The emergence of words from vocal imitations

Journal:	*Proceedings B*
Manuscript ID	RSPB-2017-2709
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Edmiston, Pierce; University of Wisconsin-Madison, Department of Psychology Perlman, Marcus; University of Birmingham Lupyan, Gary; University of Wisconsin Madison, Psychology
Subject:	Cognition < BIOLOGY, Evolution < BIOLOGY, Behaviour < BIOLOGY
Keywords:	language evolution, iconicity, vocal imitation, transmission chain
Proceedings B category:	Behaviour

SCHOLARONE™
Manuscripts

Running head: WORDS FROM IMITATIONS   1

1   The emergence of words from vocal imitations

2   Pierce Edmiston[1], Marcus Perlman[2], & Gary Lupyan[1]

3   [1] University of Wisconsin-Madison

4   [2] University of Birmingham

5   Author Note

6   Pierce Edmiston and Gary Lupyan, Department of Psychology, University of

7   Wisconsin-Madison, Madison, Wisconsin. Marcus Perlman, Department of English Language

8   and Applied Linguistics, University of Birmingham, United Kingdom.

9   Correspondence concerning this article should be addressed to Pierce Edmiston, 1202

10 W. Johnson St., Madison, WI, 53703. E-mail: pedmiston@wisc.edu

11   Abstract

12   People have long pondered the origins of language, especially the words that compose them.

13   Here, we report a series of experiments investigating how conventional spoken words might

14   emerge from imitations of environmental sounds. Does the repeated imitation of an

15   environmental sound gradually give rise to more word-like forms? In what ways do these

16   words resemble the original sounds that motivated them (i.e., iconicity)? Participants played

17   a version of the children's game "Telephone". The first generation of participants imitated

18   recognizable environmental sounds (e.g., glass breaking, water splashing). Subsequent

19   generations imitated the previous generation of imitations for a maximum of 8 generations.

20   The results showed that the imitations became more stable and word-like, and later

21   imitations were easier to learn as category labels. At the same time, even after 8 generations,

22   both spoken imitations and their written transcriptions could be matched above chance to

23   the category of environmental sound that motivated them. These results show how repeated

24   imitation can create progressively more word-like forms while continuing to retain a

25   resemblance to the original sound that motivated them, and speak to the possible role of

26   human vocal imitation in explaining the origins of at least some spoken words.

27   *Keywords:* language evolution, iconicity, vocal imitation, transmission chain

28   Word count: 6913

₂₉　　　　　　　　　The emergence of words from vocal imitations

₃₀　　　Most vocal communication of non-human primate species is based on species-typical

₃₁ calls that are highly similar across generations and between populations [1] [2]. In contrast,

₃₂ human languages comprise a vast repertoire of learned meaningful elements (words and other

₃₃ morphemes) which can number in the tens of thousands or more [3]. Aside from their

₃₄ number, the words of different natural languages are characterized by their extreme diversity

₃₅ [4–6]. The words used within a speech community change relatively quickly over generations

₃₆ compared to the evolution of vocal signals [7]. At least in part as a consequence of this

₃₇ divergence, most words appear to bear a largely arbitrary relationship between their form

₃₈ and their meaning — seemingly, a product of their idiosyncratic etymological histories [8,9].

₃₉ The apparently arbitrary nature of spoken vocabularies presents a quandary for the study of

₄₀ language origins. If words of spoken languages are truly arbitrary, by what process were the

₄₁ first words ever coined?

₄₂　　　While the origin of most spoken words is hard to discern, the situation is somewhat

₄₃ different for signed languages. In signed languages, the origins of many signs are relatively

₄₄ transparent. Although signed languages rely on the same type of referential symbolism as

₄₅ spoken languages, many individual signs have clear iconic roots, formed from gestures that

₄₆ resemble their meaning [10–12]. For instance, [13] noted the iconic origins of the American

₄₇ Sign Language (ASL) sign for bird, which is formed with a beak-like handshape articulated

₄₈ in front of the nose. Another example is steal, derived from a grabbing motion to represent

₄₉ the act of stealing something. [14] identified about 25% of American Sign Language signs to

₅₀ be iconic, and reviewing the remaining 75% of ASL signs, [15] determined that about

₅₁ two-thirds of these seemed plausibly derived from iconic origins. Further support for iconic

₅₂ origins of signed languages comes from observations of deaf children raised without exposure

₅₃ to a signed language, who develop homesign systems to use with their family. These

₅₄ communication systems are generally built from a process in which the children establish

₅₅ conventional gestures through the use of pantomimes and various iconic and indexical

gestures [16]. Participants in laboratory experiments utilize a similar strategy when they communicate with gestures in iterated communication games [17].

In contrast to the visual gestures of signed languages, many have argued that iconic vocalizations could not have played a significant role in the origin of spoken words because the vocal modality simply does not afford much resemblance between form and meaning [18–23]. It has also been argued that the human capacity for vocal imitation is a domain-specific skill, geared towards learning to speak, rather than the representation of environmental sounds. For example, [24] suggested that, "most humans lack the ability... to convincingly reproduce environmental sounds... Thus 'capacity for vocal imitation' in humans might be better described as a capacity to learn to produce speech" (p. 209). Consequently, it is still widely assumed that vocal imitation — or more broadly, the use of any sort of resemblance between form and meaning — cannot be important to understanding the origin of spoken words.

Although most words of contemporary spoken languages are not clearly imitative in origin, there has been a growing recognition of the importance of iconicity in spoken languages [25,26] and the common use of vocal imitation and depiction in spoken discourse [27,28]. This has led some to argue for the importance of imitation for understanding the origin of spoken words [29–33]. In addition, counter to previous assumptions, people are highly effective at using vocal imitations to refer to environmental sounds such as coins dropping in a jar or mechanical events such as scraping — in some cases, even more effective than when using conventional words [34]. Recent work has also shown that people are able to create novel imitative vocalizations for more abstract meanings (e.g. "slow", "rough", "good", "many") that are understandable to naïve listeners [33]. These imitations are effective not because people can mimic environmental sounds with high fidelity, but because people are able to produce imitations that capture the salient features of sounds in ways that are understandable to listeners [35]. Similarly, the features of onomatopoeic words might highlight distinctive aspects of the sounds they represent. For example, the initial voiced,

83 plosive /b/ in "boom" represents an abrupt, loud onset, the back vowel /u/ a low pitch, and

84 the nasalized /m/ a slow, muffled decay [36].

