Creating words from iterated vocal imitation

We report the results of a large-scale (*N* = 1571) experiment to investigate whether spoken words can emerge from the process of repeated imitation. Participants played a version of the children’s game “Telephone”. The first generation was asked to imitate recognizable environmental sounds (e.g., glass breaking, water splashing); subsequent generations imitated the imitators, for a total of 8 generations. We then examined whether the vocal imitations became more stable and word-like, retained a resemblance to the original sound, and became more suitable as learned category labels. The results showed (1) the imitations became progressively more word-like, (2) even after 8 generations, could be matched above chance to the environmental sound that motivated them, and (3) imitations from later generations were more effective as learned category labels. These results show how repeated imitation can create progressively word-like forms while retaining a resemblance to the original sound that motivated it.

People have long pondered the origins of languages especially the words that compose them. For example, both Plato in his *Cratylus* dialogue (Plato and Reeve 1999) and John Locke in his *Essay Concerning Human Understanding* (Locke 1948) examined the "naturalness" of words--whether they are somehow imitative of their meaning. Some theories of language evolution have hypothesized that vocal imitation played an important role in generating the first words of spoken languages (e.g., Brown, Black, and Horowitz 1955; Donald 2016; Imai and Kita 2014; Perlman, Dale, and Lupyan 2015); early humans may originally have referred to a predatory cat by imitating its roar, or to the discovery of a stream by imitating the sound of rushing water. Such vocal imitation might have served to clarify the referent of a vocalization and eventually establish a mutually understood word. In this study, we investigate the formation of onomatopoeic words--imitative words that resemble the sounds to which they refer. We ask whether onomatopoeic words can be formed gradually and without instruction simply from repeating the same imitation over generations of speakers.

Onomatopoeic words appear to be a universal lexical category found across the world’s languages (Dingemanse 2012). Languages all have conventional words for animal vocalizations and various environmental sounds. Rhodes (1994), for example, documented a repertoire of over 100 onomatopoeic words in English, which he notes exist along a continuum from "wild" to "tame". Wild words have a more imitative phonology whereas tame words take on more standard phonology of other English words. In some cases, words that begin as wild imitations of sounds become fully lexicalized and integrated into the broader linguistic system, when they behave like more "ordinary" words that can undergo typical morphological processes. Examples are English words like "crack" or the recently adapted "ping". Dingemanse and Akita (2016) found that the degree of expressiveness of an imitative word is inversely related to its degree of morphosyntactic integration.

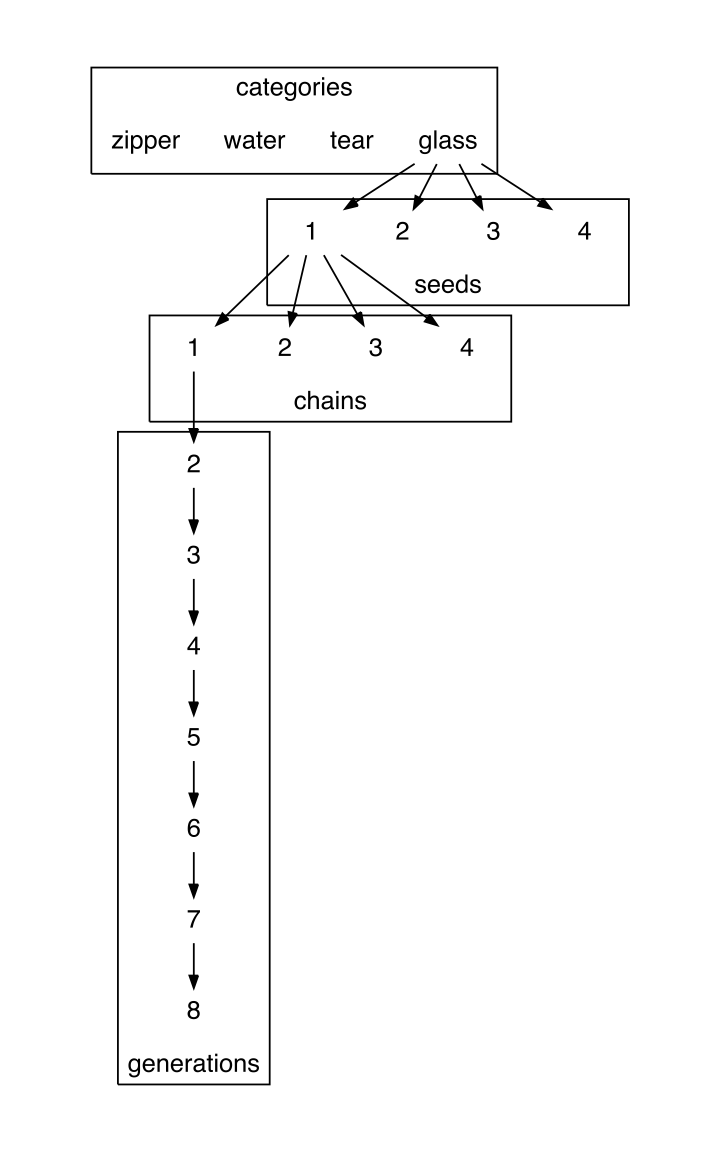
People often use more wild vocal imitations and other sound effects during demonstrative discourse, especially when producing quotations (Blackwell, Perlman, and Tree 2015; Clark and Gerrig 1990). Vocal imitation might be especially prominent in some communities. For example, an ethnographic study of Mbendjele Pygmies living in the Congo found that Mbendjele speakers frequently use vocal imitation when narrating dramatic events (Lewis 2009). For example, when describing an event such as an encounter with a dangerous animal in the jungle, they use vocal imitation to emphasize acoustic details of the event, including the sound of the animal and also other inanimate details like the thrashing of trees.

However, not all researchers agree that vocal imitation has any significant role in language. For instance, Pinker and Jackendoff (2005) suggested that, “Humans are not notably talented at vocal imitation in general, only at imitating speech sounds (and perhaps melodies). For example, most humans lack the ability (found in some birds) to convincingly reproduce environmental sounds … Thus ‘capacity for vocal imitation’ in humans might be better described as a capacity to learn to produce speech.” Nevertheless, experiments show that people can actually be quite effective at using vocal imitation. For example, Lemaitre and Rocchesso (2014) collected imitations and verbal descriptions of various mechanical and synthesized sounds. When participants listened to these and were asked to identify the source, they were more accurate with imitation than verbal descriptions of those sounds. A subsequent study found that vocal imitations tend to focus on a few salient features of the sound rather than a high fidelity representation, which aids identification of the source (Lemaitre et al. 2016).

