

Contents

Figures	2
Tables	9
Additional methods	10

Figures

Fig S1. Seed selection procedure.

Fig S2. Algorithmic measures of acoustic distance using MFCCs.

Fig S3. Additional orthographic distance measures.

Fig S4. Alternative measures of orthographic distance.

Fig S5. Error rates for category learning experiment.

Fig S6. Results for transcriptions of seed sounds.

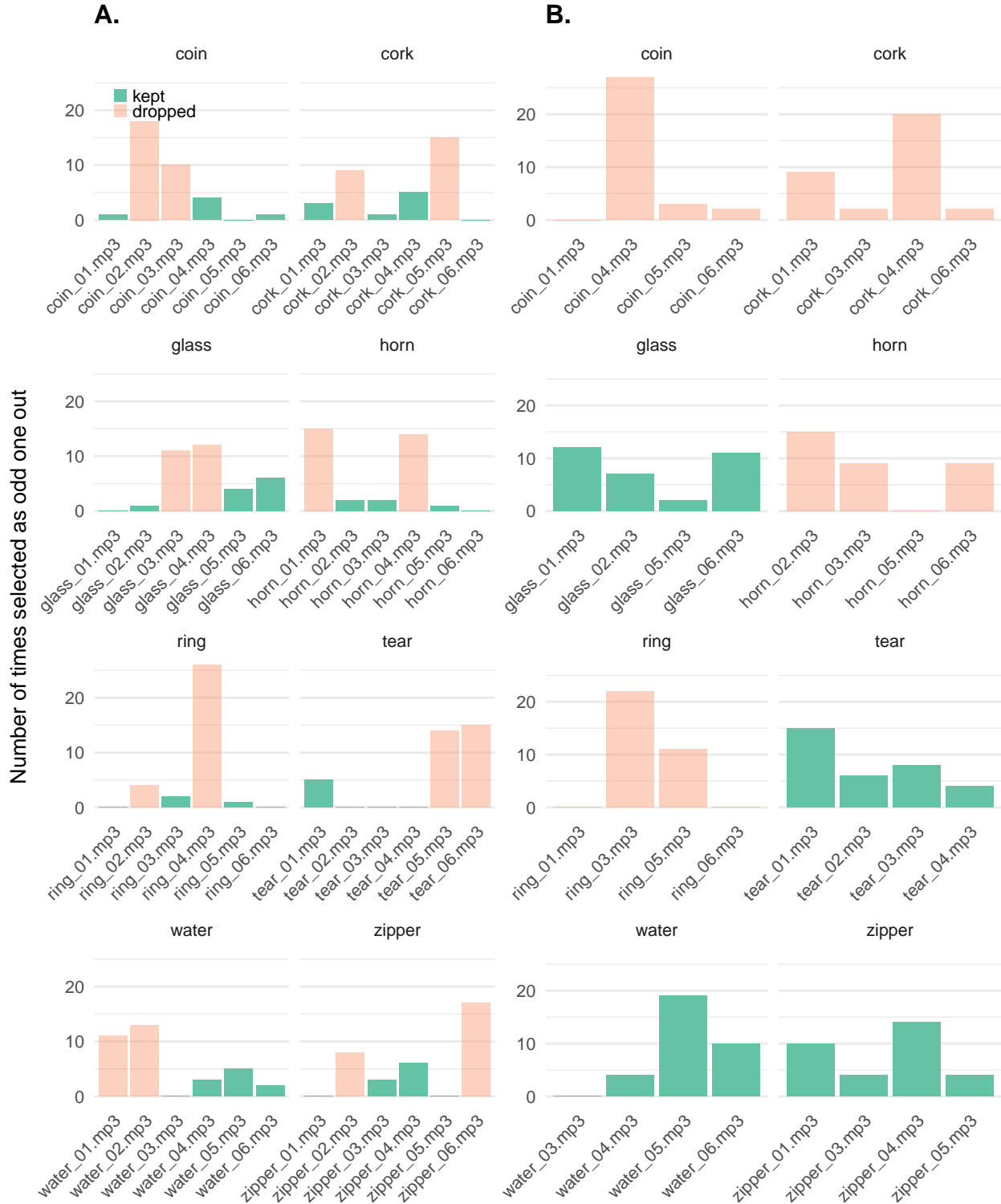


Figure S1: Seed selection norming procedure. Results of the norming procedure designed to reduce the initial set of 36 seed sounds in six categories to a final set of 16 sounds in four categories. A. Results of the first round of odd-one-out norming. After collecting these responses, two sounds that were the most different from the others in each category were removed, and the norming procedure was conducted again. B. Results of the second round of odd-one-out norming. After collecting these responses, four categories of sounds were selected for use as seeds in the transmission chain experiment.