85   Thus, converging evidence suggests that people can use vocal imitation as an effective

86 means of communication. At the same time, vocal imitations are not words. If vocal

87 imitation played a role in the origin of some spoken words, then it is necessary to identify the

88 minimal conditions under which vocal imitations can give rise to more word-like vocalizations

89 that can eventually be integrated into a vocabulary of a language. In the present set of

90 studies we ask whether vocal imitations can transition to more word-like forms through sheer

91 repetition — without an explicit intent to communicate. To answer this question, we

92 recruited participants to play an online version of the children's game of "Telephone". In the

93 children's game, a spoken message is whispered from one person to the next. In our version,

94 the original message or "seed sound" was a recording of an environmental sound. The initial

95 group of participants (first generation) imitated these seed sounds, the next generation

96 imitated the previous imitators, and so on for up to 8 generations.

97   Our approach uses a transmission chain methodology similar to that frequently used in

98 experimental studies of language evolution [37]. As with other transmission chain studies

99 (and iterated learning studies more generally), we seek to discover how various biases and

100 constraints of individuals change the nature of a linguistic signal. Importantly, while typical

101 transmission chain studies focus on the impact of learning biases [38], the present studies

102 involve iterated reproduction that does not involve any learning. Participants simply

103 attempt to imitate a sound as best as they can. The biases we hypothesize to drive

104 vocalizations to become more word-like are therefore not related to any learning process, but

105 instead are expected to emerge from constraints on the reproducibility of vocalizations. Our

106 aim is thus to determine whether iterated reproduction, even without learning, is a sufficient

107 enough constraint to enable the emergence of more word-like signals.

108   After collecting the imitations, we conducted a series of analyses and additional

109 experiments to systematically answer the following questions: First, do imitations stabilize in

form and become more word-like as they are repeated? Second, do the imitations retain a resemblance to the original environmental sound that inspired them? If so, it should be possible for naïve participants to match the emergent words back to the original seed sounds. Third, do the imitations become more suitable as categorical labels for the sounds that motivated them? For example, does the imitation of a particular water-splashing sound become, over generations of repeated imitation, a better label for the more general category of water-splashing sounds?

Experiment 1: Stabilization of imitations through repetition

In the first experiment, we collected the vocal imitations, and assessed the extent to which repeating imitations of environmental sounds results in progressive stabilization toward more word-like forms in three ways. First, we measured changes in the perception of acoustic similarity between subsequent generations of imitations. Second, we used algorithmic measures of acoustic similarity to assess the similarity of imitations sampled within and between transmission chains. Third, we obtained transcriptions of imitations, and measured the extent to which later generation imitations were transcribed with greater consistency and agreement. The results show that repeated imitation results in vocalizations that are easier to repeat with high fidelity and more consistently transcribed into English orthography.

Methods

Selecting seed sounds. To avoid sounds with lexicalized or conventionalized onomatopoeic forms in English, we used inanimate categories of environmental sounds. To select sounds that were equally distinguishable within each category, we used an odd-one-out norming procedure (N=105 participants; see Fig. S1), resulting in a final set of 16 sounds, 4 in each of 4 categories: glass, tear, water, zipper.

Collecting vocal imitations. Participants (N=94) recruited from Amazon Mechanical Turk were paid to participate in an online version of the children's game of "Telephone". Participants were instructed that they would hear some sound and their task

136 was to reproduce it as accurately as possible using their computer microphone. Full

137 instructions are provided in the Supplemental Materials.

138   Each participant listened to and imitated four sounds: one from each of the four

139 categories of environmental sounds. Sounds were assigned at random such that participants

140 were unlikely to imitate the same person more than once. Participants were allowed to listen

141 to each target sound as many times as they wished, but were only allowed a single recording

142 in response. Recordings that were too quiet (less than -30 dBFS) were not accepted.

143   Imitations were monitored by an experimenter to remove poor quality recordings (e.g.,

144 loud background sounds), and recordings that violated the rules of the experiment (e.g., an

145 utterance in English). A total of 115 (24%) imitations were removed. The final sample

146 contained 365 imitations along 105 contiguous transmission chains (Fig. 1).
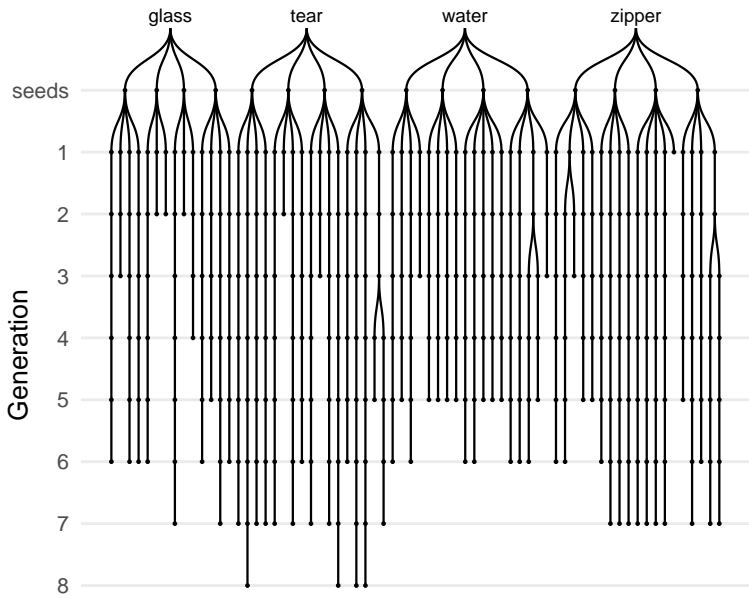

Figure 1. Vocal imitations collected in the transmission chain experiment. Seed sounds (16) were sampled from four categories of environmental sounds: glass, tear, water, zipper. Participants imitated each seed sound, and then the next generation of participants imitated the imitations, and so on, for up to 8 generations. Chains are unbalanced due to random assignment and the exclusion of some low quality recordings.

WORDS FROM IMITATIONS   8

147   **Measuring acoustic similarity.**   Acoustic similarity judgments were obtained from

148 five research assistants who listened to pairs of sounds (approx. 300) and rated their

149 subjective similarity. On each trial, raters heard two sounds from subsequent generations

150 played in random order, and indicated the similarity between the sounds on a 7- point Likert

151 scale from *Entirely different and would never be confused* to *Nearly identical.* Full

152 instructions and inter-rater reliability measures are provided in the Supplemental Materials.