Thus humans can be effective at communicating with vocal imitation, it can play an important role in narration and discourse, and it appears to be the basis for substantial inventories of sound-imitative vocabulary across languages. But the process by which onomatopoeic words like “meow”, “ping” and “buzz” emerge from vocal imitations has yet to be observed. Here we examine whether simple repeated imitations of environmental sounds become more wordlike even in the absence of explicit communication intent or the intent to create a word-like token. Alternatively does the claimed limited fidelity of human vocal imitation restrict the formation of stable words. To test this, we recruited participants to engage in a large scale online version of the children's game of "Telephone" in which an acoustic message is passed from one person to the next. After obtaining these imitations, we investigated how the imitations changed over generations to determine whether they became more wordlike. We investigated the acoustic properties of the imitations as well as the orthographic properties of the imitations once transcribed into English words. Finally, we tested how quickly these invented words are learned as category labels in a category learning experiment.

# Methods

The design of the transmission chain experiment is shown in Fig. 1. After collecting the vocal imitations, we then assessed changes in the imitations over generations in a series of experiments listed in Table 1. First we assessed the extent to which the imitations could be matched back to the original sounds. Then we collected transcriptions of a subset of imitations, and tested whether these transcriptions could be matched back to the original sounds. Finally, we used a set of transcriptions taken from first and last generation imitations as novel labels in a simple category learning experiment.



The design of the transmission chain experiment. 16 seed sounds were used, 4 in each category of nonverbal environmental sounds. Each seed sound was imitated by 4 different participants, resulting in at least 4 chains off of each seed sound. Subsequent participants imitated the imitations and so on for a maximum of 8 generations.

## Materials

We initially selected a set of 36 sounds in 6 different categories of environmental sounds. Inanimate categories of sounds were selected because they were less likely to have lexicalized onomatopoeic forms already in English, and they were assumed to be less familiar and more difficult to imitate. We used an odd-one-out norming procedure (*N* = 105 participants) to reduce this initial set to 16 “seed” sounds: 4 sounds in each of 4 categoriesseeds using an designed to remove any sounds that were outliers within each category. The four final categories included:

## Participants

In Experiments 1-4, participants were recruited via Amazon Mechanical Turk and paid to participate. Participants in Experiment 5 were University of Wisconsin-Madison undergraduates who received course credit in exchange for participation.

## Procedures

The procedures for each experiment listed in Table 1 are described below.

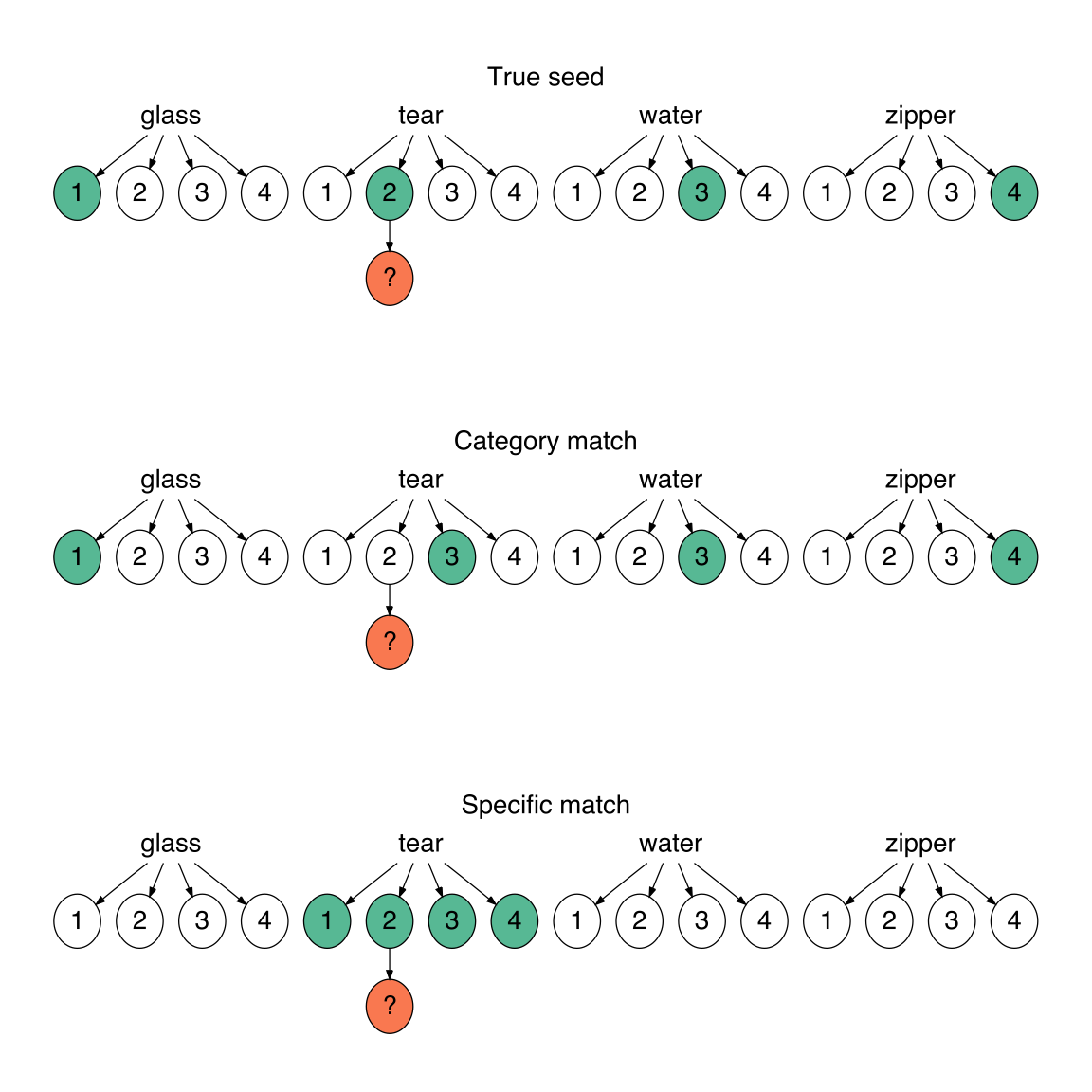
### Collecting vocal imitations

Participants were paid to participate in an online version of the children's game of "Telephone". The instructions informed participants that they would hear some sound and their task is to reproduce it as accurately as possible using their computer microphone. Participants listened to and imitated 4 sounds. Participants received one message from each of the four categories of sounds drawn at random such that participants were unlikely to hear the same person more than once.

Imitations were monitored as they were received by an experimenter. The purpose of this monitoring was to catch any gross errors in recording before they were passed on to the next generation of imitators. For example, most sounds required trimming the recording more closely to the vocalization--removing mouse clicks and other ambient sounds--before it was suitable for the next generation. The experimenter also blocked sounds that violated the rules of the experiment, e.g., by saying something in English.