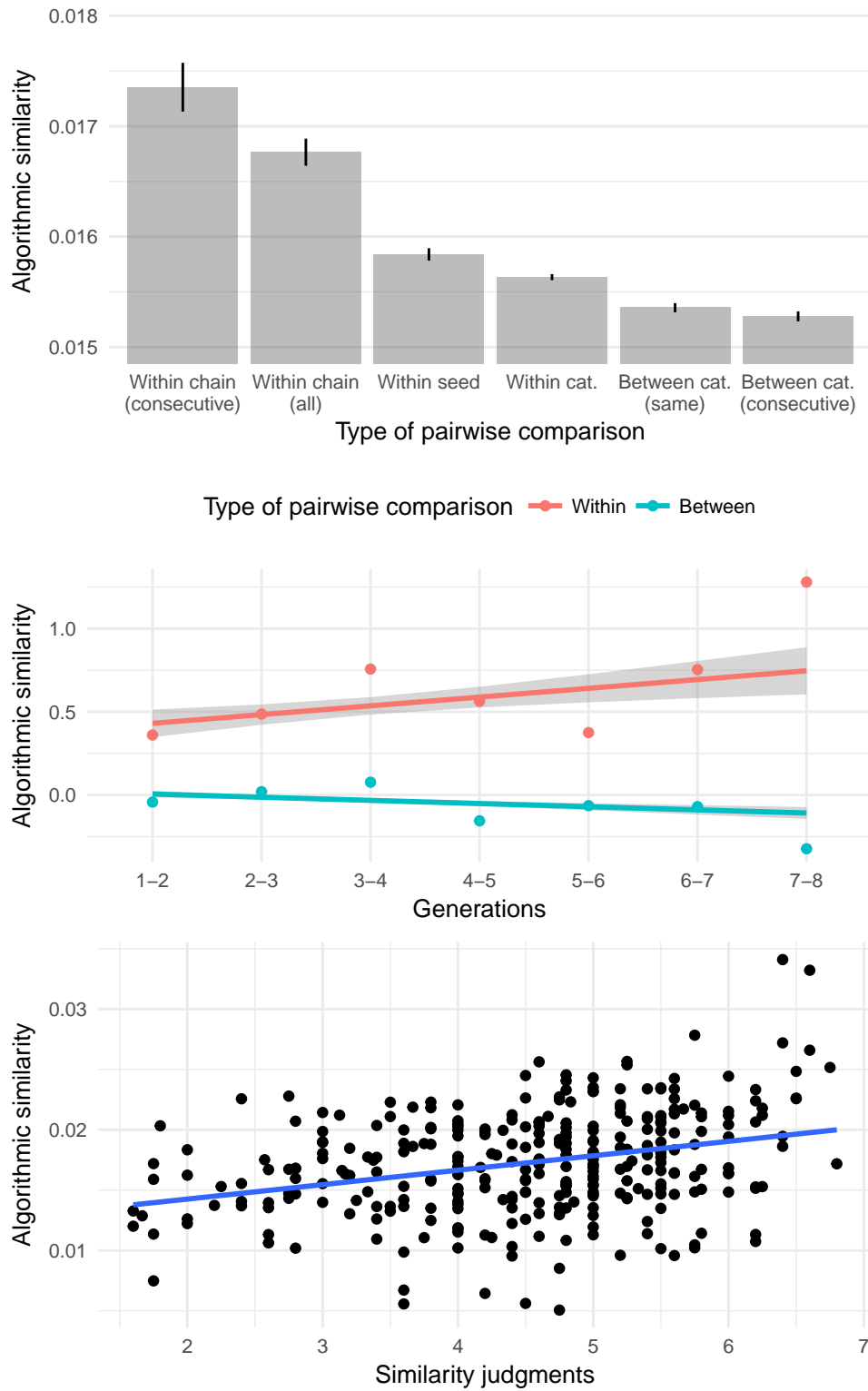


Figure S2: **Algorithmic measures of acoustic distance using MFCCs.** **a.** Average acoustic distance between pairs of sounds grouped by type of pairwise comparison. **b.** Change in algorithmic acoustic distance over generations of imitations. **c.** Correlation between similarity judgments and algorithmic measures.

