and 4. The liberal criterion resembled the conservative criterion, producing scores from 0 to 4, except that instead of requiring exclusive choice, a single choice of the nonpreferred color was allowed, provided that choice was exclusive in the preceding generation. Thus, applying the liberal criterion to Session 2 with the 3-min time-outs (squares in Fig. 1), we obtained the following scores: The third generation, in which blue was chosen exclusively, was scored a 1, the fourth generation, in which red was chosen exclusively, was scored a 1, the fifth generation, in which blue was chosen once, was scored a 2, the sixth generation (exclusive choice of red) was scored a 3, the seventh generation (one choice of blue) was scored a 4, and the eighth generation, in which blue was chosen twice, was scored a 0. Even if blue had been chosen only once in the eighth generation, it still would have been scored a zero. The liberal criterion was suggested by the participants' tendency often to socialize the newcomer about the consequences of choosing the nonpreferred color by actually choosing it once at the beginning of the generation. With time-out 1 min, this occurred in 8 out of 77 generations; with time-out 2 min, it occurred in 9 out of 75 generations; with time-out 3 min, it occurred in 13 out of 72 generations.

We tried three methods of summarizing traditionality: (1) frequency distributions of the scores; (2) averaging across generations within each session; and (3) averaging across sessions, generation by generation. In all three methods, the conservative criterion failed to distinguish differences in strength of tradition among the three different time-out durations. The liberal criterion, however, distinguished among the three in accord with the appearances in Fig. 1, so we focused on the liberal criterion.

The third approach, averaging across sessions, generation by generation, produced the clearest results and had the advantage of charting growth as well. Fig. 3 shows mean traditionality across sessions cumulated across generations. Each point represents the average of at least five sessions. The 3-min time-out produced the strongest traditions, and the 2-min time-out produced traditions only slightly stronger than those for the 1-min time-out. Multiple regression with cumulative tradition as the dependent variable and generation and time-out as independent variables revealed a highly significant contribution of time-out ($\beta = 1.73$; $t = 4.76$; $p = .00005$). Pairwise t tests confirmed that the difference between the lines for the 1- and 3-min time-outs was statistically significant ($p = .03$). Thus, the results shown in Fig. 3 accord with the trends one can see in Fig. 1. The positive curvature of the cumulative records shows that traditionality grew gradually during the first

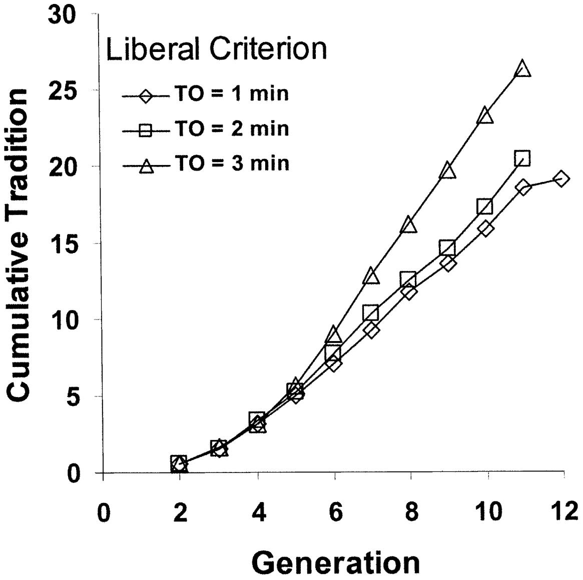

Fig. 3. Tradition averaged across sessions and cumulated across generations. Each point represents the average of six or (for later generations) five sessions. The graph shows the results of applying the liberal criterion. The different symbols indicate different time-out durations. The steeper the slope of the curve, the greater was traditionality.

five generations; the linearity after Generation 6 indicates that once established the tradition tended to remain at about the same strength.