### Matching imitations to seeds

To find out if the produced sounds retained some resemblance to the original environmental sound that motivated it (i.e. the seed sound), we recruited additional participants to participate in a 4-alternative forced choice test in which they hear an imitation from one of the chains and were asked to select which of the four seed sounds they thought the imitation most closely resembled. They did not receive any feedback on their performance. Each participant was assigned four different seed sounds as choices (for example…). We tested three types of matching questions that differed according to the relationship between the imitation and the four seed sounds serving as the options in the 4AFC task (Fig. 2).



Three types of matching questions depicted in relation to the original set of 16 seed sounds. For each question, participants listened to a sample imitation (orange circle) and had to guess which of 4 sound choices (green circles) they thought the person was trying to imitate. (Top) True seed questions contained the actual seed that generated the imitation in the choices, and the distrator seeds were sampled from different categories. (Middle) Category match questions also used distractor sounds from different categories but the correct seed was not the actual seed, but a different sound within the same category. (Bottom) Specific match questions pitted the actual seed against the other seeds within the same category.

### Collecting transcriptions of imitations

To determine Participants were asked to transcribe between 9 and 20 imitations each. They were instructed to write down what they heard as a word so that the written word would sound as much like the message as possible. We selected the first and final three imitations in each transmission chain to be transcribed into English orthography resulting in a total of 2182 or roughly 21 transcriptions per imitation. Only the transcriptions from participants who passed a catch question were included. All transcriptions containing actual English words were excluded from analysis.

### Matching transcriptions to seeds

Participants completed a modified version of the 4AFC described in **Matching imitations to seeds**, above. Instead of listening to imitations, participants now read a transcription of an imitation, which they were told was an invented word. They were told that the word was invented to describe one of four presented sounds, and they had to guess which one. To keep the number of required participants to a manageable number, we used only the four most frequent transcriptions for each imitation.

### Using transcriptions as category labels

Participants learned, through trial-and-error, the names for four different categories of sounds. On each trial participants listened to one of the 16 environmental sounds used as seeds and then saw a novel word--a transcription of one of the imitations. Participants responded by pressing a green button if the label was the correct label and a red button otherwise. They received accuracy feedback after each trial.

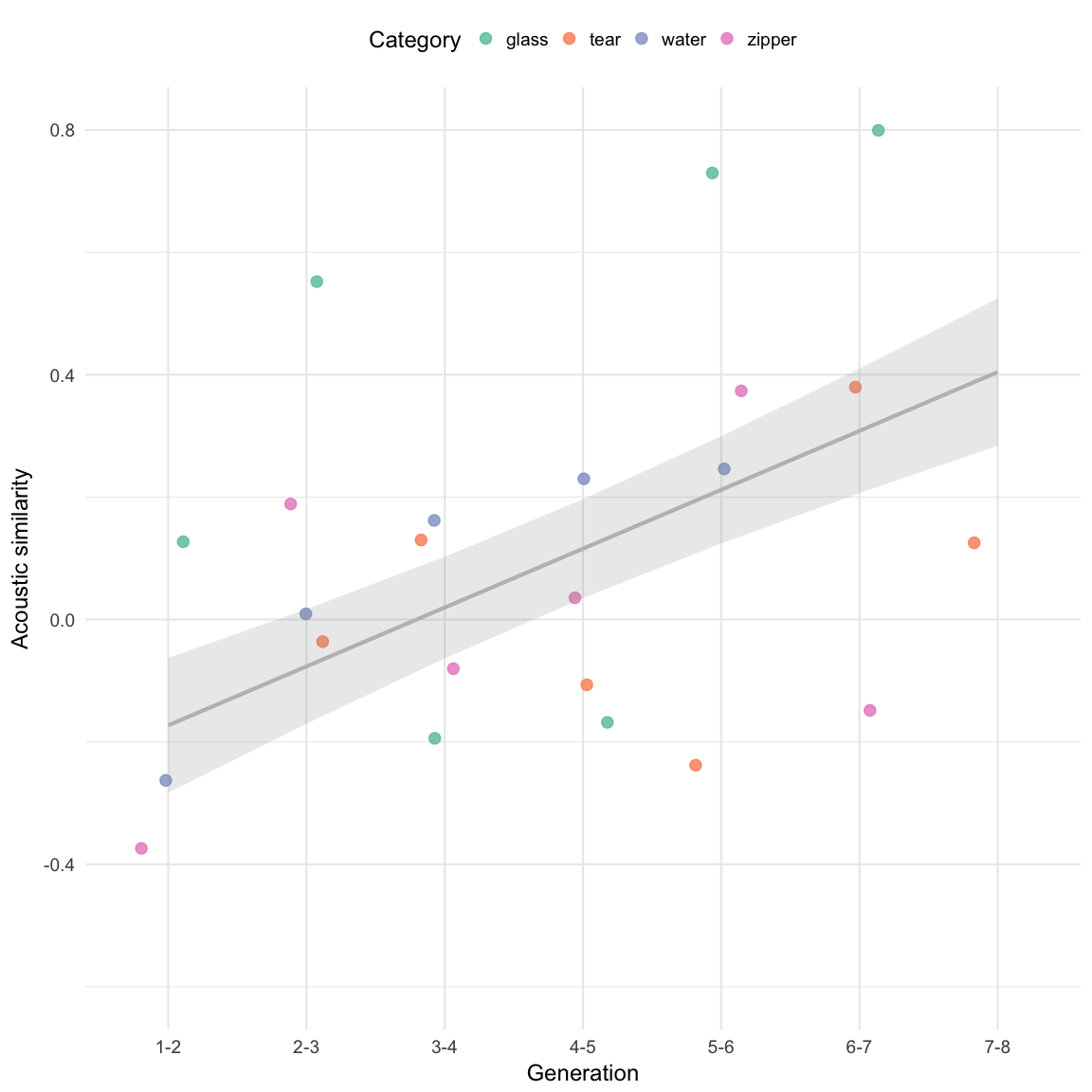
The experiment was divided into blocks so that participants had repeated exposure to each sound and the novel labels multiple times within a block. At the start of a new block, participants received four new sounds from the same four categories (e.g., a new zipping sound, a new water-splash sound, etc.) that they had not heard before, and had to associate these sounds with the same novel labels from the previous blocks. The extent to which their performance declined at the start of each block serves as a measure of how well the label they associated with the sound worked as a label for the *category*.

To determine which transcriptions to test in the experiment, we first selected only those transcriptions which had above chance matching performance when matching back to the original seeds. Then we excluded transcriptions that had less than two unique characters, and transcriptions that were over 10 characters long. From this set of eligible transcriptions, we sampled from both first and last generation imitations to reach a final set that controlled for overall matching accuracy.