153   To obtain algorithmic measures of acoustic similarity, we used the acoustic distance

154 functions included in Phonological Corpus Tools [39]. We computed Mel-frequency cepstral

155 coefficients (MFCCs) between pairs of imitations using 12 coefficients in order to obtain

156 speaker-independent estimates.

157   **Collecting transcriptions of imitations.**   Transcriptions were obtained for the

158 first and last three generations of each transmission chain. Additional "transcriptions" of the

159 original sounds used as seeds were also collected and are analyzed in the Supplementary

160 Materials (Fig. S6).

161   Participants (N=216) recruited from Amazon Mechanical Turk were paid to listen to

162 imitations and write down what they heard as a single "word" so that the written word would

163 sound as much like the sound as possible. Participants were instructed to avoid transcribing

164 the imitations into existing English words. Each participant completed 10 transcriptions.

165 **Results**

166   Imitations of environmental sounds became more stable over the course of being

167 repeated as revealed by increasing acoustic similarity judgments along individual

168 transmission chains. Acoustic similarity ratings were fit with a linear mixed-effects model

169 predicting perceived acoustic similarity from generation with random effects (intercepts and

170 slopes) for raters. To test whether the hypothesized increase in acoustic similarity was true

171 across all seed sounds and categories, we added random effects (intercepts and slopes) for

172 seed sounds nested within categories. The results showed that, across raters and seeds,

WORDS FROM IMITATIONS   9

173 imitations from later generations were rated as sounding more similar to one another than

174 imitations from earlier generations, $b = 0.10$ (SE $= 0.03$), $t(11.9) = 3.03$, $p = 0.011$ (Fig. 2).

175 This result suggests that imitations became more stable (i.e., easier to imitate with high

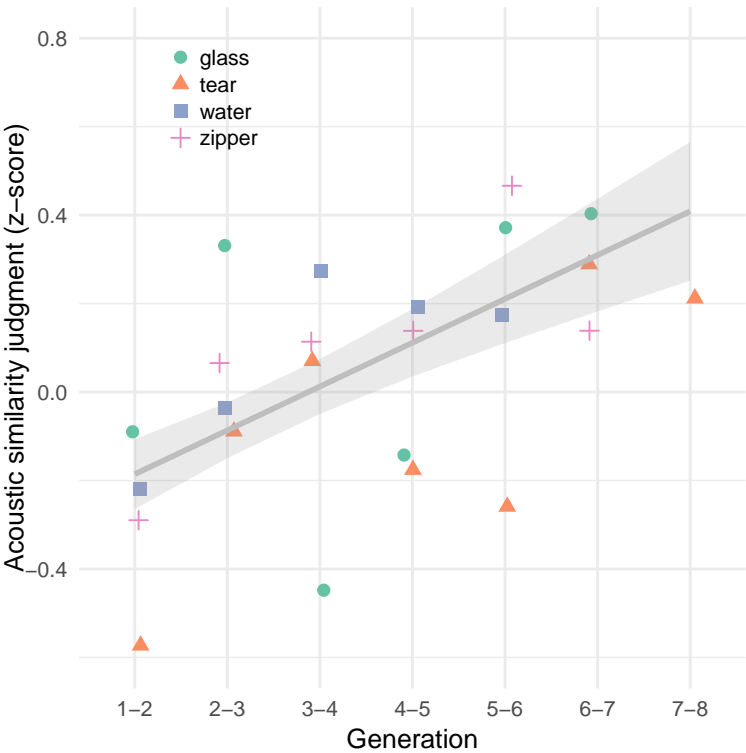176 fidelity) with each generation of repetition.

Figure 2. Change in perception of acoustic similarity over generations of iterated imitation.
Points depict mean acoustic similarity ratings for pairs of imitations in each category. The
predictions of the linear mixed-effects model are shown with ±1 SE.

177   Increasing similarity along transmission chains could also reflect the continuous

178 degradation of the signal due to repeated imitation, in which case acoustic similarity would

179 increase both within as well as between chains. To test this, we calculated MFCCs for pairs

180 of sounds sampled from within and between transmission chains across categories, and fit a

181 linear model predicting acoustic similarity from the generation of sounds. We found that

182 acoustic similarity increased within chains more than it increased between chains, $b = $ -0.07

183 (SE $= 0.03$), $t(6674.0) = $ -2.13, $p = 0.033$ (Fig. S2), indicating that imitations were

WORDS FROM IMITATIONS   10

184   stabilizing on divergent acoustic forms as opposed to converging on similar forms through

185   continuous degradation.

186   An additional test of stabilization and word-likeness was to measure whether later

187   generation imitations were transcribed more consistently than first generation imitations.

188   We collected a total of 2163 transcriptions — approximately 20 transcriptions per sound. Of

189   these, 179 transcriptions (8%) were removed because they contained English words. Some

190   examples of the final transcriptions are presented in Table 1.

Table 1

Examples of words transcribed from imitations.

Category	First generation	Last generation
glass	dirrng	wayew
tear	feeshefee	cheecheea
water	boococucuwich	galong
zipper	bzzzzup	izzip

191   To measure the similarity among transcriptions for a given imitation, we calculated the

192   average orthographic distance between the most frequent transcription and all other

193   transcriptions of the same imitation. We then fit a hierarchical linear model predicting

194   orthographic distance from the generation of the imitation (First generation, Last

195   generation) with random effects (intercepts and slopes) for seed sound nested within

196   category. The results showed that transcriptions of last generation imitations were more

197   similar to one another than transcriptions of first generation imitations, $b = -0.12$ (SE =

198   0.03), $t(3.0) = -3.62$, $p = 0.035$ (Fig. S3). The same result is reached through alternative

199   measures of orthographic distance (Fig. S4). Differences between transcriptions of human

200   vocalizations and transcriptions directly of environmental sound cues are reported in the

201   Supplementary Materials (Fig. S6).

202 **Discussion**

203   Repeating imitations of environmental sounds over generations of imitators was

204 sufficient to create more word-like forms, even without any explicit intent to communicate.