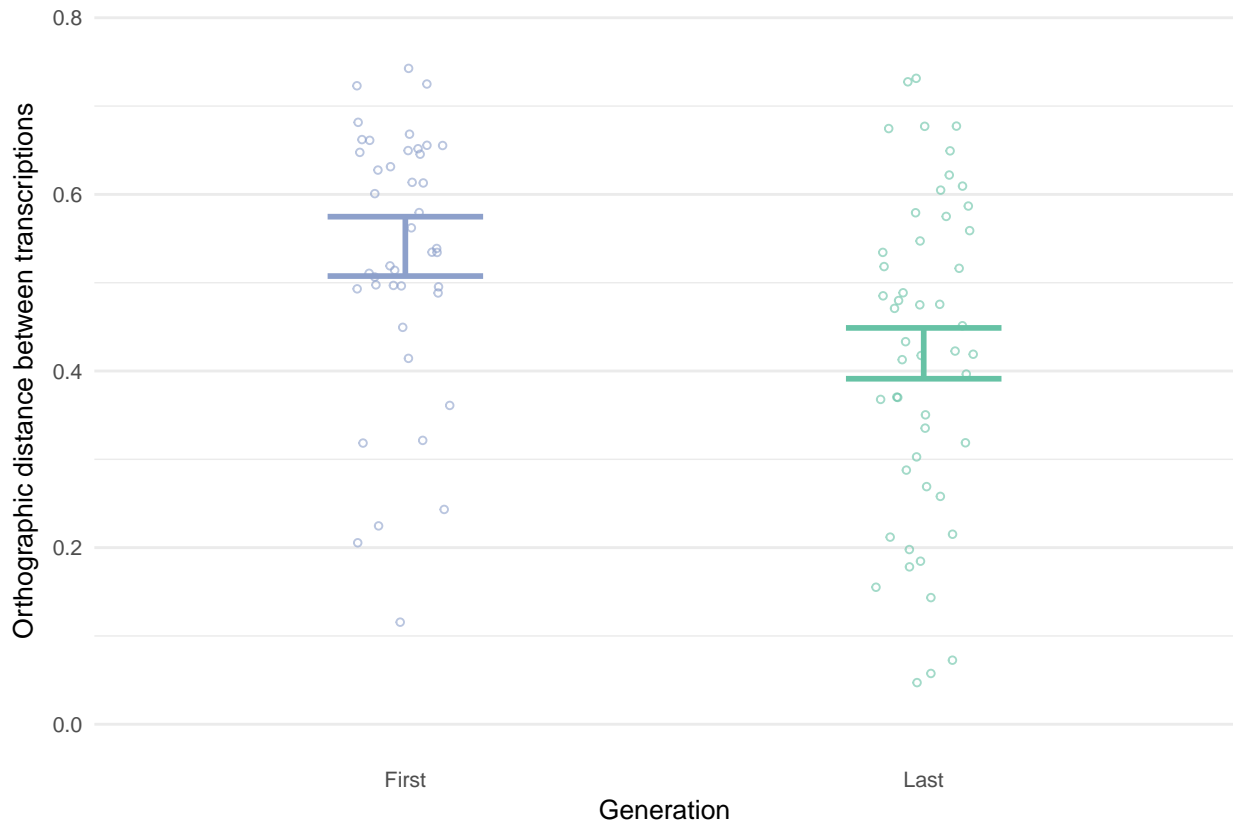


Figure S3: Orthographic distance between the most frequent transcription and all other transcriptions of a given imitation. Error bars denote ± 1 SE of the hierarchical linear model predictions.

