Cognition 237 (2023) 105467

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Auditory category learning is robust across training regimes

Chisom O. Obasih [a,b,c,*], Sahil Luthra [a,b,c], Frederic Dick [d,e], Lori L. Holt [a,b,c]

[a] Department of Psychology, Carnegie Mellon University, United States of America
[b] Neuroscience Institute, Carnegie Mellon University, United States of America
[c] Center for the Neural Basis of Cognition, Carnegie Mellon University, United States of America
[d] Experimental Psychology, University College London, United Kingdom
[e] Birkbeck/UCL Centre for NeuroImaging, United Kingdom

ARTICLE INFO

Keywords:
Categorization
Category learning
Auditory category learning
Generalization

ABSTRACT

Multiple lines of research have developed training approaches that foster category learning, with important translational implications for education. Increasing exemplar variability, blocking or interleaving by category-relevant dimension, and providing explicit instructions about diagnostic dimensions each have been shown to facilitate category learning and/or generalization. However, laboratory research often must distill the character of natural input regularities that define real-world categories. As a result, much of what we know about category learning has come from studies with simplifying assumptions. We challenge the implicit expectation that these studies reflect the process of category learning of real-world input by creating an auditory category learning paradigm that intentionally violates some common simplifying assumptions of category learning tasks. Across five experiments and nearly 300 adult participants, we used training regimes previously shown to facilitate category learning, but here drew from a more complex and multidimensional category space with tens of thousands of unique exemplars. Learning was equivalently robust across training regimes that changed exemplar variability, altered the blocking of category exemplars, or provided explicit instructions of the category-diagnostic dimension. Each drove essentially equivalent accuracy measures of learning generalization following 40 min of training. These findings suggest that auditory category learning across complex input is not as susceptible to training regime manipulation as previously thought.

1. Introduction

Is this mushroom edible? Is that a squeal of danger, or delight? Is that stranger trustworthy? Humans and other organisms readily learn complex constellations of cues that signal functionally equivalent sensory objects and events – like crying babies, for example. Cries of pain during a vaccination tend to be louder and longer, with more variable pitch and greater nonlinear acoustic characteristics compared to cries of bath time discomfort (Helmer et al., 2020; Koutseff et al., 2018). But adults' ability to categorize pain versus discomfort based on these complex cues *demands experience*; adults who have spent little time with infants categorize cries no better than chance. In contrast, parents and infant caregivers are significantly more accurate in categorizing cries, their accuracy scales with how much infant experience they have, and their categorization ability generalizes to unfamiliar infants' cries (Corvin, Fauchon, Peyron, Reby, & Mathevon, 2022). Experience molds caregivers' ability to use imperfect and complex sensory input regularities and guides behavior upon encountering novel input with similar properties. The latter ability – generalization – is a signature characteristic of effective category learning.

Cognitive science has long investigated the emergence of categories. One especially productive approach has been to utilize training paradigms to teach participants categories across novel or unfamiliar exemplars. In addition to advancing theoretical accounts of category learning and generalization, these literatures have informed real-world applications in second-language acquisition (Lim & Holt, 2011; Reetzke, Xie, Llanos, & Chandrasekaran, 2018), science learning (Eglington & Kang, 2017; Goldwater, Hilton, & Davis, 2022; Nosofsky, Sanders, & McDaniel, 2018), social group recognition through faces and voices (Lavan, Burton, Scott, & McGettigan, 2019; Retter, Jiang, Webster, & Rossion, 2020), stereotyping (Hugenberg & Sacco, 2008) and approaches to building effective educational materials (Carvalho & Goldstone, 2021; Nosofsky, Slaughter, & McDaniel, 2019). Many studies of category learning have examined aspects of training that best support

* Corresponding author at: Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States of America.
E-mail address: cobasih@andrew.cmu.edu (C.O. Obasih).

https://doi.org/10.1016/j.cognition.2023.105467
Received 11 October 2022; Received in revised form 17 March 2023; Accepted 21 April 2023

effective learning, informing both theory and application. We examine three such aspects in more depth in the next sections.

1.1. Training manipulations thought to support category learning

1.1.1. Exemplar variability

There is a longstanding appreciation that learners benefit from variability across the category exemplars experienced in training (W.K. Estes & Burke, 1953; Munsinger & Kessen, 1966; Posner & Keele, 1968). Typically, variability improves generalization to novel stimuli (see Raviv, Lupyan, & Green, 2022). The role of exemplar variability in category learning has been considered extensively in adult human second language speech category learning. For example, high variability phonetic training that uses speech exemplars from multiple speakers and across multiple word forms can improve non-native category learning and generalization (Logan, Lively, & Pisoni, 1991). More generally, greater acoustic variability of the exemplars can lead to learning improvements for speech category learning in speakers' first language (K. G. Estes & Lew-Williams, 2015; Galle, Apfelbaum, & McMurray, 2015; Rost & McMurray, 2009, 2010; Singh, 2008) and second language (Barcroft & Sommers, 2005; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Leong, Price, Pitchford, & Heuven, 2018; Lim & Holt, 2011; Lively, Logan, & Pisoni, 1993; Shinohara & Iverson, 2018). Past studies have employed various techniques to introduce high acoustic variability, such as using multiple talkers, a single talker with high acoustic variability, multiple prosodic voice affectations, multiple word forms, and sampling exmplars with variability across an acoustic dimension.

1.1.2. Sequence of category exemplar presentation

Learning can also be influenced by the sequence of category exemplars experienced in training. For example, interleaved exemplar presentation of the to-be-learned categories (e.g., ABCACBBAC) specifically benefits learning and/or generalization compared to blocked exemplar presentation (e.g., AAABBBCCC; Birnbaum, Kornell, Bjork, & Bjork, 2013; Bloom & Shuell, 1981; Kang & Pashler, 2012; Kornell & Bjork, 2008; McDaniel, Fadler, & Pashler, 2013; Taylor & Rohrer, 2010; Zulkiply, McLean, Burt, & Bath, 2012). However, putting this approach into practice has been complicated by interactions between the sequence of category exemplars and elements of the experimental design, including: (1) within- and between-category exemplar similarity and/or category structure (Carvalho & Goldstone, 2014a, 2014b; Kang & Pashler, 2012; Medin & Bettger, 1994; Noh, Yan, Bjork, & Maddox, 2016; Zulkiply et al., 2012; Zulkiply & Burt, 2013); (2) whether learning is active or passive (Carvalho & Goldstone, 2015); (3) the perceptual dimension across which exemplars are interleaved or blocked (Rau, Aleven, & Rummel, 2013); and (4) the type of test used to evaluate learning (Carvalho & Goldstone, 2021). Although few studies have investigated exemplar sequencing outside of visual category learning, there is some evidence that blocked presentation aids participants in learning non-native word pronunciations (Carpenter & Mueller, 2013) and phonetic categories (Fuhrmeister & Myers, 2020). Carvalho and Goldstone (2017) point out that blocking presentation appears to direct attention to within-category similarities, whereas interleaving appears to direct attention to between-category differences. Overall, studies have demonstrated that the order and grouping of exemplars experienced across training can influence category learning and generalization outcomes (see Brunmair & Richter, 2019 for meta-analysis; see Rohrer, 2012 for review).

1.1.3. Explicit instruction

The provision of explicit instructions may also promote category learning. Explicitly instructing learners to focus on a category-diagnostic dimension, or to direct attention away from a category non-diagnostic dimension, can result in enhanced non-native speech category learning (Chandrasekaran, Yi, Smayda, & Maddox, 2016). Moreover, when explicit instruction draws attention to a category-diagnostic dimension, it benefits non-native speech category learning and production above and beyond what is achieved with high-variability training alone (Wiener, Chan, & Ito, 2020). More nuanced, less explicit, manipulations that guide learners to category-diagnostic dimensions have also been effective in facilitating non-native speech category learning (Ingvalson, Holt, & McClelland, 2012; Iverson, Hazan, & Bannister, 2005; Jamieson & Morosan, 1986; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; McClelland, Fiez, & McCandliss, 2002).

1.2. Summary and aim of the study

In summary, examination of category learning across novel or unfamiliar categories has been useful in understanding how category training regimes affect learning and suggests means of improving real-world categorization. Indeed, an implicit assumption of category learning research has been that laboratory training tasks with relatively simple stimuli can inform real-world category learning. Studying visual category learning across simple dimensions, for example, may reveal processes available to early-career radiographers learning to categorize the subtle patterns that differentiate a benign from a cancerous tumor (Waite et al., 2019). Correspondingly, learning a simplified category characteristic of non-native speech might suggest scenarios that would improve classroom second language learning (Wiener, Murphy, Goel, Christel, & Holt, 2019).