3.3. Instructional traditions

The groups invariably began instructing newcomers, usually as soon as a new participant was brought in. Accurate instructions (e.g., "We get 25 cents for solving a blue anagram, but then we get a time-out") were distinguished from inaccurate rules, mythology (e.g., "The time-outs are 5 minutes long"), and coercion (e.g., "Just choose red"). Fig. 4 shows the frequencies of these three categories directed at newcomers. The numbers in each generation were averaged across the six (at least five, for later generations) sessions, and then those averages were cumulated across generations. The different scales on the vertical axes for instruction, mythology, and coercion reflect the different frequencies of occurrence: highest for veridical instructions, about eight times lower for mythology, and about eight times lower still for coercion. Multiple regression with cumulated instructing as the dependent variable and generation and time-out as independent variables revealed a significant effect of time-out ($\beta = 1.86$; $t = 3.45$; $p = .002$), presumably due to the greater frequency of instructing with the 3-min time-out. If instructing had evolved gradually, as did the choice traditions (Fig. 3), the lines in the top graph would have had positive curvature. Their linearity indicates that instructing began abruptly and continued at the same level. So strong was the tradition of instructing newcomers that it might be called a "social norm."

Multiple regression with cumulated mythology as the dependent variable and generation and time-out as independent variables revealed a significant effect of time-out ($\beta = 1.65$; $t = 4.95$; $p = .00003$). The frequency of mythology may be related to the strength of the choice tradition. The highest frequency occurred for the 3-min time-out, and the lowest for the 1-min time-out, with the 2-min time-out falling in between—the same ordering as in Fig. 3. As with the differences in choice tradition, only the difference in mythology between the 1-min and 3-min time-outs was statistically significant according to pairwise t tests ($p = .03$). The frequency of coercion, always low, appeared to be lowest for the 2-min time-out. None of the differences in frequency of coercion was statistically significant by multiple regression or pairwise t tests.

3.4. Group strategizing

Fig. 5 shows data from one representative session with 3-min time-out (Session 5 in Fig. 1) that illustrate a frequent coincidence we noticed between rule articulation and a shift in choice. The left vertical axis is the same as in Fig. 1: proportion of red chosen by generation. The right vertical axis shows the total number of rules (of any sort and directed at anyone) given by members of the group in a generation divided by the number of choices made by the group in that generation. It measures group strategizing, that is, rules articulated in these events as part of a group discussion of the best strategy, because its peaks resulted from rules given as proposals to the group as a whole. Its largest peak occurred in the 3rd generation, followed by smaller peaks in the 9th and 12th generations. Each peak was followed by a shift