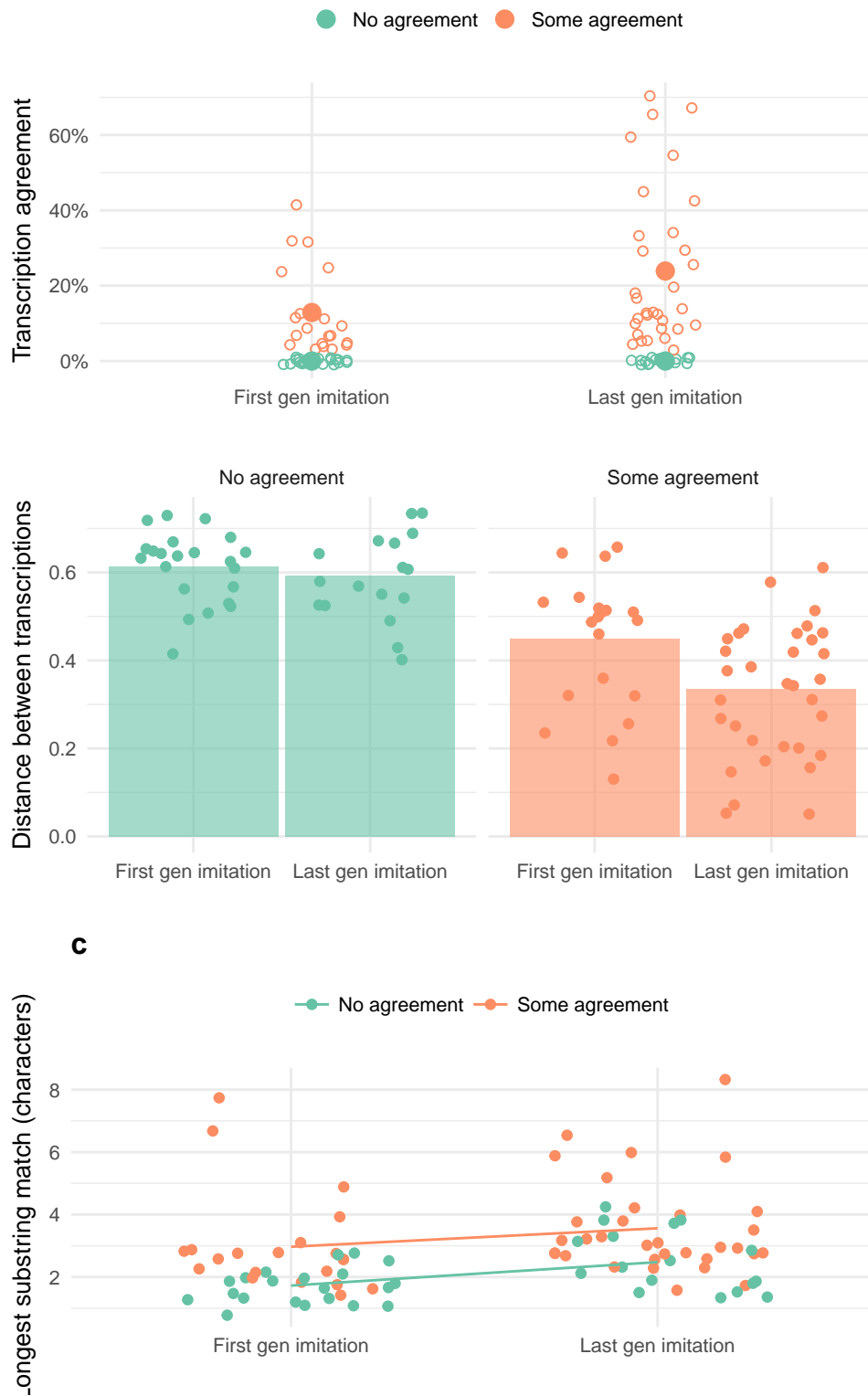


Figure S4: **Alternative measures of orthographic distance.** **a.** Percentage of exact string matches per imitation. **b.** Orthographic distance separated by whether there was any agreement among the transcriptions of a given imitation. **c.** Change in the average length of the longest substring match.