However, most category learning studies differ substantially from natural category learning challenges – often by design. For example, the number of unique exemplars in lab experiments vastly undersamples natural exemplar variation. Laboratory studies tend to model real-world exemplar variability with a Gaussian distribution for simplicity. Exemplars are often defined across just two sensory dimensions, and dimensions tend to be simple, easily verbalized sensory features (e.g., line orientation, acoustic frequency). Even when categories are defined by natural visual objects or spoken utterances, exemplar sampling tends not to truly reflect the full complexity of natural categories. As a result, much of what we know about category learning has come from studies with simplifying assumptions. This entirely reasonable approach nonetheless calls into question the implicit expectation that these studies reflect the process of category learning under more complex learning challenges, such as those posed by real-world input.

Here, we put this question to the test by creating an auditory category learning challenge that intentionally violates some common simplifying assumptions. We create a novel, nonspeech acoustic stimulus space comprising >36,000 tokens across four auditory categories. The categories rely upon natural acoustic variability from spoken language (Mandarin lexical tone across multiple talkers) with underlying regularities known to be learnable because they are derived from real speech. Despite their speech origins, these sounds are not familiar, do not convey talker information, and are not heard as speech. This is because we use signal processing to eliminate voice and linguistic information, leaving only the fundamental frequency (F0) contour thought to be the most diagnostic dimension for conveying Mandarin lexical tone category to native listeners (Ho, 1976; Howie, 1976). In tonal languages like Mandarin, F0 differences like these allow a syllable like "ma" to have four different meanings according to its intonation (Chao, 1965; Gandour, 1983). As noted, we can be confident the structure of these novel categories is learnable because they are drawn from natural categories. Further, prior research examining category learning among the same pool of nonspeech hums demonstrates robust category learning among non-Mandarin listeners (Liu, 2014).

We exaggerate the learning challenge in two ways. First, each category exemplar is composed of two streams of three hums, each stream spectrally filtered such that one is situated in a high frequency band and the other in a low frequency band. These two streams are played simultaneously, but only one carries information diagnostic to category

decisions; the other is acoustically variable and non-diagnostic. This creates a rarely examined category learning challenge: Listeners must forage the acoustic soundscape to discover category-diagnostic information as it evolves (and dissipates) over time. By design, we build this qualitative category learning challenge into our stimulus set without modeling specific details of speech per se. Instead, our approach is to create a novel version of an important puzzle present in auditory category learning: Listeners must discover category-diagnostic acoustic dimensions in the context of non-diagnostic (or less-diagnostic) acoustic variability arising from other dimensions of the same sound source (e.g., across different bands of formant frequencies) or even across simultaneous competing sounds.

Second, the hum stream in one frequency band is a concatenation of three unique hums drawn from a single Mandarin tone category. The other is a concatenation of three unique hums, each drawn from different Mandarin tone categories. In this way, one frequency band contains tone-category-diagnostic information, and the other frequency band is category uninformative. Thus, category learning requires both discovering (at least implicitly) the *category-diagnostic frequency band* that contains a statistically regular pattern derived from a single Mandarin tone category, and also recognizing the category-diagnostic, but acoustically variable, *pattern* within this band (see Fig. 1 for a schematic depiction of the stimuli). In summary, this creates a complex high-dimensional exemplar space across which four categories are defined over multiple difficult-to-verbalize dimensions and sampling distributions.

We intentionally chose a learning challenge that would not approach ceiling in a single session so that we could better capture differences that might be apparent across training regimes; ceiling performance would make this problematic. Although this approach does not measure what learners can achieve with longer training, examining category learning across a single session has been the workhorse paradigm across both the visual and auditory category learning literatures because it tracks early, online category acquisition. Further, by limiting training to a single session, we can examine effects of online category learning without influences of offline learning or consolidation (which might be productively examined in future work).

1.3. Experiments overview

Here, we first examine whether young adult participants recruited from a diverse online sample can accomplish this complex category learning challenge in a single training session that involves overt category decisions and explicit feedback. We then examine how learning is influenced by variability in three aspects of the training regime, each of which has been shown to affect learning in simpler categorization challenges: manipulations of exemplar variability, category exemplar sequencing, and explicit instructions.

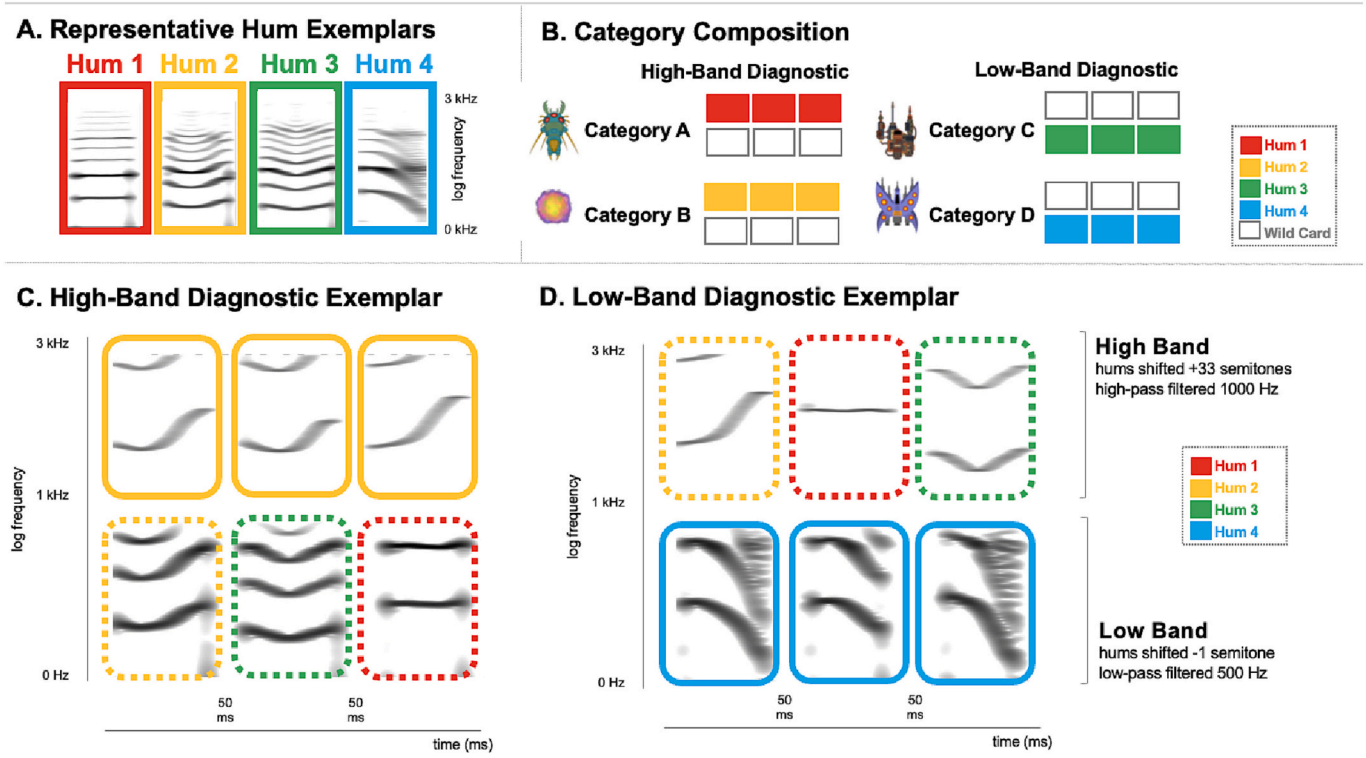

Fig. 1. Schematic of Sound Exemplars. A. Non-speech hums derived from natural utterances of four native Mandarin (2 female) speakers producing utterances varying in lexical tone, which is conveyed by fundamental frequency (F0) contours. Hums preserve only the F0 contour and do not sound like speech, yet they possess natural acoustic regularity within hum categories and distinct patterns across hum categories. Here and in subsequent panels, color conveys the hum category. B. Hums were filtered into high ($\geq$ 1000 Hz) and low ($\leq$500 Hz) frequency bands and three hums were concatenated in each band to compose a sound exemplar. For each, a diagnostic band (colored boxes) possessed within-hum-category exemplars and a non-diagnostic band had 3 hums, each drawn from a different one of the four hum categories (open "wild card" boxes). Exemplars defining the four categories were created such that listeners needed to discover the diagnostic band in the context of the simultaneous non-diagnostic band and learn the hum pattern across acoustic variability within the diagnostic band. The four aliens used to guide categorization responses are shown, as well. C. A spectrogram showing a representative exemplar drawn from Category A, for which the high-frequency band was diagnostic. Here, and in Panel D, colored rectangles indicate the lexical tone category from which the hum was created. Solid colored lines indicate the category-diagnostic frequency band; dashed lines show the category uninformative frequency band. D. Spectrogram showing a representative exemplar drawn from Category D, for which the low-frequency band was diagnostic.

2. Materials and methods

2.1. Participants

Since this was a novel categorization challenge, we conducted several pilot studies from which to estimate power. These studies revealed robust learning across ~30 participants. Here, we doubled the sample, targeting recruitment of 60 participants per experiment to improve our ability to detect subtle learning differences across learning contexts.