205 We defined word-likeness in terms of acoustic stability and orthographic agreement. With

206 each repetition, the acoustic forms of the imitations became more similar to one another,

207 indicating they became easier to repeat with high fidelity. The possibility that this similarity

208 was due to uniform degradation across all transmission chains was ruled out by algorithmic

209 analyses of acoustic similarity demonstrating that acoustic similarity increased within chains

210 but not between them. Additionally, later generation imitations were transcribed more

211 consistently into English orthography, further supporting our hypothesis that repeating

212 imitations makes them more stable and word-like.

213   The results of Experiment 1 demonstrate the ease with which iterated imitation gives

214 rise to stable wordforms. However, the results do not address how these emergent words

215 relate to the original sounds that were being imitated. As the imitations became more

216 word-like, were they stabilizing on arbitrary acoustic and orthographic forms, or did they

217 maintain some resemblance to the environmental sounds that motivated them? The purpose

218 of Experiment 2 was to assess the extent to which repeated imitations and their

219 transcriptions maintained a resemblance to the original set of seed sounds.

220   **Experiment 2: Resemblance of imitations to original seed sounds**

221   To assess the resemblance of repeated imitations to the original seed sounds, we

222 measured the ability of participants naïve to the design of the experiment to match

223 imitations and their transcriptions back to their original sound source relative to other seed

224 sounds from either the same category or from different categories (Fig. 3A). Using these

225 match accuracies, we first asked whether and for how many generations the imitations and

226 their transcriptions could be matched back to the original sounds. Second, we asked whether

227 repeated imitation resulted in a uniform degradation of the signal in each imitation, or if

repeated imitation resulted in some kinds of information degrading more rapidly than others. Specifically, we tested the hypothesis that if imitations were becoming more word-like, then they should also be interpreted more categorically, and thus we anticipated that imitations would lose information identifying the specific source of an imitation more rapidly than category information that identifies the category of environmental sound being imitated.

Methods

Matching imitations to seed sounds. Participants (N=751) recruited from Amazon Mechanical Turk were paid to listen to imitations, one at a time, and for each one, choose one of four possible sounds they thought the person was trying to imitate. The task was not speeded and no feedback was provided. Participants completed 10 questions at a time.

All imitations were tested in each of the three question types depicted in Fig. 3A. These questions differed in the relationship between the imitation and the four seed sounds provided as the choices in the question. Question types (True seed, Category match, Specific match) were assigned between-subject.

Matching transcriptions to seed sounds. Participants (N=461) recruited from Amazon Mechanical Turk completed a modified version of the matching survey described above. Instead of listening to imitations, participants now read a word (a transcription of an imitation), which they were told was invented to describe one of the four presented sounds. Of the unique transcriptions that were generated for each sound (imitations and seed sounds), only the top four most frequent transcriptions were used in the matching experiment. The distractors for all questions were between-category, i.e. true seed and category match. Specific match questions were omitted.

Results

Response accuracies in matching imitations to seed sounds were fit by a generalized linear mixed-effects model predicting match accuracy as different from chance (25%) based

254 on the type of question being answered (True seed, Category match, Specific match) and the

255 generation of the imitation. Question types were contrast coded using Category match

256 questions as the baseline condition in comparison to the other two question types, each

257 containing the actual seed that generated the imitation as one of the choices. The model

258 included random intercepts for participant, and random slopes and intercepts for seed sounds

259 nested within categories.

260   Accuracy in matching first generation imitations to seed sounds was above chance for

261 all question types, $b = 1.65$ (SE = 0.14) log-odds, odds = 0.50, $z = 11.58$, $p < 0.001$, and

262 decreased steadily over generations, $b = -0.16$ (SE = 0.04) log-odds, $z = -3.72$, $p < 0.001$.

263 We then tested whether this increase in difficulty was constant across the three types of

264 questions or if some question types became more difficult than others. The results are shown

265 in Fig. 3B. Performance decreased over generations more rapidly for questions requiring a

266 within-category distinction than for between-category questions, $b = -0.08$ (SE = 0.03)

267 log-odds, $z = -2.68$, $p = 0.007$, suggesting that between-category information was more

268 resistant to loss through repeated imitation.

269   An alternative explanation of the drop off in accuracy for within-category questions

270 but not category match questions is that the within-category questions are simply more

271 difficult because the sounds presented as choices are more acoustically similar to one another.

272 However, performance also decreased relative to the category match questions for the easiest

273 type of question where the correct answer was the actual seed generating the imitation (True

274 seed questions; see Fig. 3A). That is, the advantage of having the true seed among

275 between-category distractors decreased over generations, $b = -0.07$ (SE = 0.02) log-odds, $z =$

276 -2.77, $p = 0.006$. The observed decrease in the "true seed advantage" (the advantage of

277 having the actual seed among the choices) combined with the increase in the "category

278 advantage" (the advantage of having between-category distractors) shows that the changes

279 induced by repeated imitation caused the imitations to lose some of properties that linked

280 the earlier imitations to the specific sound that motivated them, while nevertheless

281   preserving a more abstract category-based resemblance.

282   We next report the results of matching the written transcriptions of the auditory

283   sounds back to the original environmental sounds. Remarkably, participants were able to

284   guess the correct meaning of a word that was transcribed from an imitation that had been

285   repeated up to 8 times, $b = 0.83$ (SE $= 0.13$) log-odds, odds $= $ -0.18, $z = 6.46$, $p < 0.001$

286   (Fig. 3C). This was true for True seed questions containing the actual seed generating the

287   transcribed imitation, $b = 0.75$ (SE $= 0.15$) log-odds, $z = 4.87$, $p < 0.001$, and for Category

288   match questions where participants had to associate transcriptions with a particular category

289   of environmental sounds, $b = 1.02$ (SE $= 0.16$) log-odds, $z = 6.39$, $p < 0.001$. The effect of

290   generation did not vary across these question types, $b = 0.05$ (SE $= 0.10$) log-odds, $z = 0.47$,

291   $p = 0.638$. The results of matching "transcriptions" directly of the environmental sounds are

292   shown in Fig. S6.

293   **Discussion**

294   Even after being repeated up to 8 times across 8 different individuals, vocalizations

295   retained a resemblance to the environmental sound that motivated them. This resemblance

296   remained even after the vocalizations were transcribed into orthographic forms. For vocal

297   imitations, but not for transcriptions, this resemblance was stronger for the category of

298   environmental sound than the actual seed sound, suggesting that iterated imitation produces

299   vocalizations that are interpreted by naïve listeners in a more categorical way. Iterated

300   imitation appears to strip the vocalizations of some of the characteristics that individuate

301   each particular sound while maintaining some category-based resemblance (even though

302   participants were never informed about the meaning of the vocalizations and were not trying

303   to communicate).