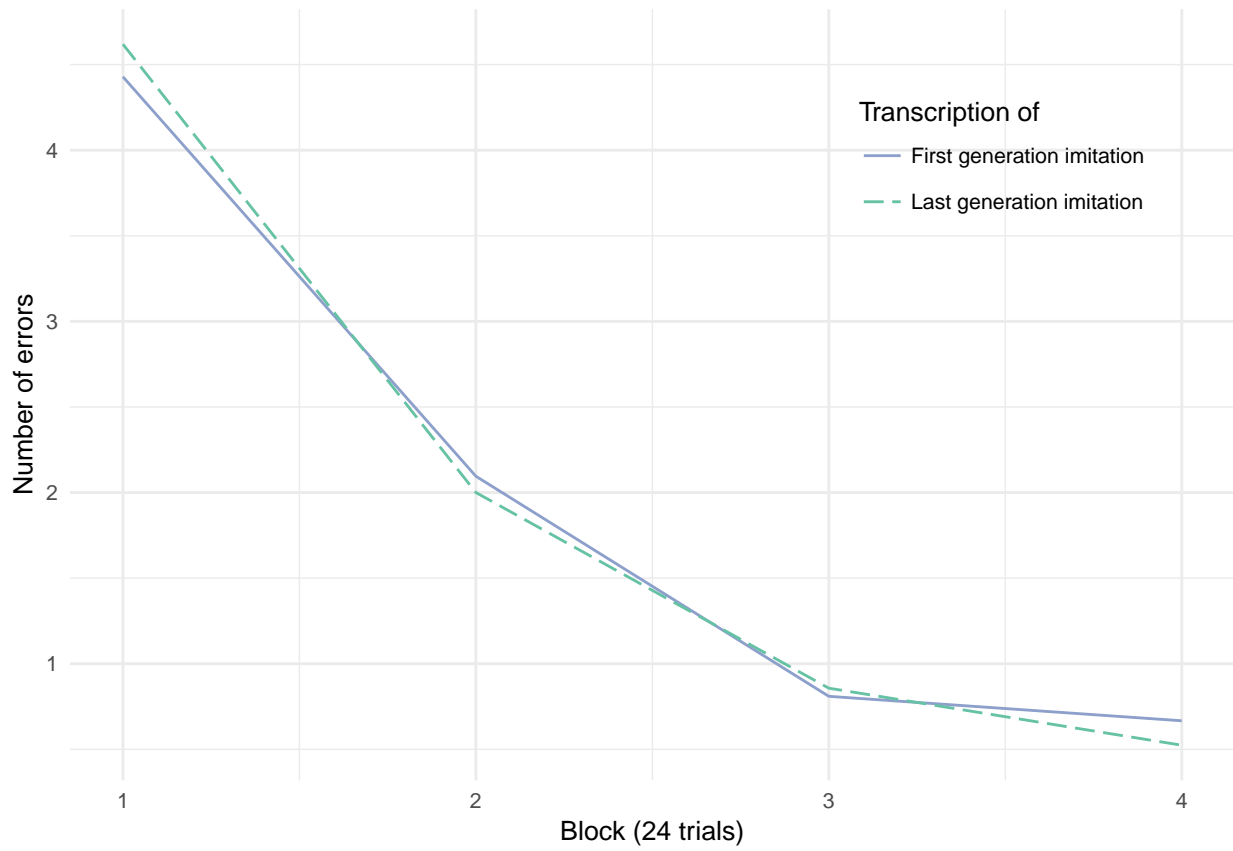


Figure S5: **Error rates for category learning experiment.** Mean number of errors per block of 24 trials, showing that accuracy performance was high even in the first block of trials, and quickly reached ceiling after the first block.