In total, 300 young adults aged 18–35 years participated online for monetary compensation via recruitment through Prolific.co. There were no restrictions on language background, and all participants self-reported normal hearing. Table 1 shares participant demographics. Given our relatively unrestricted recruitment of participants online, our sample is likely more representative of the general population than that of studies that recruit from a university student population (Henrich, Heine, & Norenzayan, 2010). Four participants were excluded due to an experimental error that duplicated trials, leaving 296 participants in the final analyses and a minimum of 58 participants per experiment. All participants provided informed consent approved by the Carnegie Mellon University Institute Review Board (IRB).

2.2. Stimuli

Fig. 1 illustrates the construction of sound exemplars. Stimuli for all experiments were drawn from the same acoustic space. The building blocks for these stimuli were nonspeech hums created by extracting the fundamental frequency (F0) contour from natural speech recordings of single-syllable words, each recorded by four native Mandarin speakers (2 male, 2 female; Liu, 2014). A screen displayed both the Mandarin Chinese character and the pinyin spelling of the word frame (with tone number 1, 2, 3, 4) to prompt native speakers to utter each word twice, with self-paced progress as utterances were digitally recorded with Praat (Boersma, 2001). Each speaker produced 20 unique word-frames (pinyin spellings: *can, chou, di, fa, ge, guo, huan, jie, kui, peng, pu, qian, shi, tuo, xi, xiang, xing, xue, yang, yu*) in each of the four lexical tones for a total of 80 utterances per talker. A native Mandarin listener checked stimuli for clarity and representativeness of the lexical tone contour.

Table 1
Participant demographics.

Experiment	N	Age Range (in years)	Mean [SE] Age (in years)	Gender	Race	# of Native Languages Represented[1]
1	59	18–33	23.1 [0.51]	68% female 29% male 3% non-binary	68% white	9
2	59	18–35	24.9 [0.61]	31% female 69% male	76% white	22
3	60	18–35	23.9 [0.60]	45% female 53% male 2% non-binary	72% white	18
4	58	18–33	24.2 [0.52]	66% female 24% male 9% non-binary 2% no response	55% white	16
5	60	18–34	24.8 [0.51]	43% female 52% male 3% non-binary 2% no response	85% white	13

[1] Based on self-reported languages when asked to "List language(s) spoken before age 2."

These speech recordings were processed in the open-source speech analysis software Praat (Boersma, 2001) to create non-speech hums by extracting the pitch contour using the *Analyse periodicity: To Pitch* function and converted into hums using the *To Sound (hum)* function. Expert listeners removed some stimuli from the pool based on poor pitch tracking and discontinuous hum outcomes (Liu, 2014).

To make a single stimulus exemplar, three unique non-speech hums drawn from the same Mandarin talker were assigned to a higher frequency band and three to a lower frequency band. As illustrated in Fig. 1B, one of the frequency bands was designated the diagnostic band; it possessed 3 unique hums drawn from a *single* lexical tone category. The other band possessed 3 unique hums from *any* lexical tone category ("wild card").

Next, the hums were processed using the audio processing software Sound eXchange (sox.sourceforge.net), with additional processing in Adobe Audition (version 13.0.7). First, hums were padded with 50 ms of silence at the beginning and end of the sound clip and high-pass-filtered at 30 Hz to remove slow drift and reduced in gain by 10 dB. Second, high- and low-frequency-band versions of these stimuli were created. To create the high-frequency-band components, hums were pitch-shifted +33 semitones in Audition and then high-pass filtered using sinc Kaiser-windowed filter in Sox to preserve all frequencies at and above 1000 Hz. To create the low-frequency-band components, the same hums were pitch-shifted by −1 semitone and low-pass filtered to preserve all frequencies at and below 500 Hz. In the process of pitch shifting, hums were simultaneously normalized to be 400 ms using the iZotope algorithm in Audition, using the high precision mode with pitch coherence set to 4. The 400-ms, pitch-shifted and high/low-pass filtered hums were RMS-matched in amplitude and normalized to be −6 dB below the maximum digital range.

As shown in Fig. 1B, the *category-diagnostic band* was created by drawing from the pool of hums derived from a single talker, choosing a frequency band (high or low), randomly selecting three hums from a single hum (lexical tone) category, and concatenating the hums with 100 ms of total silence between each token. We created all permutations in both high and low frequency bands.

Similarly, the *category-uninformative "distractor" band* was created by drawing from a pool of hums from the same talker used to create the diagnostic-band hum sequence, with hums placed into the frequency band opposite the diagnostic band. For the non-diagnostic band, hums were randomly selected from three different hum categories (selected from any of the four hum categories) and concatenated with 100 ms silences between each hum. This was repeated for all permutations. The diagnostic band and uninformative distractor band were then added together such that the onset of each of the three hums of each frequency band was temporally aligned, and stimuli evolved across 1400 ms in all.

For counterbalancing purposes, there were two sets of four categories. Fig. 1B illustrates Set 1, in which Category A and Category B are defined by high-frequency diagnostic bands whereas Category C and D are defined by low-frequency diagnostic bands. This relationship was reversed in Set 2 (e.g., low-frequency diagnostic for Categories A and B, not shown in Fig. 1). Assignment of set was counterbalanced across participants in each experiment and analyses collapse across set assignment.

Overall, the full constellation of hum permutations resulted in a stimulus pool with over 36,000 exemplars. From this exemplar space we randomly selected 2048 total exemplars (256/category/set) for the

present experiments. Half of the exemplars for each condition (128/category/set) were reserved as the training stimulus pool whereas the other half was reserved as a pool to test generalization. The 2048 stimuli selected for the present experiments are available on OSF.io.

2.3. General procedure

Five experiments shared common procedures, differing only in their approach to training. In all experiments, training blocks alternated with generalization blocks (see Fig. 2C). Only the nature of the training blocks varied across experiments. Generalization blocks were identical across experiments to facilitate cross-experiment outcome comparisons. All experiments involved training over 40 min.

Moreover, across all experiments, training involved overt category decisions and explicit feedback (see Fig. 2 for schematics). Following a 500-ms fixation, participants heard a category exemplar and matched it to one of four novel 'alien' illustrations via keyboard response at sound offset, with immediate feedback lasting 1500 ms; the next trial commenced immediately. Across experiments, each auditory category consistently mapped to a specific alien presented on the screen. In Experiments 1 and 2 all four alien creatures were visible on the screen (4-alternative force choice (4AFC)), whereas in Experiments 3–5, only pairs of alien creatures were visible (2AFC), with the other two aliens greyed out and unavailable for response.

Each of the four training blocks consisted of 120 trials (30 trials/category), totaling 480 training trials. At the commencement of each training block, 30 exemplars/category were randomly selected without replacement from the pool of 128 category exemplars. Thus, exemplars were never repeated within a single training block, and there was a low probability of any single exemplar repeating across training blocks. Each training block was divided into either three mini-blocks of 40 trials each (Experiments 1 and 2, for 4AFC training) or six mini-blocks of 20 trials each (Experiments 3, 4, and 5, for 2AFC training), to allow for brief self-timed breaks between mini-blocks. Except for Experiment 5 (see Section 7), participants were not informed of the dual-band nature of the stimuli and were simply instructed to use the feedback during training trials to learn which sounds corresponded with which alien.

Generalization was similar to training, but participants did not receive feedback. Generalization trials for all experiments were 4AFC. Each of the four generalization blocks consisted of 20 novel exemplars/category (80 total stimuli) not encountered during training. These 80 exemplars were randomly selected without replacement from the stimulus pool reserved for novel generalization prior to the experiment, and the generalization set was used consistently for each participant across each experiment. This presented the opportunity to examine cross-experiment effects of training manipulations via participants' ability to generalize category learning to novel exemplars.

Participants completed the experiment online via Gorilla, an online experiment creation and hosting website (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) on a laptop or desktop computer using the Google Chrome browser. Prior to beginning the category learning task, participants underwent a system check to ensure the auto play of sound at a comfortable listening level and a short task to ensure compliance with the use of binaural headphones (Milne et al., 2021). All sounds were presented in the lossless *.FLAC format. After the experiment, participants shared language and music training history, were invited to share notes detailing their task strategies, and received an experiment debriefing.

2.4. Approach to analyses

For each experiment, we analyzed training and generalization blocks separately, asking whether significant learning and generalization occurred with a specific training regime. For training and generalization blocks, we analyzed: (1) the overall change in performance across Blocks 1–4 using a repeated measures ANOVA and post-hoc comparison of Block 1 and Block 4; and (2) indices of early learning by examining Block 1 accuracy compared to chance. We compared training and generalization performance between select pairs of experiments using mixed model ANOVA. (Linear mixed effects modeling yielded the same results and are available on OSF.io.)