304   Transcriptions of the vocalizations, like the vocalizations themselves, were able to be

305   matched to the original environmental sounds at levels above chance. Unlike vocalizations,

306   the transcriptions continued to be matched more accurately to the true seed compared to the
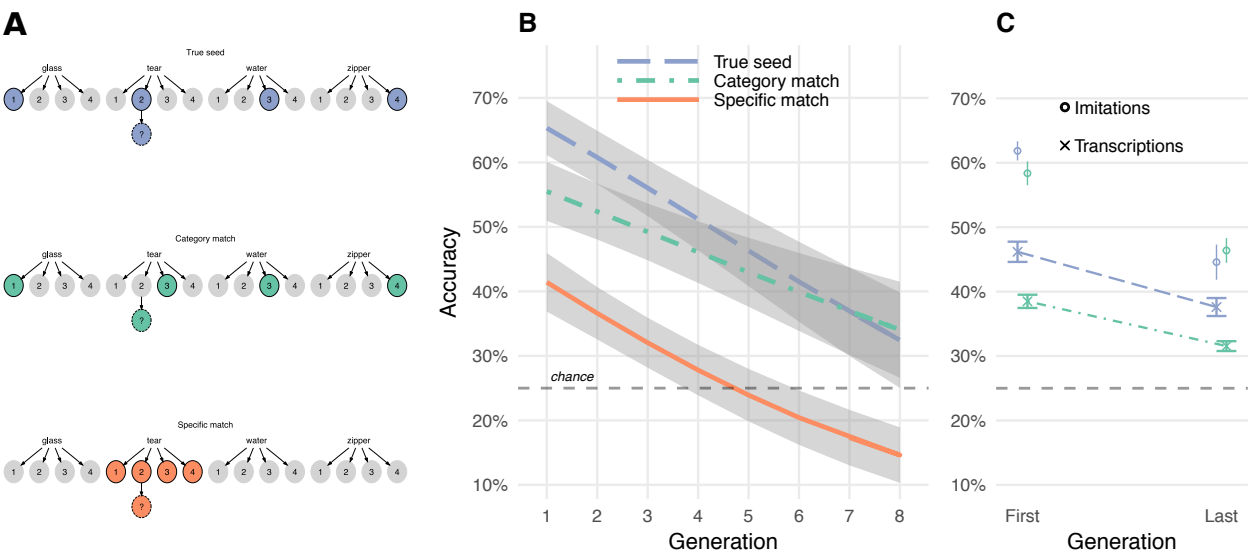
Figure 3. Repeated imitations retained category resemblance. A. Three types of matching questions. True seed and category match questions had choices from different sound categories. Specific match questions pitted the actual seed against the other seeds within the same category. B. Accuracy in matching vocal imitations to original seed sounds. Curves show predictions of the generalized linear mixed effects models with ± 1 SE of the model predictions. C. Accuracy in matching transcriptions of the imitations to original seed sounds (e.g., "boococucuwich" to a water splashing sound). Circles show mean matching accuracy for the vocal imitations that were transcribed for comparison.

307   general category. That is, transcription appears to impact specific and category-level

308   information equally. One possible explanation of the difference between the acoustic and

309   orthographic forms of this task is that the process of transcribing a non- linguistic

310   vocalization into a written word encourages transcribers to emphasize individuating

311   information about the vocalization. However, the fact that transcriptions of imitations can

312   be matched back to other category members (Category match questions) suggests that

313   transcriptions still carry some category information, so this is not a complete explanation of

314   our results. Another possible reason is that by selecting only the most frequent

315   transcriptions, we unintentionally excluded less frequent transcriptions that were nonetheless

₃₁₆ more diagnostic of category information.

₃₁₇   Experiments 1 and 2 document a process of gradual change from an imitation of an

₃₁₈ environmental sound to a more word-like form. But do these emergent words function like

₃₁₉ other words in the language? In Experiment 3, we test the suitability of words taken from

₃₂₀ the beginning and end of transmission chains in serving as category labels in a category

₃₂₁ learning task.

₃₂₂   **Experiment 3: Suitability of created words as category labels**

₃₂₃   One consequence of imitations becoming more word-like is that they may make for

₃₂₄ better category labels. For example, an imitation from a later generation, by virtue of having

₃₂₅ a more word-like form, may be easier to learn as a label for the category of sounds that

₃₂₆ motivated it than an earlier imitation, which is more closely yoked to a particular

₃₂₇ environmental sound. To the extent that repeating imitations abstracts away the

₃₂₈ idiosyncrasies of a particular category member [40,41], it may also be easier to generalize to

₃₂₉ new category members. We tested these predictions using a category learning task in which

₃₃₀ participants learned novel labels for the categories of environmental sounds. The novel labels

₃₃₁ were transcriptions of either first or last generation imitations gathered in Experiment 1.

₃₃₂ **Methods**

₃₃₃   **Selecting words to learn as category labels.**   Of the 1814 unique words created

₃₃₄ through the transmission chain and transcription procedures, we sampled 56 words

₃₃₅ transcribed from first and last generation imitations that were equated in terms of length

₃₃₆ and match accuracy with the original sounds. Our procedure for selecting otherwise-equal

₃₃₇ transcriptions is detailed in the Supplementary Materials.

₃₃₈   **Procedure.**   Participants ($N{=}67$) were University of Wisconsin undergraduates who

₃₃₉ received course credit for participation. Participants were randomly assigned four novel

₃₄₀ labels to learn for four categories of environmental sounds. Full instructions are provided in

₃₄₁ the Supplementary Materials. Participants were assigned between-subject to learn labels

342 (transcriptions) of either first or last generation imitations. On each trial, participants heard

343 one of the 16 seed sounds. After a 1s delay, participants saw a label (one of the transcribed

344 imitations) and responded yes or no using a gamepad controller depending on whether the

345 sound and the word went together. Participants received accuracy feedback (a bell sound

346 and a green checkmark if correct; a buzzing sound and a red "X" if incorrect). Four outlier

347 participants were excluded from the final sample due to high error rates and slow RTs.