# Results

## Changes in acoustic similarity

We found that imitations from later generations were more similar to one another than imitations from earlier generations (Fig. 2). Given large differences in recording quality resulting from conducting the experiment online, we were unable to use previously published techniques for calculating acoustic distance (cf. Lemaitre et al. 2016). Instead, we obtained subjective measures of acoustic similarity using a controlled, randomized norming procedure completed by research assistants. Five RAs listened to pairs of imitations while blind to generation and rated their similarity on a 7-point scale. We found that imitations from later generations were rated as being more similar to one another than imitations from earlier generations, *b* = 0.10 (0.03), *t*(3.5) = 3.86, *p* = 0.026.



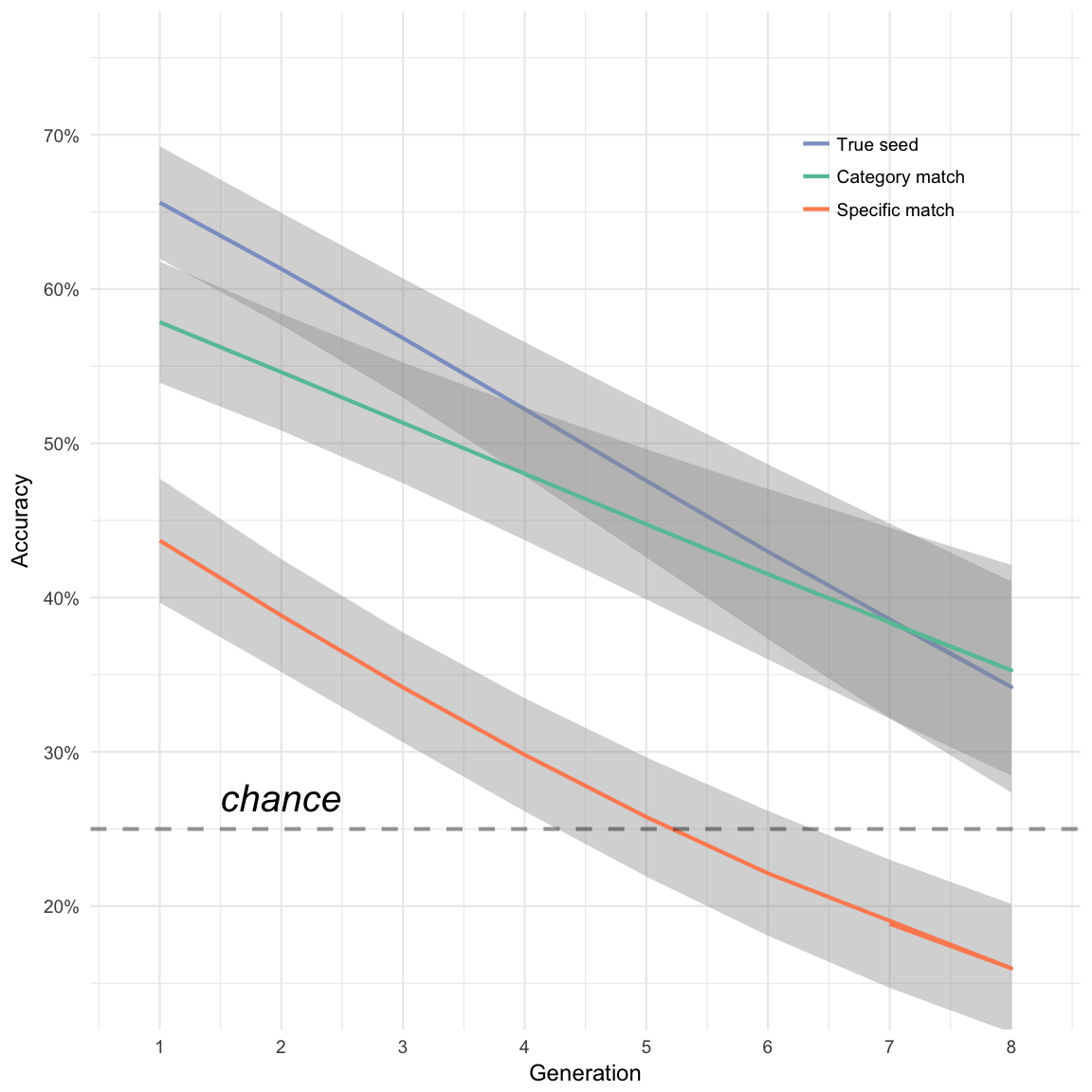
Increase in acoustic similarity over generations. Acoustic similarity judgments were collected on a 7-point scale where a 1 meant the sounds could never be confused for one another and a 7 meant the sounds were nearly identical. Prior to analysis, the similarity ratings were converted to z-scores for each rater. Points depict mean acoustic similarity ratings for imitations in each category of environmental sounds. The predictions of the linear mixed effects model with random effects for rater and for category are shown. The error bands denote +/- 1 standard error of the model predictions.

## Matching imitations to seeds

Matching accuracy for all question types (see Fig. 2) was above chance for the first generation imitations, *b* = 1.65 (0.14) log-odds, odds = 0.50, *z* = 11.58, *p* < 0.001, and decreased over generations, *b* = -0.16 (0.04) log-odds, *z* = -3.72, *p* < 0.001. We tested whether this increase in question difficulty was constant across the three types of questions or if some question types became more difficult at later generations. In particular, we hypothesized that if the imitations were becoming more like category labels as they were repeated, then performance on questions where category information enabled a correct response would be more resilient to generational decay.

The results are shown in Fig. 4. The first evidence in support of our hypothesis comes in comparing performance on questions requiring a category match to performance on questions where guessing correctly required distinguishing the true seed from other sounds within the same category. Performance decreased over generations more rapidly for these specific match questions than for category match questions, *b* = -0.05 (0.02) log-odds, *z* = -2.53, *p* = 0.012, suggesting that category information was more resistent to loss through transmission.