To ask whether training regime differentially affected generalization overall, a set of cross-experiment analyses (reported after Experiment 5) compared generalization progress from Block 1 to 4 as well as final generalization achievement in Block 4. We supplemented these analyses with Bayesian Equivalence Independent *t*-tests across all pairs of experiments, looking both at generalization progress and final generalization achievement.

3. Experiment 1: 4AFC training with full exemplar variability

Experiment 1 tested listeners' ability to learn the complex auditory

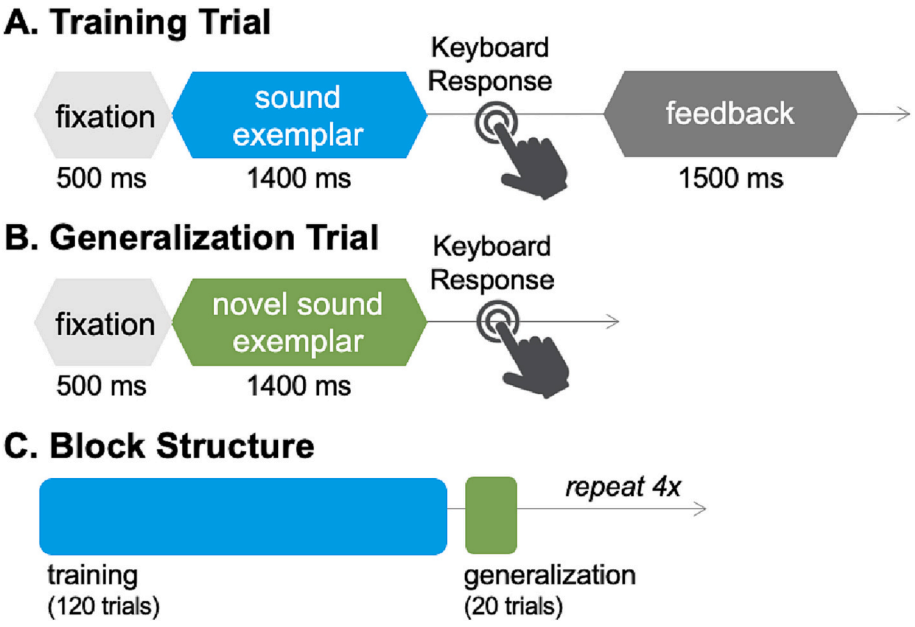

Fig. 2. Trial and Block Structure Across Experiments. A. Training trials with overt categorization decisions and immediate feedback. B. Generalization trials with novel sound exemplars not encountered in training, with no feedback C. Training regimes (defined by the nature of training trials) differed across experiments, but all experiments were comprised of four cycles of 120 training trials (A) followed by 20 generalization trials (B). Note that generalization trials were identical across experiments.

categories under conditions of full acoustic variability in a four-alternative forced-choice categorization task, with feedback (Table 2).

3.1. Methods specific to experiment 1

Here, 480 exemplars (120/category) were randomly selected from the full pool of 512 training stimuli (128/category). On each trial, participants chose which of four aliens (4AFC) corresponded to the sound they had heard; as with all experiments, they received feedback after each training trial. Participant characteristics are shown in Table 1; data are shown in Fig. 3.

3.2. Results

3.2.1. Training accuracy

A Greenhouse-Geisser (GG)-corrected repeated-measures ANOVA showed mean accuracy changed over Training Block ($F(2.03, 117.74) = 29.3$, $p = 3.7e\text{-}11$, $\eta^2_G = 0.096$). Accuracy was above chance even in the first block ($M = 0.343$, $t(58) = 8.510$, $p = 8.62e\text{-}12$, Cohen's $d = 1.108$), and accuracy significantly improved from Block 1 to Block 4 ($M_{Block4\text{-}Block1} = 0.107$, $t(58) = 6.206$, $p = 6.22e\text{-}08$, Cohen's $d = 0.8080$).

3.2.2. Generalization accuracy

Generalization of category learning to novel exemplars was evident even in Block 1 ($M = 0.374$, $t(58) = 6.895$, $p = 4.39e\text{-}09$, Cohen's $d = 0.8977$), changed significantly across blocks ($F(3, 174) = 9.295$, $p = 9.83e\text{-}06$, $\eta^2_G = 0.058$), and improved significantly from Block 1 to 4 ($M_{Block4\text{-}Block1} = 0.107$, $t(58) = 4.437$, $p = 4.14e\text{-}05$, Cohen's $d = 0.5776$).

4. Experiment 2: 4AFC training with low exemplar variability

As noted above, high exemplar variability may lead to slower and initially less accurate performance in training. However, it can yield dividends in supporting better generalization (Logan et al., 1991; Lively et al., 1993; see Raviv et al., 2022 for review). Conversely, small numbers of training exemplars may lead to faster and more accurate learning, but poorer generalization. We test this hypothesis in Experiment 2 with a limited set of training exemplars, but with the same set of novel generalization exemplars as in Experiment 1.

4.1. Methods specific to experiment 2

Here, training involved only 40 exemplars (10 exemplars/category) randomly selected from the training pool of 512 training exemplars prior to experimentation and consistent among participants. Each exemplar was encountered 12 times across training to arrive at the same number of 480 training trials as Experiment 1. Participant demographics are in Table 1. Fig. 4 shows training and generalization data.

Table 2
Training Protocols.

Experiment	Response Type	# Unique Exemplars	Feedback	Training Type
1	4AFC	480	Yes	Full exemplar variability and all 4 category response options/trial
2	4AFC	40	Yes	Restricted exemplar variability and all 4 category response options/trial
3	2AFC	480	Yes	2 category response options/trial, always grouped by high/low informative band
4	2AFC	480	Yes	2 category response options/trial, all possible pair-wise combinations
5	2AFC	480	Yes	As in Exp 3, but with explicit instructions to direct attention to the diagnostic band

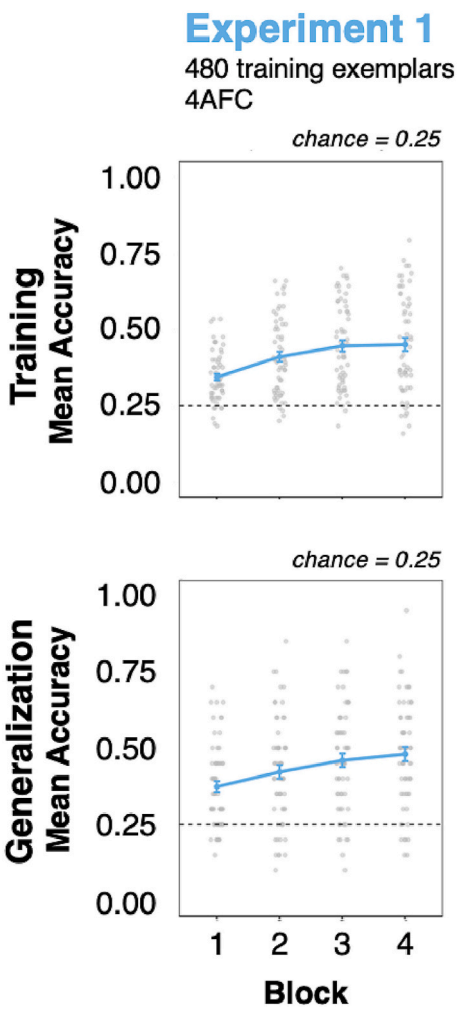

Fig. 3. Experiment 1, 4AFC Full Exemplar Variability: Training and Generalization Accuracy by Block. The top panel represents training accuracy. The bottom panel shows generalization accuracy. Dashed lines represent chance and error bars reflect standard error of the mean. Each individual gray point represents an individual participant's mean accuracy and larger, colored symbols show mean across-participant accuracy.

4.2. Results

4.2.1. Training accuracy

A Greenhouse-Geisser (GG)-corrected repeated-measures ANOVA showed accuracy changed with Training Block ($F(1.87, 108.23) = 31.257$, $p = 5.8e\text{-}11$, $\eta^2_G = 0.089$). Block 1 accuracy was above chance ($M = 0.34$, $t(58) = 8.245$, $p = 2.39e\text{-}11$, Cohen's $d = 1.073$), and accuracy significantly improved from Block 1 to Block 4 ($M_{Block4\text{-}Block1} = 0.114$, $t(58) = 6.768$, $p = 7.17e\text{-}09$, Cohen's $d = 0.8812$).

4.2.2. Generalization accuracy

Generalization of category learning to novel exemplars was evident even in Block 1 ($M = 0.381$, $t(58) = 5.970$, $p = 1.52e\text{-}07$, Cohen's $d = 0.7773$), changed across blocks ($F(3, 174) = 5.9$, $p = 0.00074$, $\eta^2_G = 0.03$), and improved from Block 1 to 4 ($M_{Block4\text{-}Block1} = 0.087$, $t(58) = 3.766$, $p = 0.000389$, Cohen's $d = 0.4903$).