348 Participants categorized all 16 seed sounds over the course of the experiment, but they

349 learned them in blocks of 4 sounds at a time. Within each block of 24 trials, participants

350 heard the same four sounds and the same four words multiple times, with a 50% probability

351 of the sound matching the word on any given trial. At the start of a new block of trials,

352 participants heard four new sounds they had not heard before, and had to learn to associate

353 these new sounds with the words they had learned in the previous blocks.

354 **Results**

355 Participants began by learning through trial-and-error to associate four written labels

356 with four categories of environmental sounds. The small number of categories made this an

357 easy task (mean accuracy after the first block of 24 trials was 81%; Fig. S5). Participants

358 learning transcriptions of first or last generation imitations did not differ in overall accuracy,

359 $p = 0.887$, or reaction time, $p = 0.616$.

360 After this initial learning phase (i.e. after the first block of trials), accuracy

361 performance quickly reached ceiling and did not differ between groups $p = 0.775$. However,

362 the response times of participants learning last generation transcriptions declined more

363 rapidly with practice than participants learning first generation transcriptions, $b = $ -114.13

364 (SE = 52.06), $t(39.9) = $ -2.19, $p = 0.034$ (Fig. 4A). These faster responses suggest that, in

365 addition to becoming more stable both in terms of acoustic and orthographic properties,

366 repeating imitations makes them easier to process as category labels. We predict that given

367 a harder task (i.e., more than four categories and 16 exemplars) would yield differences in

368 initial learning rates as well.

369   Next, we examined whether transcriptions from last generation imitations were easier

370 to generalize to novel category exemplars. To test this hypothesis, we compared RTs on

371 trials immediately prior to the introduction of novel sounds (new category members) and the

372 first trials after the block transition (±6 trials). The results revealed a reliable interaction

373 between the generation of the transcribed imitation and the block transition, $b = -110.77$

374 (SE = 52.84), $t(39.7) = -2.10$, $p = 0.042$ (Fig. 4B). This result suggests that transcriptions

375 from later generation imitations were easier to generalize to new category members.
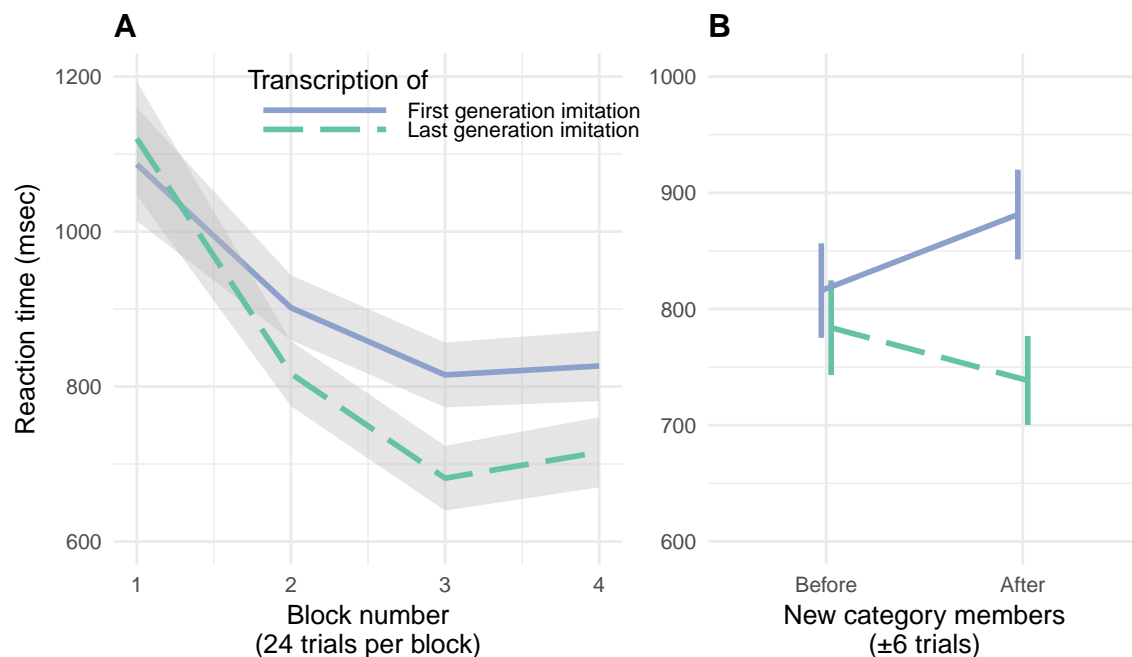

Figure 4. Repeated imitations made for better category labels. A. Mean RTs for correct responses in the category learning experiment with ±1 SE. B. Cost of generalizing to new category members with ±1 SE.

376 **Discussion**

377   The results of a simple category learning experiment demonstrate a possible benefit to

378 the stabilization of repeated imitations on more word-like forms. As a consequence of being

379 more word-like, repeated imitations were responded to more quickly, and generalized to new

380　category members more easily. These results suggest an advantage to repeating imitations

381　from the perspective of the language learner in that they afford better category

382　generalization.

General Discussion

384　　　Accumulating evidence shows that iconic words are prevalent across the spoken

385　languages of the world [25,26,32]. And counter to past assumptions about the limitations of

386　human vocal imitation, people are surprisingly effective at using vocal imitation to represent

387　and communicate about the sounds in their environment [35] and more abstract meanings

388　[33]. These findings raise the hypothesis that early spoken words originated from vocal

389　imitations, perhaps comparable to the way that many of the signs of signed languages

390　appear to be formed originally from pantomimes [33,42]. Here, we examined whether simply

391　repeating an imitation of an environmental sound — with no intention to create a new word

392　or even to communicate — produces more word-like forms.

393　　　Our results show that through unguided repetition, imitative vocalizations became

394　more word-like both in form and function. In form, the vocalizations gradually stabilized

395　over generations, becoming more similar from imitation to imitation. The standardization

396　was also found when the words were transcribed into the English alphabet. Even as the

397　vocalizations became more word-like, they maintained a resemblance to the original

398　environmental sounds that motivated them. Notably, this resemblance appeared to be

399　greater with respect to the category of sound (e.g., water-splashing sounds), rather than to

400　the specific exemplar (a particular water-splashing sound). After eight generations the

401　vocalizations could no longer be matched to the particular sound from which they originated

402　any more accurately than they could be matched to the general category of environmental

403　sound. Thus, information that distinguished an imitation from other sound categories was

404　more resilient to transmission decay than exemplar information within a category.