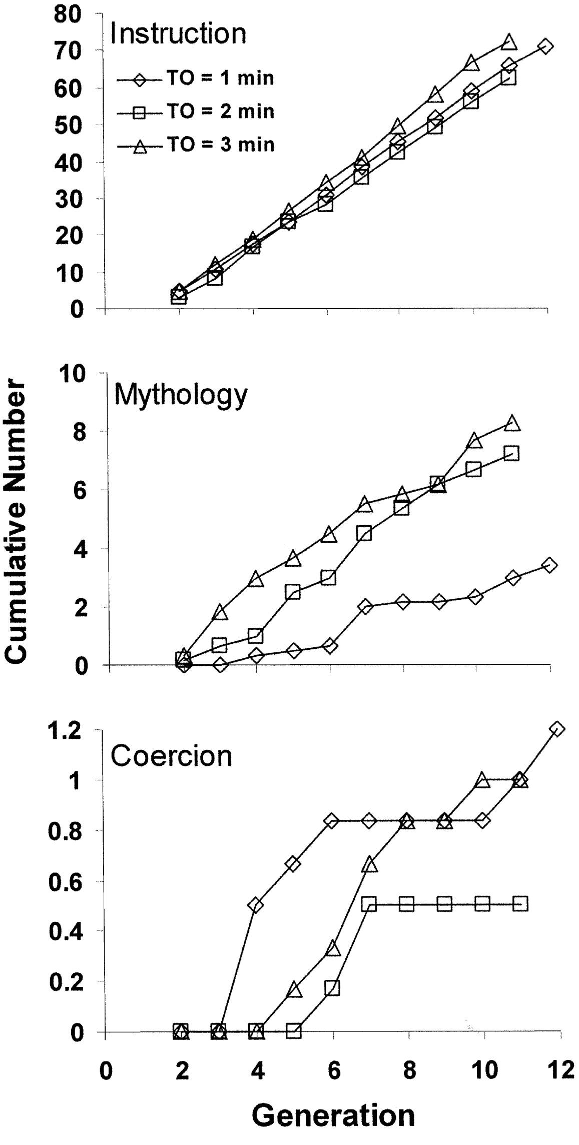

Fig. 4. Cumulative rule giving for three different types of rules directed at newcomers. Each point represents the average of six or (for later generations) five sessions. The top graph shows cumulative frequency of accurate instructing. The middle graph shows cumulative frequency of inaccurate rules (myths) that promoted the newcomer's compliance with the group's tradition. The bottom graph shows cumulative frequency of coercive statements that directed the newcomer's behavior without explanation. The different symbols represent different levels of time-out duration. The *y* axes of the three graphs were scaled according to the overall frequencies of the three types of rules.

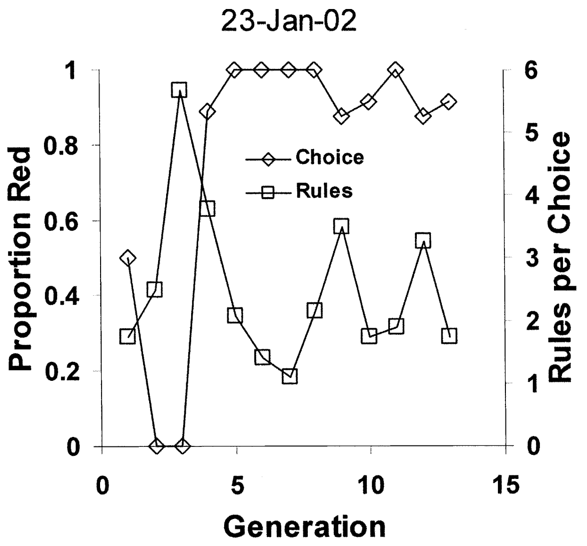

Fig. 5. A representative session with 3-min time-out illustrating the effect of group strategizing on choice. A peak in rule giving often accompanied or immediately preceded a shift in choice. Three peaks occurred in this session, each followed in the next generation by a shift in choice. The diamonds show the group's choice for each generation. The squares (y axis on the right) show the number of rules given divided by the number of choices in the generation.

in choice (toward choosing red in this example) in the next generation, and the bigger the peak the bigger the shift.

To assess the frequency of this sort of coincidence between group strategizing and shift in choice, we calculated two series of measures across generations, starting with the second ($n = 2$): We paired rule giving (rules per choice, as in Fig. 5) in generation n with the absolute change in choice from generation $n-1$ to generation n. The latter variable disregarded whether the shift in choice went toward choosing red or toward choosing blue; only its magnitude was considered. A correlation (Pearson's r) was calculated between the two series. The shifts in choice in Fig. 5, however, all follow the peaks in rule giving by a lag of one generation. Accordingly, correlations were calculated at different generation lags between the two series. The first correlation was considered to have a lag of 0, and, for example, a lag of -1 meant that shift in choice from generation n to $n+1$ was paired with rule giving in generation n, a lag of $+1$ meant that shift in choice from generation $n-2$ to $n-1$ was paired with rule giving in generation n, and so on. Fig. 6 plots the correlation coefficients as a function of the lag for the three different time-out durations. Correlations marked with an asterisk were statistically significant at the .05 level or better according to a two-tailed z test. Those marked with two asterisks were significant at the .000001 level. For 2- and 3-min time-outs, the highest correlations were for lags of 0 and -1; correlations at other lags were noticeably lower. This means that a peak in rule giving indicating group strategizing tended either to accompany (lag 0) or immediately precede (lag -1, as in Fig. 5) a shift in choice indicating a change in strategy. That the correlations were lower for the 3-min time-out