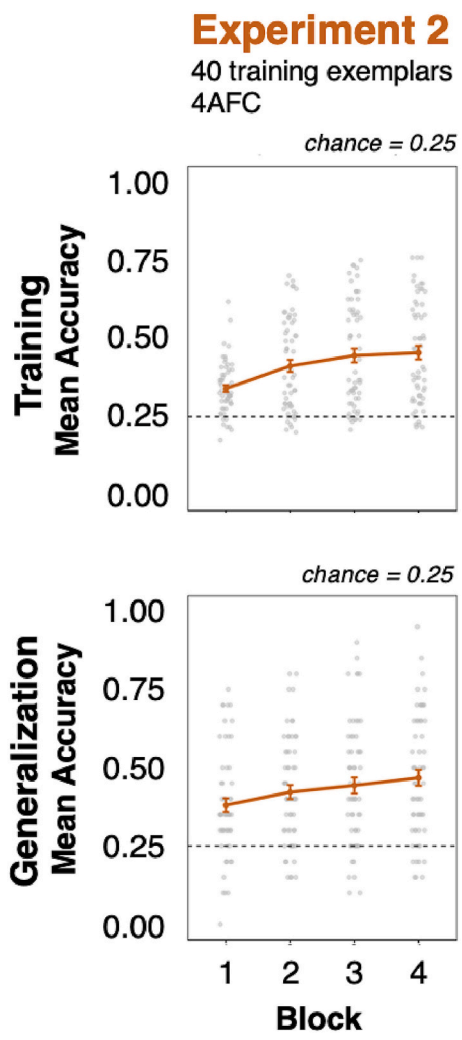

Fig. 4. Experiment 2, 4AFC Low Exemplar Variability: Training and Generalization Accuracy by Block. The top panel represents training accuracy. The bottom panel shows generalization accuracy. Dashed lines represent chance and error bars reflect standard error of the mean. Each individual gray point represents an individual participant's mean accuracy and larger, colored symbols show mean across-participant accuracy.

4.2.3. Comparison of experiments 1 and 2

A mixed-model ANOVA across Training Regime (Experiment) and Training Block showed no significant effect of exemplar variability in training ($F(1, 116) = 0.000338$, $p = 0.985$, $\eta^2_G = 2.35\text{e-}6$) and no interaction (GG-corrected $F(2.01, 233.03) = 0.067$, $p = 0.936$, $\eta^2_G = 1.12\text{e-}4$). Likewise, neither improvements across Block 1 to 4 ($t(115.9814) = -0.3435$, $p = 0.732$, Cohen's d = −0.06324) nor final Block 4 achievement in training ($t(115.9996) = -0.1391$, Bonferroni-adjusted p = 1, Cohen's d = −0.0256) differed as a function of exemplar variability. In all, exemplar variability did not produce differential training outcomes.

In a similar manner, there was no influence of exemplar variability on generalization accuracy ($F(1, 116) = 0.048$, $p = 0.828$, $\eta^2_G = 0.000271$), nor an interaction of exemplar variability with generalization block ($F(3, 348) = 0.258$, $p = 0.855$, $\eta^2_G = 0.000753$). Changes in generalization accuracy from Block 1 to 4 did not differ with exemplar variability experienced in training ($t(115.8347) = 0.5834$, $p = 0.561$, Cohen's d = 0.1074) nor did generalization achievement in Block 4 ($t(114.3794) = 0.3689$, Bonferroni-adjusted p = 1, Cohen's d = 0.06792).

5. Experiment 3: 2AFC training with pairs grouped by category-diagnostic band

Recall that the auditory category exemplars confront participants with two learning challenges: (1) to identify the diagnostic frequency band in the context of a simultaneous, non-diagnostic band and (2) to learn the pattern of hums present in the diagnostic band despite their within-category acoustic variability. In Experiment 3, we block categorization decisions according to the category-diagnostic band, thereby potentially (implicitly) encouraging selective attention to the category-relevant frequency band within blocks of trials (Carvalho & Goldstone, 2017).

5.1. Methods specific to experiment 3

Here, training trials were blocked as 2AFC category decisions. Like Experiment 1, participants completed 480 training trials with feedback, where the 480 trials (120/category) were randomly selected from the full pool of 512 training exemplars (128/category). This was accomplished by dividing each training block (120 trials) into six 20-trial mini-blocks. Half of the mini-blocks were grouped by high-frequency diagnostic band and half by low-frequency diagnostic band. For example, as shown in Fig. 1B, Category A and B stimuli were presented in one half of the mini-blocks, and Category C and D were presented in the other half. Mini-blocks alternated between category pairs differentiated in either the high- and low-frequency diagnostic band, with order counterbalanced across participants. Generalization blocks mirrored Experiments 1 and 2. Participant demographics are shown in Table 1. Data are plotted in Fig. 5.

5.2. Results

5.2.1. Training accuracy

Accuracy changed across Training Block (GG-corrected $F(2.05, 120.68) = 18.488$, $p = 7.75\text{e-}08$, $\eta^2_G = 0.065$), with improvement from Block 1 to Block 4 ($M_{Block4\text{-}Block1} = 0.087$, $t(59) = 5.380$, $p = 1.34\text{e-}06$, Cohen's d = 0.6946) and as previously, above-chance accuracy in Block 1 (M = 0.588, $t(59) = 9.712$, $p = 7.59\text{e-}14$, Cohen's d = 1.254; chance = 0.50).

5.2.2. Generalization accuracy

Generalization accuracy changed with Training Block ($F(3, 177) = 12.934$, $p = 1.12\text{e-}07$, $\eta^2_G = 0.061$) with significant improvement in generalization from Block 1 to Block 4 ($M_{Block4\text{-}Block1} = 0.137$, $t(59) = 5.438$, $p = 1.08\text{e-}06$, Cohen's d = 0.7021), and above-chance generalization accuracy in Block 1 (M = 0.355, $t(59) = 4.534$, $p = 2.88\text{e-}05$, Cohen's d = 0.5853; chance = 0.25).

6. Experiment 4: 2AFC training with all category pairs

As a counterpart to Experiment 3, Experiment 4 examines whether category learning with 2AFC training is successful without category-diagnostic blocking. Here, all six possible pairs of categories were presented in separate training blocks (e.g., AB/AC/AD/BC/BD/CD). We hypothesized that without implicit direction to the diagnostic band, participants would be forced to discover the two learning challenges simultaneously and that this would, akin to interleaved presentation, exaggerate between-category differences (Carvalho & Goldstone, 2017). After Experiment 4 findings are reported, results from Experiments 3 and 4 are directly compared.

6.1. Methods specific to experiment 4

Experiment 4 used full exemplar variability (like Experiments 1 and 3) and presented 2AFC training across six 20-trial mini-blocks per training block (like Experiment 3). The order of category pair mini-

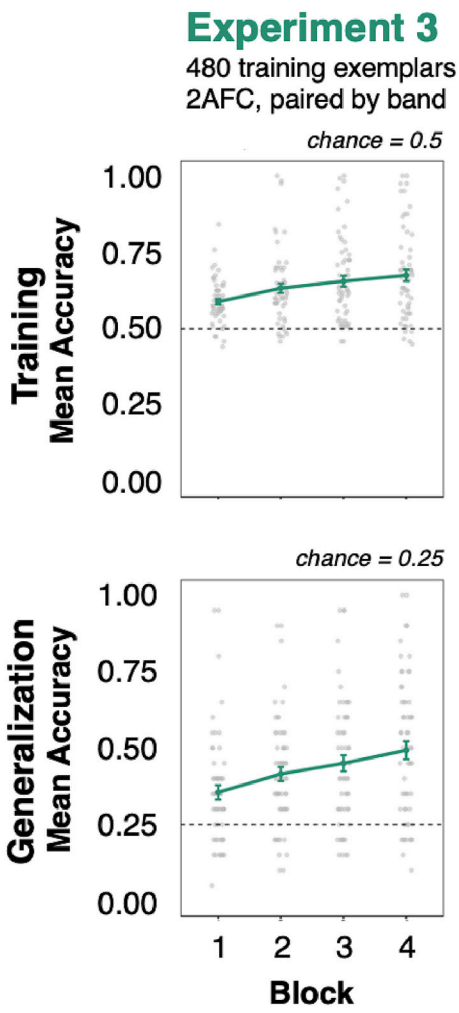

Fig. 5. Experiment 3, 2AFC Pairs Grouped by Category-Diagnostic Band: Training and Generalization Accuracy by Block. The top panel represents training (2AFC) accuracy. The bottom panel shows generalization (4AFC) accuracy. Dashed lines represent chance and error bars reflect standard error of the mean. Each individual gray point represents an individual participant's mean accuracy, and larger, colored symbols show mean across-participant accuracy.