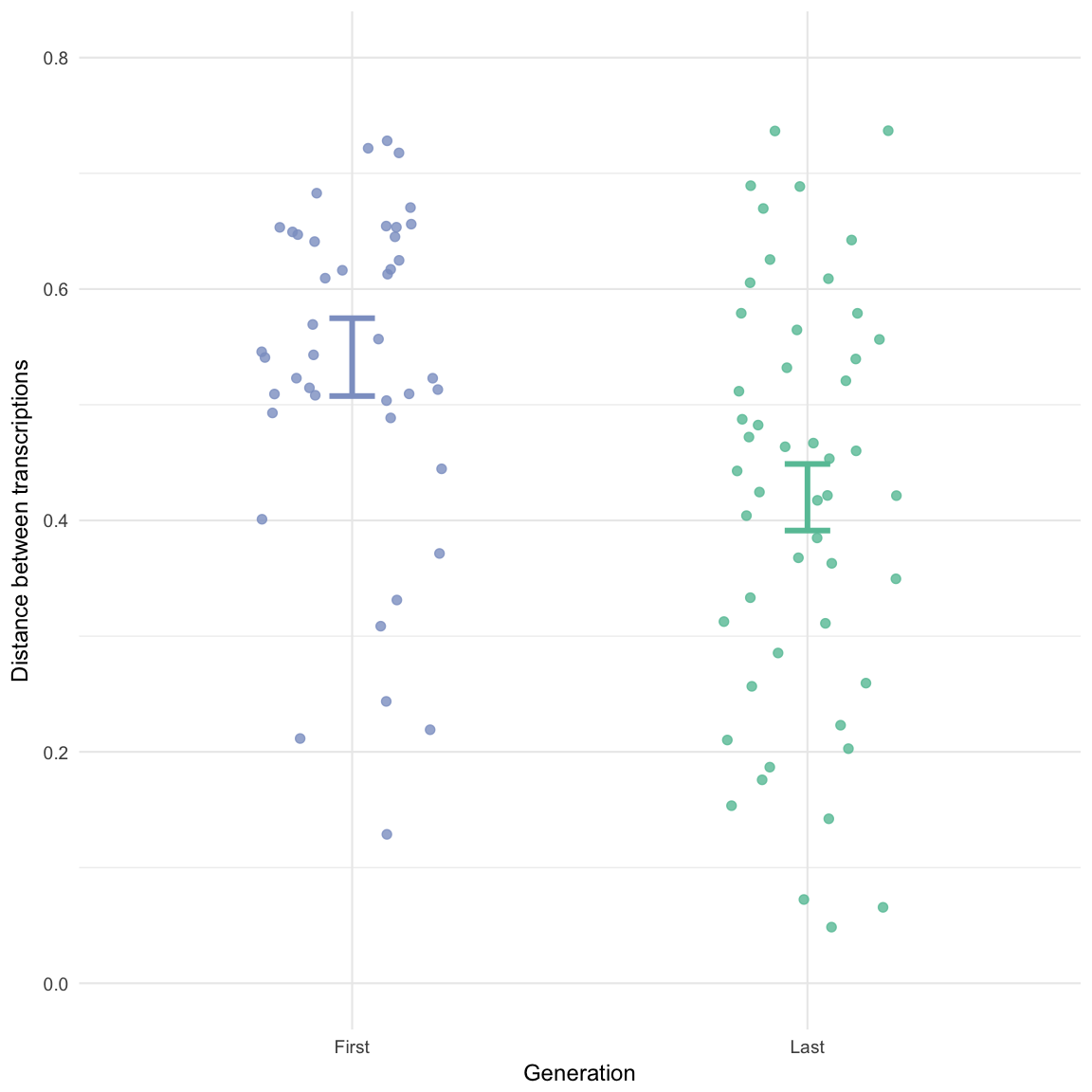
One explanation for this result is that the specific match questions are simply harder than the category match questions. However, performance also decreased more rapidly for the easiest type of question where the correct answer was the actual seed generating the imitation. The advantage for having the true seed among the options decreased over generations, *b* = -0.07 (0.02) log-odds, *z* = -2.83, *p* = 0.005. These results indicate that later generation imitations were more likely to be recognized as identifiers of a particular category than they were of particular exemplars within each category.



Accuracy in matching imitations back to seed sounds. Perforamnce is separated by question type concerning the relationship between the imitation and the options in the question (see Fig. 2). Lines depict predictions from the generalized linear mixed effects model along with +/- 1 standard error. The advantage of having a True seed among the options (difference between True seed and Category match lines) decreased over generations, and the advantage of having distractors from different categories (difference between Category match and Specific match lines) increased over generations.

## Changes in orthographic agreement

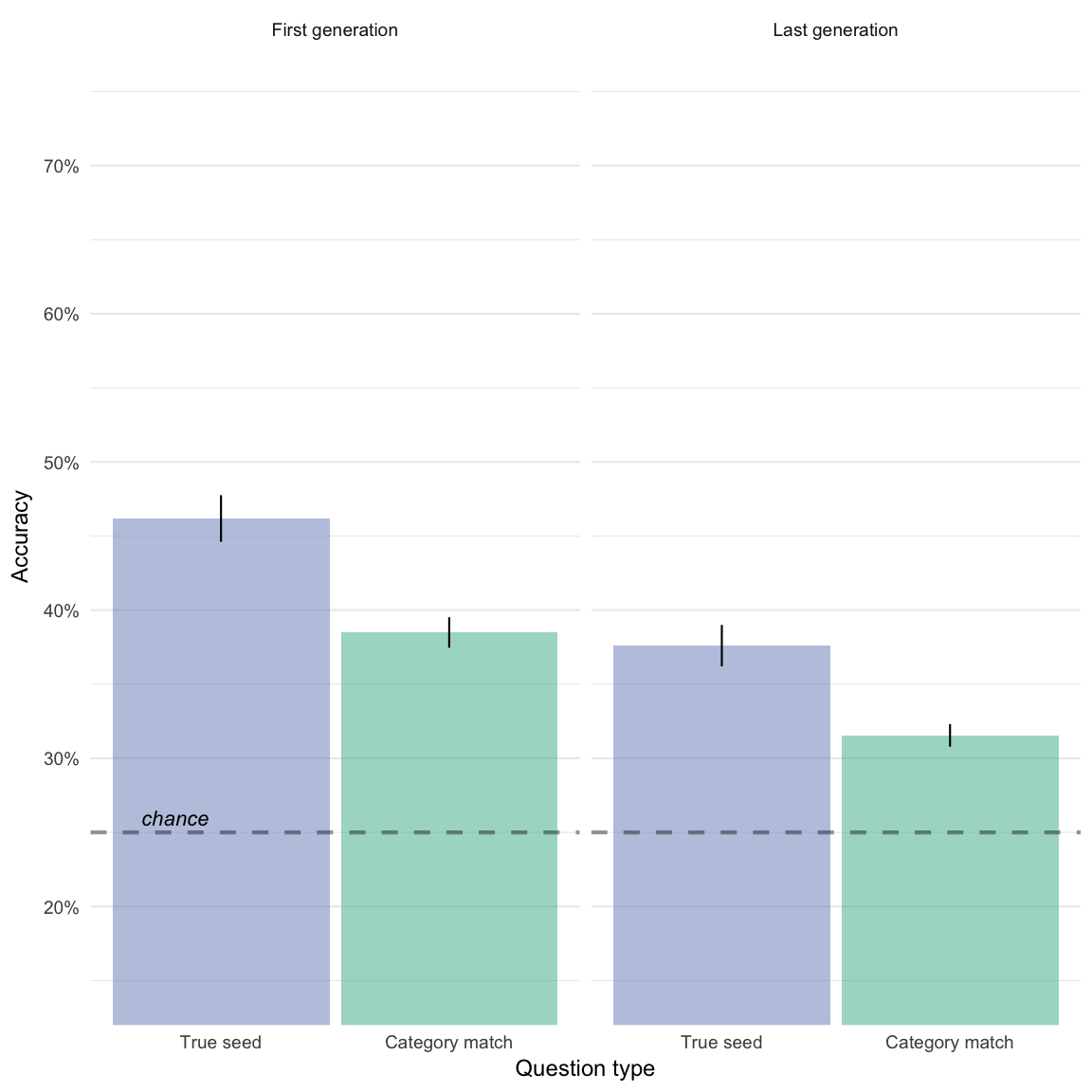
We next investigated how agreement among orthographic forms of the imitations changed over generations. A total of 106 imitations were transcribed, with approximately 21 transcriptions per imitation. Analyzing changes in orthographic agreement over generations paralleled what was observed in the analysis of acoustic similarity: Transcriptions from later generation imitations were more similar to one another in terms of orthographic distance than transcriptions from earlier generations, *b* = -0.12 (0.03), *t*(-3.6) = 3.05, *p* = 0.035 (Fig. 5). This result supports our hypothesis that the imitations were becoming more stable in both acoustic and orthographic forms.