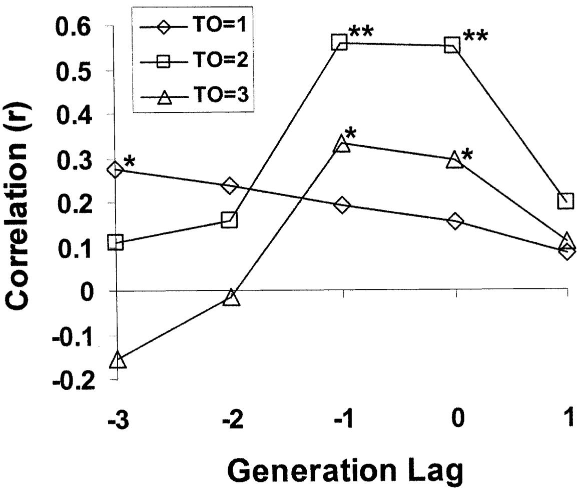

Fig. 6. Correlation (Pearson's r) between absolute change in choice and rules given per choice for several lags of rule giving relative to choice change. A lag of zero means that choice change from the preceding generation to the present generation was correlated with rule giving in the present generation. A lag of -1 means that choice change from the preceding generation to the present generation was correlated with rule giving in the preceding generation. The different symbols represent different durations of time-out. Correlations were high for lags of zero and -1, but only for time-outs of 2 and 3 min. Points marked with a single asterisk indicate correlations significant at the .05 level or better. Those marked with two asterisks indicate correlations significant at the .000001 level.

probably resulted from a preponderance of zero shifts in choice, because of the strong traditions in that condition. In contrast, the correlations for the 1-min time-out were uniformly low, indicating little effect of group strategizing on choice, in accord with the perception that traditions were weak in that condition (Fig. 1).

3.5. Generalized linear model

All the measures studied were incorporated into a generalized linear model with maximum-likelihood estimation of parameters (Burnham & Anderson, 2002). The analysis used logistic regression, in which the dependent variable was the logarithm of the ratio of probability of optimal choice (red or blue) to probability of nonoptimal choice (red or blue) in each generation. We used Akaike *c* (Burnham & Anderson, 2002) to compare among a set of possible models incorporating the following predictor variables: time-out duration (1, 2, or 3 min), rules per choice (as in Fig. 5), generation (1 to 15), conservative traditionality, and liberal traditionality. Table 1 shows the results. The full model fitted best, with time-out duration and the two measures of traditionality showing substantial parameter weights (b). The confidence intervals were relatively small. The Akaike weights (w) reveal that the full model fitted best by far; w was almost 1.0 for it and almost 0 for all the others. Comparing among rows shows that omitting any of the parameters with substantial weight (b) destroyed the model's effectiveness. Although small in magnitude, the weight for rules per choice (b_{rpc})

Table 1
Applying general linear models to choice (logarithm of the ratio of optimal choices to nonoptimal choices)

b_{int}	b_{int} CI	b_{TO}	b_{TO} CI	b_{rpc}	b_{rpc} CI	b_{gen}	b_{gen} CI	$b_{\text{tr(con)}}$	$b_{\text{tr(con))}}$ CI	$b_{\text{tr(lib)}}$	$b_{\text{tr(lib)}}$ CI	w
−0.84	[−1.34, −0.36]	0.86	[0.66, 1.07]	−0.04	[−0.12, 0.03]	−0.02	[−0.06, 0.01]	0.98	[0.75, 1.23]	0.30	[0.19, 0.41]	>0.99
−0.25		1.06		−0.23		0.05						<0.01
−0.92		1.03		−0.07				1.18				<0.01
−0.62		0.84		−0.11						0.56		<0.01
−0.71		1.25										<0.01
2.25				−0.34								<0.01
1.30						0.04						<0.01
0.72								1.36				<0.01
0.45										0.72		<0.01
0.08		1.06		−0.23								<0.01
−1.00		1.25				0.04						<0.01
−1.19		1.09						1.22				<0.01
−1.04		0.93								0.60		<0.01

The maximum-likelihood parameter estimates (b) are shown for intercept (int), time-out duration (TO), rules per choice (rpc), generation (gen), conservative traditionality [tr(con)], and liberal traditionality [tr(lib)]. Each row represents one model.

Where a parameter estimate is missing, that parameter was omitted from that model.

Confidence intervals (95% CI) are shown only for the full model.

The Akaike weights (w) appear in the last column.

in the full model is negative, consistent with the observation that instruction and strategizing sometimes included suboptimal choice.