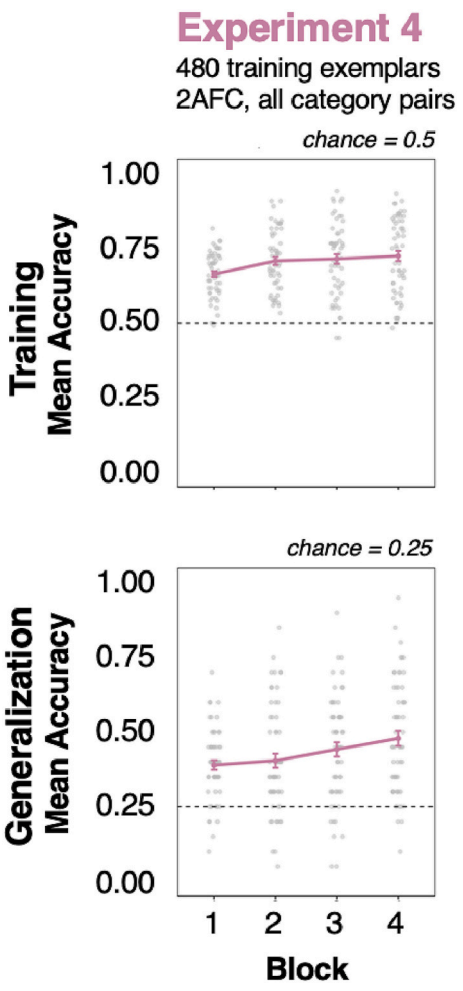

Fig. 6. Experiment 4, 2AFC All Category Pairs: Training and Generalization Accuracy by Block. Top panel shows training accuracy, and bottom panel shows generalization accuracy. Dashed lines represent chance and error bars reflect standard error of the mean. Each individual gray point represents an individual participant's mean accuracy and larger, colored symbols show mean across-participant accuracy.

blocks was randomized for each training block, for each participant. Generalization blocks mirrored previous experiments. Table 1 provides demographic information, and data are plotted in Fig. 6.

6.2. Results

6.2.1. Training accuracy

Accuracy changed with Block (GG-corrected $F(2.37, 134.96) = 9.673$, $p = 4.27e\text{-}05$, $\eta^2_G = 0.044$), with improvement from Block 1 to Block 4 ($M_{Block4\text{-}Block1} = 0.061$, $t(57) = 4.118$, $p = 0.000125$, Cohen's d = 0.5407) and above-chance accuracy in Block 1 (M = 0.664, $t(57) = 16.46$, $p = 2.51e\text{-}23$, Cohen's d = 2.162; chance = 0.50).

6.2.2. Generalization accuracy

Accuracy changed across Block ($F(3, 171) = 6.654$, $p = 0.000282$, $\eta^2_G = 0.041$), with significant improvements from Block 1 to 4 ($M_{Block4\text{-}Block1} = 0.089$, $t(57) = 3.973$, $p = 0.000202$, Cohen's d = 0.5216) and above-chance generalization in Block 1 (M = 0.39, $t(57) = 8.929$, $p = 2.02e\text{-}12$, Cohen's d = 1.172).

6.2.3. Comparison of experiments 3 and 4

We asked how training that paired categories according to diagnostic band (Experiment 3) compared to pairing categories randomly regardless of diagnostic band (Experiment 4). A mixed-model ANOVA revealed that there was a significant effect of Training Regime ($F(1, 116) = 12.130$, $p = 0.000701$, $\eta^2_G = 0.073$) but no interaction between Block and Experiment (GG-corrected $F(2.22, 257.56) = 1.07$, $p = 0.35$, $\eta^2_G = 0.002$).

Random pairing of categories without regard to the diagnostic band in Experiment 4 resulted in significantly better Block 1 training accuracy ($t(114.7529) = -5.5759$, Bonferroni-adjusted $p = 6.64e\text{-}7$, Cohen's d = −1.027) compared to Experiment 3. However, there was no significant difference in final training achievement in Block 4 ($t(114.6223) = -1.9417$, Bonferroni-adjusted $p = 0.2184$, Cohen's d = −0.3571) nor a difference in overall training improvement from Block 1 to 4 ($t(115.4895) = 1.1408$, $p = 0.256$, Cohen's d = 0.2099).

There was no influence of category pairing on generalization accuracy ($F(1, 116) = 0.000729$, $p = 0.979$, $\eta^2_G = 4.23e\text{-}06$) nor an interaction with block ($F(3, 348) = 1.019$, $p = 0.384$, $\eta^2_G = 0.003$). Likewise, there were no significant differences in generalization progress from Block 1 to 4 ($t(114.942) = 1.4117$, $p = 0.161$, Cohen's d = 0.2597) or final generalization achievement in Block 4 ($t(113.1273) = 0.3403$, Bonferroni-adjusted $p = 1$, Cohen's d = 0.06255).

7. Experiment 5: 2AFC training with pairs grouped by category-diagnostic band and explicit instructions

Experiment 3 blocked categories according to their diagnostic frequency band in a manner that might implicitly guide discovery of category-relevant dimensions. Experiment 5 takes a more explicit approach, asking whether category learning is facilitated by providing instructions about the category-relevant frequency band.

7.1. Methods specific to experiment 5

Experiment 5 used full exemplar variability (like Experiment 1) and a 2AFC training task with trials blocked according to a shared diagnostic band (like Experiment 3). In addition, participants were informed that "previous participants […] found it beneficial to listen to the higher [or lower] pitched sounds when learning which sounds go with which alien." Before each mini-block of 20 trials, participants were presented with a blank screen with the text "Listen high!" or "Listen low!" in accordance with the diagnostic frequency band of the category pairs in the mini-block. Otherwise, the procedure followed that of Experiment 3. Table 1 shows participant demographics. Data are plotted in Fig. 7.

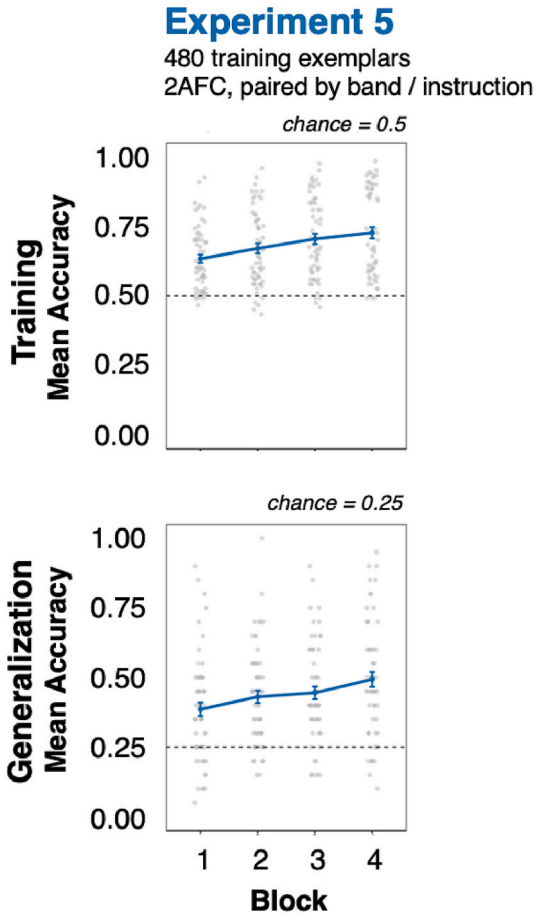

Fig. 7. Experiment 5 2AFC, Pairs Grouped by Category-Diagnostic Band + Explicit Instructions: Training and Generalization Accuracy by Block. The top panel represents training accuracy. The bottom panel shows generalization accuracy. Dashed lines represent chance and error bars reflect standard error of the mean. Each individual gray point represents an individual participant's mean accuracy and larger, colored symbols show mean across-participant accuracy.

7.2. Results

7.2.1. Training accuracy

Accuracy changed with Training Block (GG-corrected $F(2.48, 146.23) = 35.011$, $p = 3.93e\text{-}15$, $\eta^2_G = 0.063$). Performance was above chance in Block 1 ($M = 0.633$, $t(59) = 8.977$, $p = 1.25e\text{-}12$, Cohen's $d = 1.159$) and improved from Block 1 to 4 ($M_{Block4\text{-}Block1} = 0.093$, $t(59) = 7.736$, $p = 1.53e\text{-}10$, Cohen's $d = 0.9987$).

7.2.2. Generalization accuracy

Generalization accuracy changed over block (GG-corrected $F(2.55, 150.28) = 9.629$, $p = 2.52e\text{-}05$, $\eta^2_G = 0.041$). Block 1 generalization accuracy was above chance ($M = 0.386$, $t(59) = 5.570$, $p = 6.61e\text{-}07$, Cohen's $d = 0.7190$) and improved significantly from Block 1 to 4 ($M_{Block4\text{-}Block1} = 0.107$, $t(59) = 4.284$, $p = 6.87e\text{-}05$, Cohen's $d = 0.5531$).