405　Remarkably, the resemblance to the original sounds was maintained even when the

vocalizations were transcribed into a written form: participants were able to match the transcribed vocalizations to the original sound category at levels above chance.

We further tested the hypothesis that repeated imitation led to vocalizations becoming more word-like by testing the ease with which people learned the (transcribed) vocalizations as category labels (e.g., "pshfft" from generation 1 vs. "shewp" from generation 8 as labels for tearing sounds) (Exp. 3). Labels from the last generation were responded to more quickly than labels from the first generation. More importantly the labels from the last generation generalized better to novel category members. This fits with previous research showing that the relatively arbitrary forms that are typical of words (e.g. "dog") makes them better suited to function as category labels compared to direct auditory cues (e.g., the sound of a dog bark) [40,41,43].

Even as the vocalizations became more word-like, they nevertheless maintained an imitative quality. After eight generations they could no longer be matched to the particular sound from which they originated any more accurately than they could be matched to the general category of environmental sound. Thus, information that distinguished an imitation from other sound categories was more resilient to transmission decay than exemplar information within a category. Remarkably, even after the vocalizations were transcribed into English orthography, participants were able to guess their original sound category from the written "words". In contrast to the vocalizations, participants continued to be more accurate at matching late generation transcriptions back to their particular source sound relative to other exemplars from the same category.

Unlike the large number of iconic signs in signed languages [10], the number of iconic words in spoken languages may appear to be very small [44,45]. However, increasing evidence from disparate language suggests that vocal imitation is, in fact, a widespread source of vocabulary. Cross-linguistic surveys indicate that onomatopoeia—iconic words used to represent sounds—are a universal lexical category found across the world's languages [46]. Even English, a language that has been characterized as relatively limited in iconic

433   vocabulary [47], is documented as having hundreds of onomatopoeic words not only for

434   animal and human vocalizations ("meow", "tweet", "slurp", "babble", murmur"), but also for

435   a variety of environmental sounds (e.g., "ping", "click", "plop") [36,48]. Besides words that

436   directly resemble sounds — the focus of the present study — many languages contain

437   semantically broader inventories of ideophones. These words comprise a grammatically and

438   phonologically distinct class of words that are used to express various sensory-rich meanings,

439   such as qualities related to manner of motion, visual properties, textures and touch, inner

440   feelings and cognitive states [46,49,50]. As with onomatopoeia, ideophones are often

441   recognized by naïve listeners as bearing a degree of resemblance to their meaning [51].

442   Our study focused on imitations of environmental sounds, and more work remains to

443   be done to determine the extent to which vocal imitation can ground de novo vocabulary

444   creation in other semantic domains [33,52]. Notably, our hypothesis that vocal imitation may

445   have played a role in the origin of some of the first spoken words does not preclude that

446   gesture played an equal or more important role in establishing the first linguistic conventions

447   [10,11,53]. What the present results make clear is that the transition from imitation to word

448   can be a rapid and simple process: the mere act of repeated imitation can drive vocalizations

449   to become more word-like in both form and function while still retaining some resemblance

450   to the real world referents.

Ethics

452   This was approved by the University of Wisconsin-Madison's Educational and

453   Social/Behavioral Sciences Institutional Review Board and conducted in accordance with the

454   principles expressed in the Declaration of Helsinki. Informed consent was obtained for all

455   participants.

Data, code, and materials

457   Our data along with all methods, materials, and analysis scripts, are available in public

458   repositories described on the Open Science Framework page for this research here:

osf.io/3navm.

Competing interests

Authors' contributions

Funding

References

1. Seyfarth RM, Cheney DL. 1986 Vocal development in vervet monkeys. *Animal Behaviour* **34**, 1640–1658.

2. Crockford C, Herbinger I, Vigilant L, Boesch C. 2004 Wild chimpanzees produce group-specific calls: a case for vocal learning? *Ethology* **110**, 221–243.

3. Brysbaert M, Stevens M, Mandera P, Keuleers E. 2016 How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology* **7**, 55–11.

4. Wierzbicka A. 1996 *Semantics: Primes and universals: Primes and universals.* Oxford University Press, UK.

5. Evans N, Levinson SC. 2009 The myth of language universals: Language diversity and its importance for cognitive science. *Brain and Behavioral Sciences* **32**, 429–492.

6. Lupyan G, Dale R. 2016 Why are there different languages? The role of adaptation in linguistic diversity.

7. Pagel M, Atkinson QD, Meade A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720.

8. Sapir E. 1921 *Language: An introduction to the study of speech.* New York: Harcourt, Brace; Company.

9. Labov W. 1972 *Sociolinguistic patterns.* University of Pennsylvania Press.

10. Goldin-Meadow S. 2016 What the hands can tell us about language emergence. *Psychonomic Bulletin & Review* **24**, 1–6.

11. Kendon A. 2014 Semiotic diversity in utterance production and the concept of 'language'. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130293–20130293.

12. Klima ES, Bellugi U. 1980 *The signs of language.* Harvard University Press.

13. Frishberg N. 1975 Arbitrariness and Iconicity: Historical Change in American Sign Language. *Language* **51**, 696–719.

14. Stokoe W. 1965 *Dictionary of the American Sign Language based on scientific principles.* Gallaudet College Press, Washington.

15. Wescott RW. 1971 Linguistic iconism. *Linguistic Society of America* **47**, 416–428.

16. Goldin-Meadow S, Feldman H. 1977 The development of language-like communication without a language model. *Science* **197**, 401–403.

17. Fay N, Lister CJ, Ellison TM, Goldin-Meadow S. 2014 Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology* **5**, 663.

18. Arbib MA. 2012 *How the brain got language: The mirror system hypothesis.* Oxford University Press.

19. Armstrong DF, Wilcox S. 2007 *The gestural origin of language.* Oxford University Press.

20. Corballis MC. 2003 *From hand to mouth: The origins of language.* Princeton University Press.

21. Hewes GW. 1973 Primate Communication and the Gestural Origin of Language. *Current Anthropology* **14**, 5–24.

22. Hockett CF. 1978 In search of Jove's brow. *American speech* **53**, 243–313.

23. Tomasello M. 2010 *Origins of human communication.* MIT press.

510 24. Pinker S, Jackendoff R. 2005 The faculty of language: what's special about it? *Cognition*

511   **95**, 201–236.