4. Discussion

The procedure succeeded as an experimental model of cultural evolution. Traditions of two sorts evolved: traditions of choosing and traditions of rule giving or instructing. The traditions of choosing tended to reflect the choice that was optimal. Fig. 1 shows that blue was, as expected, initially attractive because this choice paid more; most sessions included at least one early generation in which blue was chosen exclusively. For the 2-min time-outs, shifts toward choosing blue occurred even after the red-choosing tradition was strong, and even for the 3-min time-out two temporary shifts toward choosing blue occurred in later generations. When more money could be earned by choosing red anagrams (2- and 3-min time-outs; Fig. 2), traditions of choosing red tended to evolve (Fig. 1), although only after some generations in some groups. When more money could be earned by choosing blue (the 1-min time-out; Fig. 2), traditions of choosing blue tended to evolve (Fig. 1). (See also a discussion below of groups that evolved counteradaptive all-blue traditions during pilot runs.) The strength of these traditions across and within groups (Fig. 1) depended on the perceptibility of the relation between choice and earnings (Fig. 2): The lower the correlation, the weaker the traditions of choosing (Fig. 3).

The generalized linear modeling (Table 1) confirmed that choice of the optimal alternative was strongly related to time-out duration and strength of tradition. It confirmed also that generation (Fig. 1) and group strategizing (rules per choice; Figs. 5 and 6) contributed to variation in choice in other ways. Figs. 1 and 3 show that choice strategy and tradition evolved as generations went on. Figs. 5 and 6 show that group strategizing shifted choice strategy (with 2- and 3-min time-outs), thus explaining some of the variation in choice, but Table 1 shows that group strategizing had little relation to optimality. The absence of a relation could have been due to absence of a lag in the generalized linear model, but additional analysis (not shown) ruled this possibility out. Instead the absence most likely arose because group strategizing represented experimentation that often resulted in sampling the nonoptimal alternative.

Although no bold generalizations about “evoked culture” (dispositions) versus “transmitted culture” (Tooby & Cosmides, 1989) can be made from one experiment, both effects were measurable in this experiment. In this experiment, however, adaptive traditions generally outweighed the intrinsic attractiveness of the 25-cent payoff for blue. In other words, both consequences—the amount of money and the duration of the time-out—mattered to choice. In behavior-analytic terms, the amount of reinforcement due to the money and the amount of punishment due to the time-out combined to determine choice. Presumably they affected the groups’ choosing as a result of their effects on the groups’ earnings (Fig. 2).

Of the criteria that we applied to the choice proportions to quantify strength of tradition, the liberal criterion, which permitted one choice of the nonpreferred color following a generation of exclusive choice, proved to be the most successful, particularly when we averaged across sessions (Fig. 3). Table 1 shows, however, that the conservative and liberal criteria were not redundant with one another; both received weight (b), and omitting either from the full model reduced the Akaike weight to zero. If one judged success in quantifying strength of tradition on the basis of agreement with qualitative impressions resulting from visual inspection of the choices (Fig. 1), then the liberal criterion was superior. The liberal criterion probably succeeded because of the groups' tendency not only to instruct newcomers but occasionally also to demonstrate the consequences of the nonpreferred choice by choosing it once. The best single quantifier was the slope of the cumulative graph (Fig. 3), with the 3-min time-out producing the highest slope, the 1-min time-out producing the lowest slope, and the 2-min time-out producing a slope between those two.

Apart from the role of making suboptimal choices to instruct newcomers, occasionally choosing the low-payoff alternative might have another function. Although seemingly maladaptive in our particular experiment, such behavior would have detected any changes that the experimenters might have made to the environmental contingencies. For example, we might have removed the time-out midway through the experiment. Since one of the main adaptive advantages of culture in theoretical models is to permit populations to respond rapidly to environmental variation while at the same time economizing on the costs of individual learning (Boyd & Richerson, 1989), a certain frequency of testing the non-preferred variant makes sense in the larger context. Variations of our experiment could be used to test this and many similar theoretical questions.