7.2.3. Comparison of experiments 3 and 5

There was a significant influence of the presence of explicit instructions on training accuracy ($F(1, 118) = 4.311$, $p = 0.04$, $\eta^2_G = 0.03$) but no interaction between Block and Experiment (GG-corrected $F(2.28, 268.87) = 0.304$, $p = 0.766$, $\eta^2_G = 0.000416$). There was a significant difference in training accuracy in Block 1 with an advantage for learning with explicit instructions ($t(98.1304) = 2.551$, Bonferroni-adjusted $p = 0.0492$, Cohen's $d = 0.4657$), but no difference in learning progress from Block 1 to 4, $t(109.5165) = 0.3311$, $p = 0.741$, Cohen's $d = 0.0604$) or in the final training accuracy achievement ($t(117.7899) = 1.8456$, Bonferroni-adjusted $p = 0.27$, Cohen's $d = 0.3370$).

There was no influence of explicit instructions on generalization ($F(1, 118) = 0.126$, $p = 0.723$, $\eta^2_G = 0.000769$) and there was no interaction with Block (GG-corrected $F(2.79, 329.31) = 0.556$, $p = 0.632$, $\eta^2_G = 0.001$). There was no significant difference in generalization progress from Block 1 to 4 ($t(117.9932) = 0.8422$, $p = 0.401$, Cohen's $d = 0.1538$) or final generalization achievement in Block 4 ($t(116.7032) = 0.0208$, Bonferroni-adjusted $p = 1$, Cohen's $d = 0.003804$) as a function of providing explicit instruction.

8. Comparing generalization across training regimes

As described above, each experiment involved generalization testing blocks comprised of the same 80 exemplars, not heard during training. This allows for direct comparison of the influence of different training regimes on generalization of category learning. To this end, we conducted a two-way mixed model ANOVA of generalization accuracy across Block versus all five Training Regimes (Experiments). The significant effect of Block (GG-corrected $F(2.91, 845.94) = 42.678$, $p = 5.05e\text{-}25$, $\eta^2_G = 0.044$) is indicative of improvement of generalization with training, consistent with the results from individual experiments. Crucially, there was no overall difference in generalization accuracy across Training Regimes ($F(4, 291) = 0.058$, $p = 0.994$, $\eta^2_G = 0.000542$) and no interaction (GG-corrected $F(11.63, 845.94) = 0.513$, $p = 0.903$, $\eta^2_G = 0.002$). Neither generalization progress (examined with a one-way ANOVA comparing the Block 4 – Block 1 difference in generalization accuracy; $F(4, 291) = 0.698$, $p = 0.594$, $\eta^2_G = 0.009$; Fig. 8A) nor generalization achievement (examined with a one-way ANOVA comparing Block 4 accuracy; $F(4, 291) = 0.164$, $p = 1$, $\eta^2_G = 0.002$; Fig. 8B) differed across training regimes.

Given the similarity among experimental outcomes, we also conducted Bayesian equivalence testing to examine the strength of the evidence that training regime manipulations have essentially equivalent effects. We again focus on generalization progress along with final generalization achievement, setting the equivalence region from −0.05 to 0.05 in Cohen's d units using Bayesian Independent Samples Equivalence *t*-test (JASP Team, 2022).

Fig. 9 shows Bayes Factor (BF) comparing the equivalence hypothesis (i.e., that the effect falls within our equivalence interval) versus the

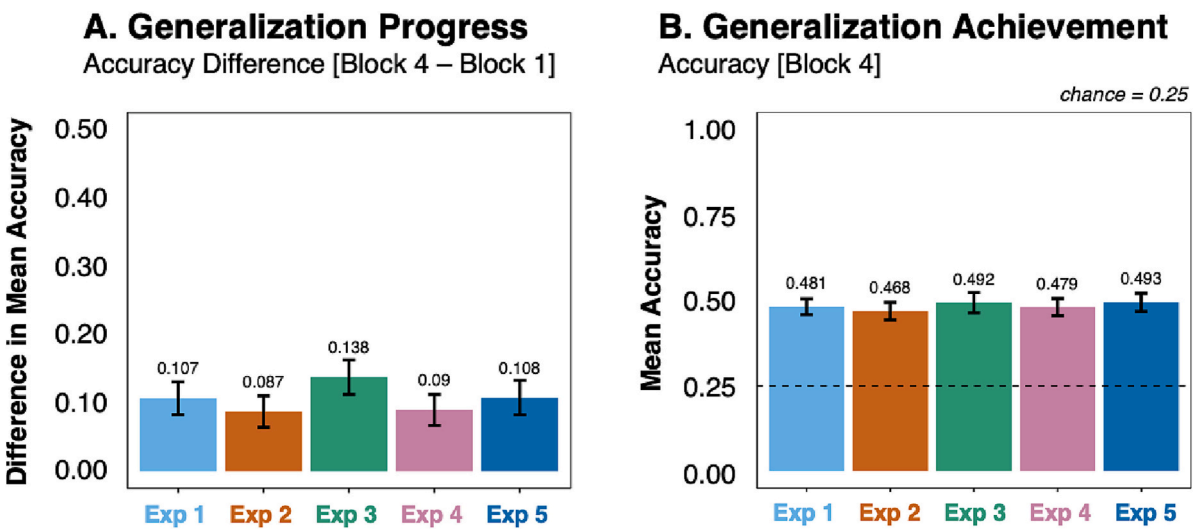

Fig. 8. Generalization Progress and Achievement Across Training Regimes. Generalization of category learning was very robust. Training regime manipulations across experiments did not influence generalization progress from Block 1 to Block 4 (panel A), nor did they influence ultimate generalization achievement in Block 4 (panel B). Error bars indicate standard error.

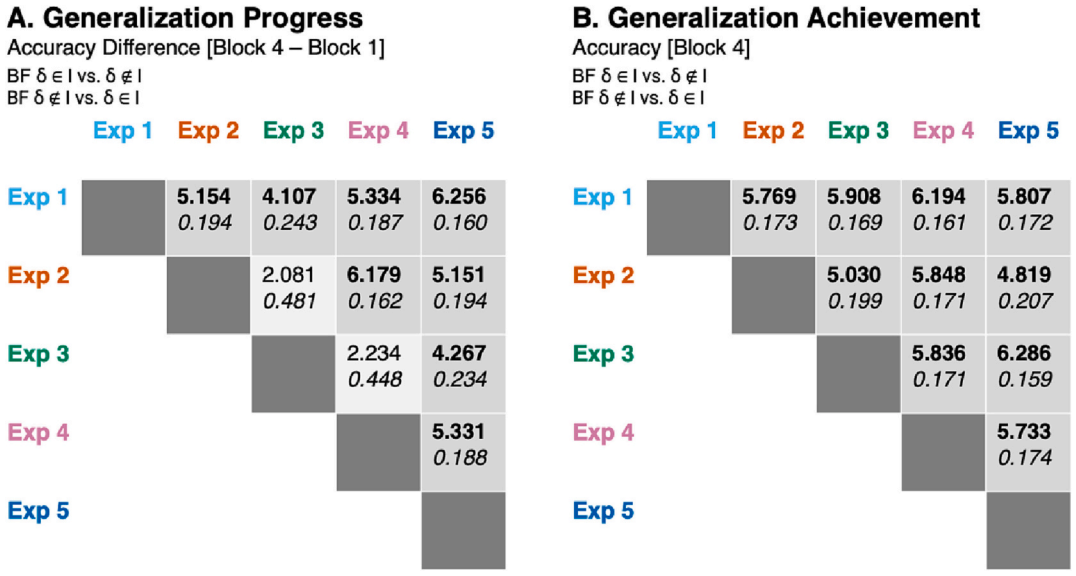

Fig. 9. Generalization Across Training Regimes, Bayesian Equivalence Testing. Each panel shows comparison of two Bayes Factors (BF) across experiments: the top number indicates the evidence that the difference lies within the equivalence region, and the bottom number indicates the evidence that the difference lies outside the equivalence region. A. BF results from Generalization Progress (Block 4 – Block 1 accuracy). B. BF results from Generalization Achievement (Block 4). For ease of interpretation, comparisons where BF > 4 are in bold font, and BF < 1 in italics.

hypothesis that the effect lies outside this interval. For each pairwise comparison, the evidence is stronger for equivalence. Using criteria suggested by Andraszewicz et al. (2015), there is moderate evidence that generalization progress and ultimate achievement are not differentially influenced by the training regimes that manipulate exemplar variability, exemplar sequencing, or explicit instruction.

Finally, we examined the potential influence of four (Exp 1–2) versus two (Exp 3–5) response options on generalization outcomes. Bayesian equivalence testing on generalization results pooled across 4AFC and 2AFC training reveals moderate evidence in favor of their equivalence, suggesting that 4AFC ($n = 118$) versus 2AFC ($n = 178$) training regimes did not differentially influence generalization of category learning (BF $\delta \in I$ vs. $\delta \notin I = 7.994$; BF $\delta \notin I$ vs. $\delta \in I = 0.125$) or final generalization achievement (BF $\delta \in I$ vs. $\delta \notin I = 8.441$; BF $\delta \notin I$ vs. $\delta \in I = 0.118$).