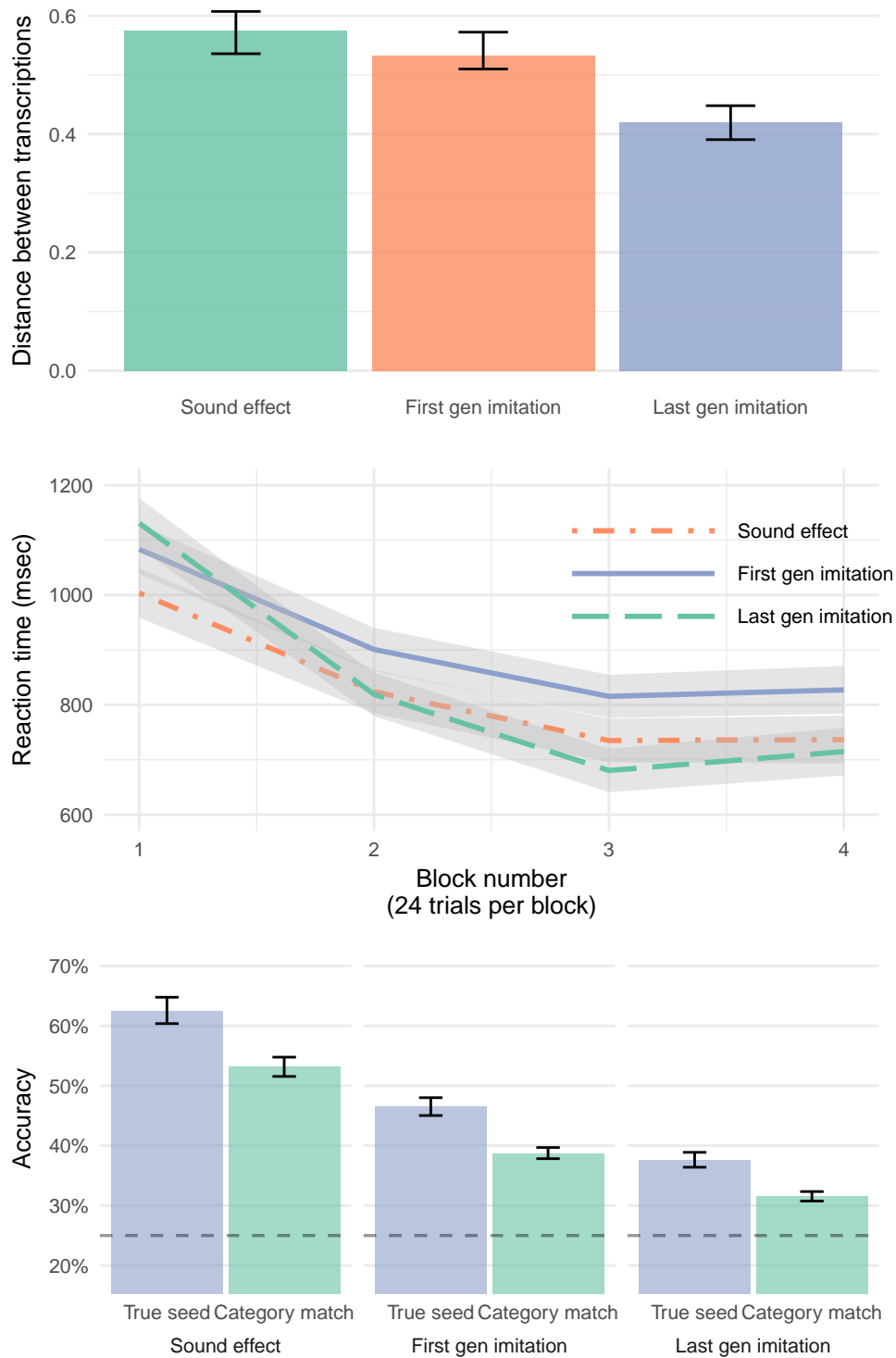


Figure S6: **Results of transcriptions directly of seed sounds.** As a control, we also had participants generate “transcriptions” directly from the seed sounds. **a.** Transcriptions of environmental sounds were the most variable in terms of orthographic distance. **b.** The most frequent of the transcriptions were the easiest to match back to the original seeds. **c.** When learning these transcriptions as category labels, participants were the fastest to learn them in the first block, but they did not generalize to new category members as fast as transcriptions taken from last generation imitations.

Tables

Model results for algorithmic measures of acoustic similarity.

Table 1: Model results predicting algorithmic acoustic similarity from type of pairwise comparison. The six types of pairwise comparisons depicted in Fig. S2 were contrast coded for the linear model with Helmert contrasts. The intercept term corresponds to the overall acoustic similarity along consecutive generations of imitations within individual transmission chains. The first Helmert contrast compares the acoustic similarities of consecutive generation imitations to all other comparisons within the same transmission chain. The second contrast compares the similarity among imitations from an individual transmission chains to the similarity of all imitations from a given seed sound (between chains). The third contrast compares the similarity among imitations of a given seed sound to imitations of the same category. The fourth contrast compare within category similarity to between category similarity, with imitations sampled from the same generation. The fifth contrast compares all other contrast types to the least similar type of comparisons, where between category imitations from consecutive generations are compared.

term	estimate	std.error	statistic	p.value
(Intercept)	0.0160376	0.0000403	398.299246	0.0000000