Average orthographic distance among transcriptions of imitations taken from first and last generations. String distances were calculated between the most frequent transcription for each imitation compared to all others using the SequenceMatcher class available in the python standard library. Each point represents the average distance among all transcriptions for a single imitation. Error bars are +/- 1 standard error of the linear mixed effects model predictions.

## Matching transcriptions to seeds

To investigate whether the invented words retained any resemblance to the original seed sounds, we selected the top 4 most frequent transcriptions for each imitation and presented them in a modified version of the matching game. Participants were able to guess the correct meaning of the transcribed word above chance even after 8 generations of repetition, *b* = 0.83 (0.13) log-odds, odds = -0.18, *z* = 6.46, *p* < 0.001 (Fig. 6). This was true both for true seed questions, *b* = 0.75 (0.15) log-odds, odds = -0.28, *z* = 4.87, *p* < 0.001, and for category match questions, *b* = 1.02 (0.16) log-odds, odds = 0.02, *z* = 6.39, *p* < 0.001. The effect of generation did not vary across these question types, *b* = 0.05 (0.10) log-odds, *z* = 0.47, *p* = 0.637.

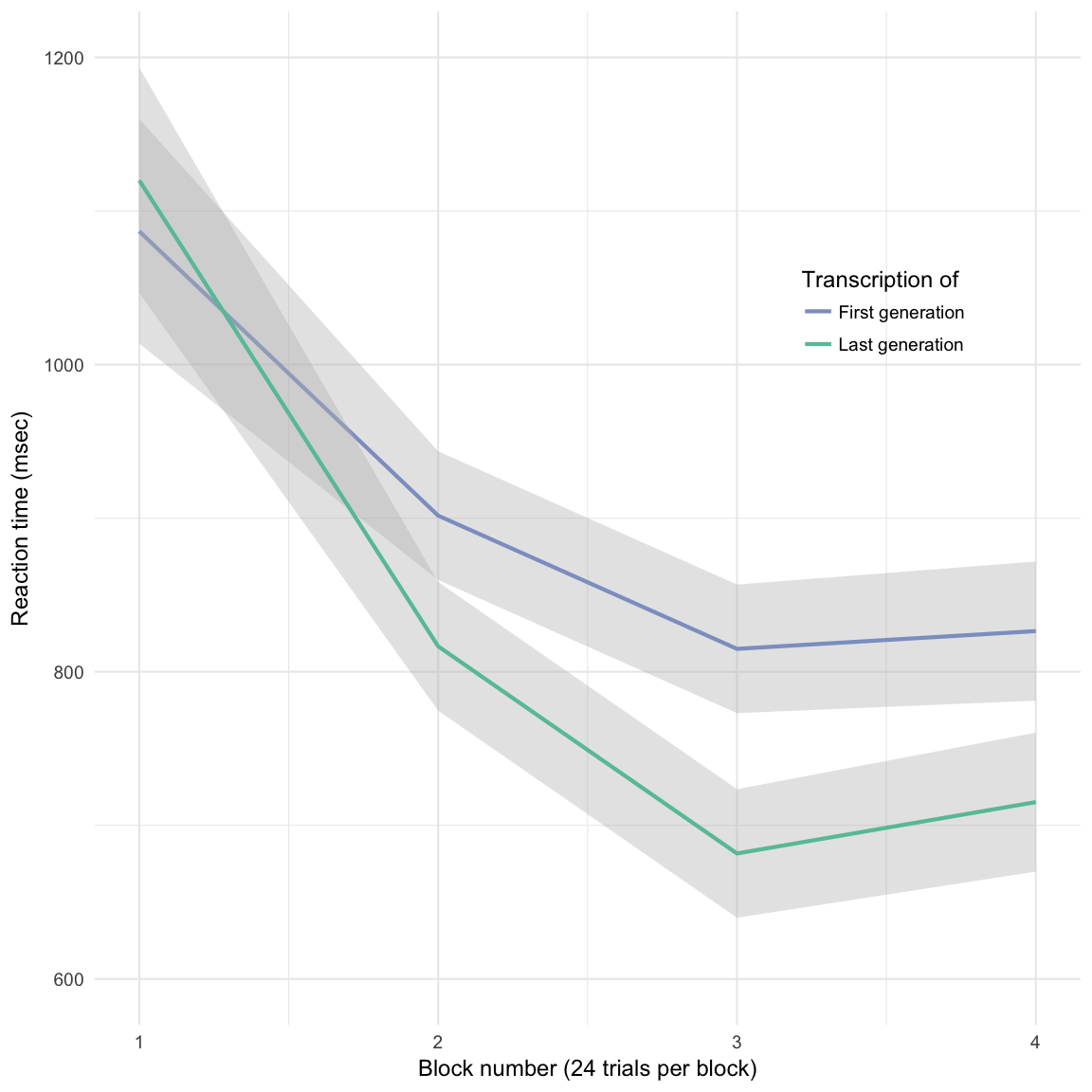


Matching accuracy for transcriptions of imitations taken from first and last generations. True seed questions contained transcriptions of the actual seed generating the transcribed word. Category match questions contained transcriptions of imitations of other seeds from the same category.