In keeping with this line of theorizing, Figs. 5 and 6 show that group strategizing produced shifts in choice policy, sometimes toward optimality, but sometimes away from optimality. Group strategizing thus explains some of the variation in choice shown in Fig. 1. Indeed, it led to a degree of instability, because traditions would sometimes be dramatically and unpredictably broken. Fig. 6 shows that group strategizing affected tradition when traditions tended to be strong, with the 2- and 3-min time-outs. The strongest effect, however, occurred with the 2-min time-out, suggesting that behavior–payoff relations of intermediate perceptibility (Fig. 2) might be most susceptible to group strategizing. No effect of group strategizing was apparent for the 1-min time-out, which also produced no strong traditions. With the 1-min time-out subjects rarely were able to reach consensus regarding rules that optimized their earnings because the unsystematic variability in earnings masked the small systematic variability (Fig. 2, top panel). These results support Boehm's (1996) ethnographic argument that collective decision making is a common force for cultural change.

Since it seemed to be a good indicator, the slope of the cumulative graph was used to assess the strength of the traditions of rule giving (Fig. 4). From the slopes of the lines in the top panel of Fig. 4, we see that a tradition of instructing the newcomer evolved quickly and reliably for all three time-out durations. The almost perfect linearity of the plots indicates that, having evolved, the tradition remained consistent. Such behavior in the group is understandable because the newcomer's cooperation was required for the others to proceed. Getting the newcomer to go along as quickly as possible allowed the group to continue earning money. Indeed, the members who undertook to instruct the newcomer invariably did so speedily. Why more instructing of the newcomers occurred with the 3-min time-out remains to be understood, although one may speculate that more rule giving is associated with stronger tradition (Fig. 3).

The results for mythology (Fig. 4, middle panel) suggest a possible relation between strength of tradition and tendency toward myth making, because the slopes of the lines in Fig. 4 are in the same order as those in Fig. 3. The three different slopes for mythology, compared with the three similar slopes for instruction (top panel), indicate that the conditions with stronger traditions (2- and 3-min time-outs) produced disproportionately more myth making. On average, 10% of rules given to newcomers were myths for both the 2- and 3-min time-outs, whereas only 4% were myths for the 1-min time-out. Possibly, rule giving under a strong choice tradition tends to become exaggerated or distorted to ensure or speed up newcomers' compliance with the group's policy. For example, the members doing the instructing would sometimes be vague about or overstate the duration of the time-outs following choices of blue. We cannot be sure that the myth giving had this effect because we made no attempt to record resistance on the part of newcomers. Whether this more frequent myth making is a reliable phenomenon and what its function might be remain to be determined by future research.

Although one might have thought that coercion would be more frequent under a strong choice tradition, this relation failed to occur (Fig. 4, bottom panel). The frequency of coercion, however, was so low that any differences were unlikely to be reliable. Future research might include operations that would increase coercion to test for this possibility. For example, given an opportunity to police group members in public-goods games, would subjects rely more on "cheap talk" or on coercion?

Our ethnographic observations lent confidence to the claim that our microsocieties exhibited true cultural evolution. For example, a tradition of using pencil and paper to try combinations of letters for more difficult anagrams developed spontaneously in several different groups. If one person tried this, other group members usually adopted the innovation for a number of generations. The technique seldom quickened solution of anagrams and hence always died out. As we predicted, variation in behavior from group to group was considerable, because individual dispositions injected much variation into group behavior. In pilot runs not used in the analyses here, we discovered that the propensity of groups to talk among themselves varied greatly in the first few generations. If early participants tended to be shy, little talk occurred until an extrovert broke the ice. To minimize this particular variation, we emphasized in the instructions that talking was permitted at all times. Even after this change, spontaneous socializing seemed to depend on chance inclusion of talkative individuals, their impact persisting after they left the group.

Often particular individuals were responsible for discovering the pattern of the environmental contingencies (i.e., the game's structure). Most participants showed relatively little curiosity about them. In two of our pilot experiments (not the final procedure, but with 2-min time-outs) groups got into a rigid pattern of choosing only blue. In both sessions, the experimenter eventually forced some red choices on the group. Despite these forced choices, the participants usually failed to notice that red choices resulted in no time-out, and the group immediately went back to choosing blue. Although at least half of participants wore watches, efforts to time time-outs and calculate their payoff consequences were relatively few and often desultory, producing the myth making shown in Fig. 4. Typically, an individual who correctly guessed the length of the time-out and its payoff consequences was responsible for initiating a tradition of accurate rule giving. Even among these individuals, a highly analytical approach was unusual. The making of new rules seemed to rely usually on an intuitive balancing of the time-out punishment against the monetary reinforcement, rather than on a more formal approach. These results agree with Henrich's (in press) contention that maintaining complex cultural traits in a population depends on rare gifted innovators. Perhaps if participants were drawn from courses in economics or engineering, rather than psychology, results would differ.