edge_type_helmert1	-0.0002942	0.0001126	-2.613679	0.0089608
edge_type_helmert2	-0.0004065	0.0000418	-9.732521	0.0000000
edge_type_helmert3	-0.0002547	0.0000205	-12.427253	0.0000000
edge_type_helmert4	-0.0002081	0.0000144	-14.494862	0.0000000
edge_type_helmert5	-0.0001519	0.0000111	-13.731716	0.0000000

Additional methods

Instructions for the collecting vocal imitations.

Participants played a version of the children's game of Telephone via a web-based interface. The instructions given to participants were as follows.

We are researchers at the University of Wisconsin-Madison studying how audio messages are passed on from person to person, much like in the children's game Telephone. If you choose to participate, we will ask you to listen to an audio message recorded by someone else, and then record yourself imitating the message that you heard using your computer's microphone.

Unlike the children's game of Telephone, the sounds you will hear will not be recognizable English words, but will be various nonspeech sounds. Your task is the same, however: to recreate the sound you heard as accurately as you can.

The interface you will use for playing the game is shown below. Before you begin, make sure you are using a computer with a working microphone, and that your computer speakers are turned on (or, preferably, that you are using headphones). If there is TV or music playing in the background, please turn it off or move to a quieter space. To play, first you will click on the top sound wave icon to hear the message you are to imitate. After you listen to the message, then you will click on the bottom sound wave icon to record your imitation.



After you submit your recording, the next person who arrives at the site will now hear your recording, and record their own recreation of your recreation (and so on, and so on!). Don't worry, we will not ask you to provide any information that identifies you. Each player will be asked to imitate several sounds.