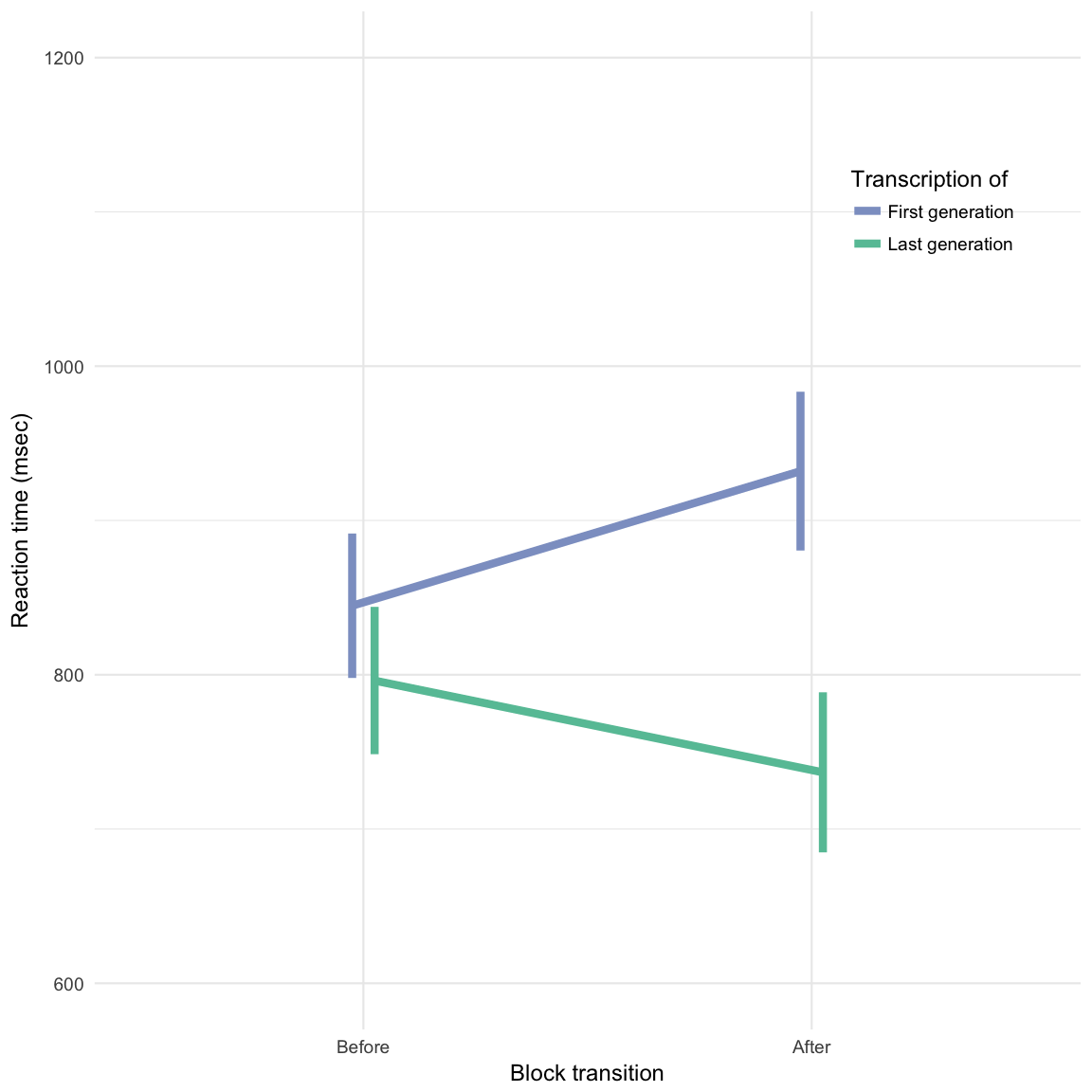
## Using transcriptions as novel category labels

Finally, we examined whether there was a learning advantage to the more wordlike imitations emerging through iterated repetition compared to direct imitations of the source of the sound. An advantage of word cues as opposed to environmental cues to a particular category is that words activate more categorical representations useful in generalizing across idiosyncratic differences between category members (Edmiston and Lupyan 2015). Therefore, we hypothesized that transcriptions of the more wordlike forms emerging through repeated imitation should be easier to generalize to new category members than transcriptions that are more imitative of a specific category exemplar.

When participants had to generalize the meaning of the novel label to new category members (new sounds), they were faster when the label came from transcriptions of later generation imitations than from transcriptions of first generation imitations, *b* = -114.13 (52.06), *t*(-2.2) = 39.92, *p* = 0.034. Comparing RTs on the trials leading up to a block transition and the trials immediately after the block transition revealed a reliable interaction between block transition and the generation of the transcribed label, *b* = -146.75 (65.47), *t*(-2.2) = 1869.72, *p* = 0.025. This suggests that in addition to becoming more stable both in terms of acoustic and orthographic properties, imitations that have been more repeated may also be easier to learn as category labels.



Response times across blocks in the category learning experiment. Participants learned the labels for sounds in each category. Each new block introduced four new sounds into the categories. Participants who learned labels that were transcriptions from imitations made in later generations were faster to generalize these labels to new category members than participants who learned labels transcribed directly from imitations of the source sounds.



Cost to generalizing to new category members at block transition. RTs on correct trials were summarized for the 6 trials prior to each block transition and the 6 trials after the transition. Participants had an easier time generalizing transcriptions from last generation imitations to the new category members presented after the block transition. Error bars show +/- 1 standard error of the predictions from the linear mixed effects model with random effects for participants.

# Discussion

We show that repeated imitation of an originally imitative vocalization gradually becomes more word-like as it is transmitted along a chain of a “Telephone” game. This claim is supported by \_\_\_\_ lines of evidence. First.. Second… Third. The first evidence provided showed that imitations became more stable over generations of repetition, both in terms of acoustic similarity as well as in orthographic agreement. But more than just becoming more stable over generations, the imitations also become more wordlike in that they began to share a fundamental property of words: that of being category labels. Category information was more resilient to transmission decay than specific information identifying a particular exemplar within a category. This category information remained even when the imitations were transcribed into lexical forms, as participants were able to guess the categorical meaning of the word at above chance levels even after 8 generations of repetition. This suggests that by becoming more stable and more wordlike these imitations also became more wordlike in terms of their consequences for cognition. One such consequence of having words is that they make categorization easier. In support of this conclusion, we found that participants naive to the transmission chain experiment were faster to learn category labels that had emerged through repeated imitation than they learned from transcriptions of direct imitations of the environmental sound, completing the transition from nonverbal imitation to a fully lexicalized word form and demonstrating the impact of this transition on communication.