512 25. Dingemanse M, Blasi DE, Lupyan G, Christiansen MH, Monaghan P. 2015 Arbitrariness,

513   Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences* **19**, 603–615.

514 26. Perniss P, Thompson RL, Vigliocco G. 2010 Iconicity as a General Property of Language:

515   Evidence from Spoken and Signed Languages. *Frontiers in Psychology* **1**.

516 27. Clark HH, Gerrig RJ. 1990 Quotations as demonstrations. *Language* **66**, 764–805.

517 28. Lewis J. 2009 As well as words: Congo Pygmy hunting, mimicry, and play. In *The cradle*

518   *of language*, The cradle of language.

519 29. Brown RW, Black AH, Horowitz AE. 1955 Phonetic symbolism in natural languages.

520   *Journal of abnormal psychology* **50**, 388–393.

521 30. Dingemanse M. 2014 Making new ideophones in Siwu: Creative depiction in

522   conversation. *Pragmatics and Society*

523 31. Donald M. 2016 Key cognitive preconditions for the evolution of language. *Psychonomic*

524   *Bulletin & Review*, 1–5.

525 32. Imai M, Kita S. 2014 The sound symbolism bootstrapping hypothesis for language

526   acquisition and language evolution. *Philosophical Transactions of the Royal Society*

527   *B: Biological Sciences* **369**.

528 33. Perlman M, Dale R, Lupyan G. 2015 Iconicity can ground the creation of vocal symbols.

529   *Royal Society Open Science* **2**, 150152–16.

530 34. Lemaitre G, Rocchesso D. 2014 On the effectiveness of vocal imitations and verbal

531   descriptions of sounds. *The Journal of the Acoustical Society of America* **135**,

532   862–873.

533 35. Lemaitre G, Houix O, Voisin F, Misdariis N, Susini P. 2016 Vocal Imitations of

534   Non-Vocal Sounds. *PloS one* **11**, e0168167–28.

535 36. Rhodes R. 1994 Aural images. *Sound symbolism*, 276–292.

536 37. Tamariz M. 2017 Experimental Studies on the Cultural Evolution of Language. *Annual*

537   *Review of Linguistics* **3**, 389–407.

538   38. Kirby S, Cornish H, Smith K. 2008 Cumulative cultural evolution in the laboratory: an

539   experimental approach to the origins of structure in human language. *Proceedings of*

540   *the National Academy of Sciences* **105**, 10681–10686.

541   39. Hall KC, Allen B, Fry M, Mackie S, McAuliffe M. 2016 Phonological CorpusTools. *14th*

542   *Conference for Laboratory Phonology*

543   40. Edmiston P, Lupyan G. 2015 What makes words special? Words as unmotivated cues.

544   *Cognition* **143**, 93–100.

545   41. Lupyan G, Thompson-Schill SL. 2012 The evocative power of words: Activation of

546   concepts by verbal and nonverbal means. *Journal of Experimental Psychology:*

547   *General* **141**, 170–186.

548   42. Fay N, Ellison TM, Garrod S. 2014 Iconicity: From sign to system in human

549   communication and language. *Pragmatics and Cognition* **22**, 244–263.

550   43. Boutonnet B, Lupyan G. 2015 Words Jump-Start Vision: A Label Advantage in Object

551   Recognition. *Journal of Neuroscience* **35**, 9329–9335.

552   44. Crystal D. 1987 *The Cambridge Encyclopedia of Language*. Cambridge Univ Press.

553   45. Newmeyer FJ. 1992 Iconicity and generative grammar. *Language*

554   46. Dingemanse M. 2012 Advances in the Cross-Linguistic Study of Ideophones. *Language*

555   *and Linguistics Compass* **6**, 654–672.

556   47. Vigliocco G, Perniss P, Vinson D. 2014 Language as a multimodal phenomenon:

557   implications for language learning, processing and evolution. *Philosophical*

558   *Transactions of the Royal Society B: Biological Sciences* **369**, 20130292–20130292.

559   48. Sobkowiak W. 1990 On the phonostatistics of English onomatopoeia. *Studia Anglica*

560   *Posnaniensia* **23**, 15–30.

561   49. Nuckolls JB. 1999 The case for sound symbolism. *Annual Review of Anthropology* **28**,

562   225–252.

563   50. Voeltz FE, Kilian-Hatz C. 2001 *Ideophones*. John Benjamins Publishing.

564　51. Dingemanse M, Schuerman W, Reinisch E. 2016 What sound symbolism can and cannot

565　　　do: Testing the iconicity of ideophones from five languages. *Language* **92**.

566　52. Lupyan G, Perlman M. 2015 The vocal iconicity challenge! In *The th biennial*

567　　　*protolanguage conference*, Rome, Italy.

568　53. Fay N, Arbib MA, Garrod S. 2013 How to Bootstrap a Human Communication System.

569　　　*Cognitive Science* **37**, 1356–1367.

570 ## Table captions

571 *Table 1.*   Examples of words transcribed from imitations.

572   **Figure captions**

573 *Figure 1.*   Vocal imitations collected in the transmission chain experiment. Seed

574   sounds (16) were sampled from four categories of environmental sounds:

575   glass, tear, water, zipper. Participants imitated each seed sound, and

576   then the next generation of participants imitated the imitations, and

577   so on, for up to 8 generations. Chains are unbalanced due to random

578   assignment and the exclusion of some low quality recordings.

579 *Figure 2.*   Change in perception of acoustic similarity over generations of iterated

580   imitation. Points depict mean acoustic similarity ratings for pairs of

581   imitations in each category. The predictions of the linear mixed-effects

582   model are shown with ± 1 SE.

583 *Figure 3.*   Repeated imitations retained category resemblance. A. Three types

584   of matching questions. True seed and category match questions had

585   choices from different sound categories. Specific match questions pitted

586   the actual seed against the other seeds within the same category. B.

587   Accuracy in matching vocal imitations to original seed sounds. Curves

588   show predictions of the generalized linear mixed effects models with ± 1

589   SE of the model predictions. C. Accuracy in matching transcriptions of

590   the imitations to original seed sounds (e.g., "boococucuwich" to a water

591   splashing sound). Circles show mean matching accuracy for the vocal

592   imitations that were transcribed for comparison.

593 *Figure 4.*   Repeated imitations made for better category labels. A. Mean RTs for

594   correct responses in the category learning experiment with ± 1 SE. B.

595   Cost of generalizing to new category members with ± 1 SE.