Instructions for acoustic similarity ratings

Research assistants rated the similarity between 324 different pairs of imitations. These pairs comprised consecutive imitations in the transmission chain design, e.g., each message was compared to its response. Message order was randomized on each trial so that participants did not know which message was the original and which message was the imitation. Raters were also blind to the overall generation of the imitations by

randomizing generation from trial to trial. To facilitate consistency in rating, pairs of sounds were blocked by category, e.g., all tearing sounds were rated before moving on to other categories of sounds. The instructions given to raters were as follows.

On each trial, you will hear two sounds played in succession. To help you distinguish them, during the first you will see the number 1, and during the second a number 2. After hearing the second sound, you will be asked to rate how similar the two sounds are on a 7-point scale.

A 7 means the sounds are nearly identical. That is, if you were to hear these two sounds played again, you would likely be unable to tell whether they were in the same or different order as the first time you heard them. A 1 on the scale means the sounds are entirely different and you would never confuse them. Each sound in the pair will come from a different speaker, so try to ignore differences due to just people having different voices. For example, a man and a woman saying the same word should get a high rating.

Please try to use as much of the scale as you can while maximizing the likelihood that if you did this again, you would reach the same judgments. If you need to hear the sounds again, you can press ‘r’ to repeat the trial. If one of the sounds is a non-verbal sound (like someone tapping on the mic), or if you only hear a single sound, or if you are otherwise unable to judge the similarity between the sounds, press the ‘e’ key to report the error. Pressing ‘q’ will quit the experiment. Your progress will be saved and you can continue later. Press the SPACEBAR to begin the experiment.

Inter-rater reliability for acoustic similarity ratings

Inter-rater reliability was calculated as the intra-class coefficient treating the group as the unit of analysis: ICC = 0.76, 95% CI [0.70, 0.81], $F(170, 680) = 4.18$, $p < 0.001$.

Instructions for category learning experiment

In the category learning experiment, participants had to learn novel labels for categories of environmental sounds through trial and error. On each trial, they heard an environmental sound (one of the 16 seed sounds) and then read a novel “word”—a transcription of an imitation collected in the transmission chain experiment. Here are the instructions these participants received.

Thanks for participating in the study. In this experiment we are going to have you learn some new words to name different types of sounds. Think about it like you are on an alien planet and you are learning words for some of the sounds that are common on their planet.

In the experiment you’ll learn which names go with which sounds through trial and error. On each trial you will hear one of the sounds and then you’ll see a word for it, and you have to decide whether the word is the correct word. If it is, you’ll press the green button on the gamepad. If it’s not the right word, press the red button.

At first you will have to guess which words go with which sounds, but since you’ll be getting feedback on your performance, you should be able to learn the correct names for the sounds pretty quickly. If you guess correctly you’ll see a green checkmark. If you guess incorrectly you’ll see a red ‘X’.

If you have any questions, now is the time to ask the experimenter. When you feel you understand these instructions and are ready to begin, press the green button.

Procedure for sampling transcriptions to use as category labels.

Our procedure for sampling transcriptions to use as category labels was as follows: First, we removed transcriptions that contained less than 3 unique characters and transcriptions that were over 10 characters long. Of the remaining transcriptions, a sample of 56 were selected that were approximately equally associated with the target category.

To measure the association between each imitation and its target category (the category of the seed sound), we used the match accuracy scores reported in Experiment 2. The reason for using this measure of association strength as a control for selecting words to learn as category labels was to be able to select words that were initially equally associated with the target categories. Equating along this dimension allowed for a more focused test of differences between the words in terms of generalization to new category members.

The final sample of transcriptions were selected using a bootstrapping procedure which involved selecting a desired mean (the average association strength for eligible transcriptions of last generation imitations) and sampling transcriptions from first generation imitations and from seed sounds until the match accuracy of those imitations matched the desired mean within 1 standard deviation.