In summary, the methods described here offer the possibility of studying evolution of cultural traditions in the laboratory, even in populations as small as four. Using these results with contingencies that encourage cooperation as a baseline, we envision many possible experiments for the future. One line of investigation would be to search for evolved predispositions that are strong enough to interfere with the evolution of adaptive traditions such as dominate our results. Another would be to study the effects of different task requirements on cultural transmission. In the experiments reported here, the need to reach consensus on choice motivated socialization. If individuals could pick and solve their own anagrams, their earnings would depend less on other members' choices, and senior participants might be less inclined to socialize newcomers, that is, to engage in rule giving on a newcomer's arrival. Henrich and Gil-White (2001), however, suggest that inexperienced individuals adaptively grant prestige to the experienced to induce instruction. Individuals in our experiments often complimented successful anagram solvers; such behavior would be adaptive (i.e., reinforced) if it induced instruction from them. Still other types of tasks might be studied, such as social prisoners' dilemma or commons games. Populations might also be set to tasks that permit the cumulative evolution of complex adaptations. Longer term traditions might be allowed to evolve by recruiting the final generation in one day's experiment to return another day to continue the population (Boyd & Richerson, 1996). The possibilities seem to be limited only by researchers' imaginations.

Acknowledgments

We are indebted to J. L. Monestes of the University of Lille for suggesting the incorporation of choice into the experiment and for suggesting the use of time-outs. We are grateful also to Ed Edsten of University of California, Davis, who helped by coding rule giving in several sessions.

References

Atran, S., Medin, D., Ross, N., Lynch, E., Coley, J., Ek, E. U., & Vapnarsky, V. (2002). Folkecology, cultural epidemiology, and the spirit of the commons—A garden experiment in the Maya, 1991–2001. *Current Anthropology*, *43*, 421–450.

Baum, W. M. (1995). Rules, culture, and fitness. *The Behavior Analyst*, *18*, 1–21.

Baum, W. M. (2000). Being concrete about culture and cultural evolution. In N. Thompson, & F. Tonneau (Eds.), *Perspectives in ethology* (vol. 13, pp. 181–212). New York: Kluwer Academic/Plenum.

Baum, W. M. (2003). *Understanding behaviorism: science, behavior, and culture*. Oxford: Blackwell (Original work published 1994).

Boehm, C. (1996). Emergency decisions, cultural-selection mechanics, and group selection. *Current Anthropology*, *37*, 763–793.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Boyd, R., & Richerson, P. J. (1989). Social learning as an adaptation. *Lectures on Mathematics in the Life Sciences*, *20*, 1–26.

Boyd, R., & Richerson, P. J. (1996). Why culture is common but cultural evolution is rare. *Proceedings of the British Academy*, *88*, 73–93.

Boyer, P. (1994). *The naturalness of religious ideas: a cognitive theory of religion*. Berkeley: University of California Press.

Brandon, R. N. (1990). *Adaptation and environment*. Princeton, NJ: Princeton University Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference* (2nd ed.). New York: Springer-Verlag.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: a quantitative approach*. Princeton, NJ: Princeton University Press.

Durham, W. H. (1991). *Coevolution: genes, culture, and human diversity*. Stanford, CA: Stanford University Press.

Endler, J. A. (1986). *Natural selection in the wild*. Princeton, NJ: Princeton University Press.

Henrich, J. (2001). Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. *American Anthropologist*, *103*, 992–1013.

Henrich, J. (in press). Demography and cultural evolution, why adaptive cultural processes produced maladaptive losses in Tasmania. *American Antiquity*.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige—Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*, 165–196.