One result that did not fit squarely with imitations becoming more word-like is that with transcriptions, there was no difference over generations between question types. If the results of matching transcriptions back to seed sounds mirrored the results of matching imitations back to seed sounds we would have expected the difference between True seed questions and Category match questions to decrease over generations. Instead we found a main effect of question type indicating that although participants were able to match transcriptions to categories of sounds even after 8 generations of repetition, it was still easier for them to match a transcription to the actual seed that generated the transcription.

Our study focused on the formation of onomatopoeia—sound-imitative words— but in addition to onomatopoeia, many languages have semantically rich systems of ideophones. These words comprise a grammatically and phonologically distinct class of words that are used to express a variety of sensory-rich meanings (Dingemanse 2012; Voeltz and Kilian-Hatz 2001). Notably, these words are often recognized by native speakers to be somehow imitative of their meaning. For example, in Japanese, the word ‘koron’ – with a voiceless [k] refers to a light object rolling once, the reduplicated ‘korokoro’ to a light object rolling repeatedly, and ‘gorogoro’ – with a voiced [g] – to a heavy object rolling repeatedly (Imai & Kita, 2014). The iconicity of ideophones was verified by an experiment that tested the ability of naïve listeners to guess the meanings of words sampled from five different languages (Dingemanse, Schuerman, and Reinisch 2016). Although words for sounds were guessed more accurately than the rest, listeners were better than chance at guessing the synonyms of ideophones that expressed meanings from all five semantic categories tested -- color/visual, motion, shape, sound, and texture. In addition, laboratory experiments show that people are able to generate imitative vocalizations for a variety of non-sound concepts, and that these are also understandable to naïve listeners (Perlman, Dale, and Lupyan 2015). Thus vocal imitation has the potential to play a role in word formation that extends beyond just the imitation of sounds.

Our findings from an online game of Telephone suggest that the formation of words from vocal imitation can be a simple process. The results show how repeated imitation can create progressively more word-like forms while retaining a resemblance to the original sound that motivated it. This raises the possibility that onomatopoeic words can be created from the repetition of one-shot vocal imitations of an original sound.

# References

Blackwell, Natalia L, Marcus Perlman, and Jean E Fox Tree. 2015. “Quotation as a multimodal construction.” *Journal of Pragmatics* 81 (May). Elsevier B.V.: 1–7. doi:[10.1016/j.pragma.2015.03.004](https://doi.org/10.1016/j.pragma.2015.03.004).

Brown, R W, A H Black, and A E Horowitz. 1955. “Phonetic symbolism in natural languages.” *Journal of Abnormal Psychology* 50 (3): 388–93. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=14381156&retmode=ref&cmd=prlinks>.

Clark, Herbert H, and Richard J Gerrig. 1990. “Quotations as Demonstrations.” *Language, Cognition, and Neuroscience* 66 (4). Linguistic Society of America: 764–805. doi:[10.2307/414729?ref=search-gateway:df58c19715a2e512d551c6fd62e27164](https://doi.org/10.2307/414729?ref=search-gateway:df58c19715a2e512d551c6fd62e27164).

Dingemanse, Mark. 2012. “Advances in the Cross-Linguistic Study of Ideophones.” *Language and Linguistics Compass* 6 (10): 654–72. doi:[10.1002/lnc3.361](https://doi.org/10.1002/lnc3.361).

Dingemanse, Mark, and Kimi Akita. 2016. “An inverse relation between expressiveness and grammatical integration: On the morphosyntactic typology of ideophones, with special reference to Japanese.” *Journal of Linguistics*, October. University of Wisconsin-Madison Libraries, 1–32. doi:[10.1017/S002222671600030X](https://doi.org/10.1017/S002222671600030X).

Dingemanse, Mark, W Schuerman, and E Reinisch. 2016. “What sound symbolism can and cannot do: testing the iconicity of ideophones from five languages.” *Science*. <http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:2286810>.

Donald, Merlin. 2016. “Key cognitive preconditions for the evolution of language.” *Psychonomic Bulletin & Review*, June. Psychonomic Bulletin & Review, 1–5. doi:[10.3758/s13423-016-1102-x](https://doi.org/10.3758/s13423-016-1102-x).

Edmiston, Pierce, and Gary Lupyan. 2015. “What makes words special? Words as unmotivated cues.” *Cognition* 143 (C). Elsevier B.V.: 93–100. doi:[10.1016/j.cognition.2015.06.008](https://doi.org/10.1016/j.cognition.2015.06.008).

Imai, M, and S Kita. 2014. “The sound symbolism bootstrapping hypothesis for language acquisition and language evolution.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1651): 20130298–8. doi:[10.1098/rstb.2013.0298](https://doi.org/10.1098/rstb.2013.0298).

Lemaitre, Guillaume, and Davide Rocchesso. 2014. “On the effectiveness of vocal imitations and verbal descriptions of sounds.” *The Journal of the Acoustical Society of America* 135 (2): 862–73. doi:[10.1121/1.4861245](https://doi.org/10.1121/1.4861245).

Lemaitre, Guillaume, Olivier Houix, Frédéric Voisin, Nicolas Misdariis, and Patrick Susini. 2016. “Vocal Imitations of Non-Vocal Sounds.” *PloS One* 11 (12): e0168167–28. doi:[10.1371/journal.pone.0168167](https://doi.org/10.1371/journal.pone.0168167).

Lewis, J. 2009. “As well as words: Congo Pygmy hunting, mimicry, and play.” In *The Cradle of Language*. The cradle of language. <http://books.google.com/books?hl=en&lr=&id=IVRkzK1VX1oC&oi=fnd&pg=PA236&dq=As+well+as+words+Congo+Pygmy+hunting+mimicry+and+play&ots=vCI59rdgYY&sig=d2svCjyUidFhnlfHmje9SNjgs9s>.

Locke, John. 1948. “An essay concerning human understanding.” In *Readings in the History of Psychology*, edited by Wayne Dennis. Norwalk, CT.

Perlman, Marcus, R Dale, and Gary Lupyan. 2015. “Iconicity can ground the creation of vocal symbols.” *Royal Society Open Science* 2 (8): 150152–16. doi:[10.1098/rsos.150152](https://doi.org/10.1098/rsos.150152).

Pinker, Steven, and Ray Jackendoff. 2005. “The faculty of language: what’s special about it?” *Cognition* 95 (2): 201–36. doi:[10.1016/j.cognition.2004.08.004](https://doi.org/10.1016/j.cognition.2004.08.004).

Plato, and C D C Reeve. 1999. *Cratylus*. Indianapolis: Hackett.

Rhodes, Richard. 1994. “Aural images.” *Sound Symbolism*. Cambridge University Press: Cambridge, UK, 276–92.

Voeltz, FK Erhard, and Christa Kilian-Hatz. 2001. *Ideophones*. Vol. 44. John Benjamins Publishing.