Hewlett, B. S., & Cavalli-Sforza, L. L. (1986). Cultural transmission among Aka pygmies. *American Anthropologist*, *88*, 922–934.

Heyes, C. M., & Galef, B. G. (1996). *Social learning in animals: the roots of culture*. San Diego, CA: Academic Press.

Insko, C. A., Gilmore, R., Drenan, S., Lipsitz, A., Moehle, D., & Thibaut, J. (1983). Trade versus expropriation in open groups: a comparison of two type of social power. *Journal of Personality and Social Psychology*, *44*, 977–999.

Insko, C. A., Gilmore, R., Moehle, D., Lipsitz, A., Drenan, S., & Thibaut, J. W. (1982). Seniority in the generational transition of laboratory groups: the effects of social familiarity and task experience. *Journal of Experimental Social Psychology*, *18*, 577–580.

Insko, C. A., Thibaut, J. W., Moehle, D., Wilson, M., Diamond, W. D., Gilmore, R., Solomon, M. R., & Lipsitz, A. (1980). Social evolution and emergence of leadership. *Journal of Personality and Social Psychology*, *39*, 431–448.

Ireland-Galman, M. M., Padilla, G. J., & Michael, W. B. (1980). The relationship between performance on the Mazes subtest of the Wechsler Intelligence Scale for Children—Revised (WISC-R) and speed of solving anagrams with simple and difficult arrangements of letter order. *Educational and Psychological Measurement*, *40*, 513–524.

Jacobs, R. C., & Campbell, D. T. (1961). The perpetuation of an arbitrary tradition through several generations of laboratory microculture. *Journal of Abnormal and Social Psychology 62*, 649–658.

Labov, W. (2001). *Principles of linguistic change: social factors* (vol. 29). Malden, MA: Blackwell.

Laland, K. N., & Plotkin, H. C. (1990). Social learning and social transmission of digging for buried food in Norway rats (*Rattus norvegicus*). *Animal Learning and Behavior*, *18*, 246–251.

Laland, K. N., & Plotkin, H. C. (1992). Further experimental analysis of the social learning and transmission of foraging information amongst Norway rats. *Behavioural Processes*, *27*, 53–64.

Martindale, C. (1975). *Romantic progression: the psychology of literary history*. Washington, DC: Hemisphere.

Martindale, C. (1990). *The clockwork muse: the predictability of artistic change*. New York: Basic Books.

Mayzner, M. S., & Tresselt, M. E. (1958). Anagram solution times: a function of letter order and word frequency. *Journal of Experimental Psychology*, *56*, 376–379.

McElreath, R. In the pastures and the fields: ecology, community and cultural microevolution in Usangu, Tanzania. *Evolutionary Anthropology*. Submitted for publication.

Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of innovations: a cross-cultural approach* (2nd ed.). New York: Free Press.

Roof, W. C., & McKinney, W. (1987). *American mainline religion: its changing shape and future*. New Brunswick, NJ: Rutgers University Press.

Rosenthal, T. L., & Zimmerman, B. J. (1978). *Social learning and cognition*. New York: Academic Press.

Schotter, A. (2003). Decision-making with naïve advice. *AEA Papers and Proceedings*, *93*, 196–201.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Skinner, B. F. (1981). Selection by consequences. *Science*, *213*, 501–504.

Sperber, D. (1996). *Explaining culture: a naturalistic approach*. Oxford, UK: Blackwell.

Stark, R. (1997). *The rise of Christianity: how the obscure, marginal Jesus movement became the dominant religious force in the western world in a few centuries* (1st HarperCollins pbk. ed.). San Francisco, CA: HarperSanFrancisco.

Thomason, S. G. (2001). *Language contact*. Washington, DC: Georgetown University Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teacher's College, Columbia University.

Tooby, J., & Cosmides, L. (1989). Evolutionary psychology and the generation of culture: 1. Theoretical considerations. *Ethology and Sociobiology*, *10*, 29–49.

Wilson, D. S. (2002). *Darwin's cathedral: evolution, religion, and the nature of society*. Chicago: University of Chicago Press.

Zucker, L. G. (1977). The role of institutionalization in cultural persistence. *American Sociological Review*, *42*, 726–743.