9. General discussion

Category learning studies have often taken the entirely reasonable approach of examining simplified category-learning challenges; one or a few often easily verbalizable diagnostic dimensions with low exemplar variability and a small number of category exemplars have been typical (e.g., Gabay, Dick, Zevin, & Holt, 2015; Lim & Holt, 2011; Maddox, Koslov, Yi, & Chandrasekaran, 2016; Roark, Lehet, Dick, & Holt, 2022). This has been as true for natural exemplars, like non-native speech sounds as well as for novel objects and events. Overall, these studies have informed theories of category learning and have significantly driven our understanding of both basic processes and application. Yet we do not completely understand how factors that impact simplified category learning challenges might play out in more real-world category learning. Here, we developed a novel space of auditory categories that

embodied some of the natural complexity and variability typically encountered in real-world stimuli. Within this space, categories were characterized by many unique exemplars, difficult-to-characterize dimensions, and simultaneous non-diagnostic information.

We observed strong evidence that these categories are learnable even over short-term training. Moreover, this learning generalizes readily to novel exemplars. Across five independent experiments involving 296 listeners, adult participants learned these challenging auditory categories above chance accuracy at the group level. Learning was rapid. There was evidence of learning as early as the first block across all training regimes; for most participants, categorization improved across the 40–45 min of total training. The learning curves across training are consistent with results from a wide variety of category learning studies with simpler category learning challenges. Typically, these studies show evidence of significant learning early in training followed by relatively slow, incremental increases in accuracy across subsequent blocks (e.g., Reetzke et al., 2018; Roark & Holt, 2019; Zeithamova & Maddox, 2006).

As is often the case in category learning studies, there were substantial range individual differences in learning outcomes (e.g., Baese-Berk, Chandrasekaran, & Roark, 2022). We informally examined two potential contributors to these individual differences across our sample of almost 300 participants: (1) experience with Mandarin or another tonal language and (2) musical expertise. Neither was predictive of generalization outcomes (supplemental information can be found at OSF.io).

With this learning and generalization as a baseline, we examined the extent to which manipulations of exemplar variability (Logan et al., 1991), exemplar blocking (Carvalho & Goldstone, 2017), and provision of explicit instruction (Chandrasekaran et al., 2016) – each shown to impact category learning outcomes in prior research – modulate generalization of category learning in a more complicated stimulus space. Under the present category learning challenge, learning was surprisingly consistent across training regimes. As demonstrated by the Bayesian analyses, generalization progress and final generalization achievement were essentially equivalent.

This is quite unexpected given the prior literature. Even participants left to discover diagnostic dimensions implicitly via feedback did not fare more poorly in generalization of category learning than participants provided explicit instruction about where to find category-relevant information. Next, we consider the findings from prior literature and how they diverge from and inform our findings by examining the three training manipulations.

9.1. Exemplar variability

The expectation that training with high variability exemplars produces more robust generalization of category learning has a long history and continues to have a substantial impact on theory and application. As we noted in the introduction, the implications of high variability training have been especially well-investigated in non-native speech category learning (e.g., Logan et al., 1991). Brekelmans et al. (2022) review this literature thoroughly and make a case that evidence is mixed regarding an advantage of high versus low exemplar variability. Moreover, in this well-powered replication of Logan et al. (1991) and Lively et al. (1993), Brekelmans and colleagues observed no learning differences across high and low exemplar variability.

Other studies have shown that the benefit from high variability acoustic training interacts strongly with participants' individual characteristics and perceptual abilities. For example, Perrachione, Lee, Ha, and Wong (2011) demonstrated that high-variability training benefited only learners with already strong perceptual abilities and indeed impeded learners with weaker perceptual abilities. Several other studies have reported variation in the effectiveness of high-variability training for different learners, with some studies finding no beneficial effect of the high-variability condition, and others finding that high exemplar variability in training hinders learning (Fuhrmeister & Myers, 2017, 2020; Sadakata & McQueen, 2014). Further, another recent study has demonstrated that high variability training sets could confer an advantage *or* a disadvantage in voice-identity category learning, depending on stimulus type, the dimension that is varied, and the nature of the post-test (Lavan, Knight, Hazan, & Mcgettigan, 2019).

In summary, emerging evidence challenges the strength and/or consistency of effects of exemplar variability on category learning outcomes. The present results echo these concerns. Here, there was no advantage to generalization progress or ultimate achievement across training with high exemplar variability (480 unique exemplars) versus low exemplar variability (40 unique exemplars).

9.2. Exemplar sequence

A recent meta-analysis revealed that interleaved exemplar presentation tends to benefit learning (Brunmair & Richter, 2019), but vanishingly few studies have examined exemplar sequencing in the auditory modality. Studies examining learning across auditory input of non-native speech sounds – though few in number – have found benefits of blocking, rather than interleaving, category exemplars (Carpenter & Mueller, 2013; Fuhrmeister & Myers, 2020). These studies also found that participants learned to rely on the category-diagnostic dimensions and made error judgments based on category-irrelevant dimensions.

In the present study, exemplars blocked according to the category-diagnostic frequency band initially led to significantly poorer training performance than randomly paired category exemplars. Even so, by the end of training there was no difference in learning outcomes or generalization across training regimes. Any influence of blocked versus interleaved presentation of exemplars in training was ephemeral and contrary to expectations that category-diagnostic blocking would support learning. Participants left to discover category-relevant dimensions through trial-and-error tuned by explicit feedback fared no better or worse than learners who were supported by blocking according to the category-diagnostic dimension.

9.3. Explicit instruction

Explicitly instructing learners about the nature of category-diagnostic dimensions can improve categorization accuracy for non-native speech categories (Chandrasekaran et al., 2016). Other studies have more implicitly "instructed" participants via training methods that exaggerate category-relevant dimensions; these appear to enhance learning compared to control conditions (Ingvalson et al., 2012; Iverson et al., 2005; Jamieson & Morosan, 1986; McCandliss et al., 2002; McClelland et al., 2002).

In the present study, explicit instruction improved early training accuracy compared to implicit support to learning via blocking by category-diagnostic frequency band. But that advantage was fleeting. By the culmination of training, groups' learning and generalization achievements were equivalent. It is possible that simple instructions such as "Listen high!" or "Listen low!" may not be informative enough to direct listeners to the diagnostic dimension. However, we modeled our instructions after those of Chandrasekaran et al. (2016), who instructed listeners that previous participants had succeeded in listening to a specific dimension of sound, and listeners are fully capable of paying explicit attention to one of two spectrally separated dimensions in a range of tasks (Dick et al., 2017; Holt, Tierney, Guerra, Laffere, & Dick, 2018).

9.4. Conclusions

In sum, the present results underscore the robustness of auditory category learning, regardless of training regime. A large, diverse sample of online research participants exhibited the ability to acquire novel auditory categories drawn from a complex acoustic space within 40 min and to generalize this knowledge robustly to novel exemplars. At a group

level, participants across experiments began to categorize at above-chance levels even in the first 10 min of training and generalized this learning to novel exemplars along a similarly speedy timeline. Adult listeners are capable of quickly acquiring complex novel categories that involve substantial simultaneous, non-diagnostic exemplar variability.

Previous work suggested that training regime should have meaningful influence on the speed and generalizability of category learning. Instead, we observed remarkable consistency in learning and generalization across manipulations of exemplar variability, exemplar sequencing, and explicit instruction. To put this in context, consider Experiment 1 versus Experiment 5. In Experiment 1, participants were left to discover category-relevant regularities by foraging the sounds' multiple dimensions and utilizing feedback to shape future responses across four category alternatives. In contrast, Experiment 5 learners had the opportunity to benefit from blocking by (high/low) category-diagnostic information and explicit instruction about the nature of that information. Nonetheless, the two groups' generalization performance was statistically indistinguishable.

What we take away from these studies is that auditory category learning is robust to the introduction of challenging input distributions, despite our best efforts to influence its progress. This rather surprising finding suggests a modicum of caution when drawing conclusions from studies of simpler learning challenges (including many of our own) and inspires us to explore and create new category spaces whose perceptual and dimensional richness begin to approximate that of the natural world.

CRediT authorship contribution statement

Chisom O. Obasih: Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Sahil Luthra:** Conceptualization, Formal analysis, Writing – review & editing, Visualization, Supervision. **Frederic Dick:** Conceptualization, Methodology, Formal analysis, Resources, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Lori L. Holt:** Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors have no competing interests to declare.

Data availability

Research data and supplemental information is available at https://osf.io/bnk9e/. Research materials used to run the study online is available at https://app.gorilla.sc/openmaterials/462043.

Acknowledgements

This work was supported by the National Science Foundation (BCS1950054 to LLH and FD) and the National Institutes of Health (R01DC017734 to LLH and FD and a R01DC017734S1 Research Supplement to COO). COO has been supported by a National Science Foundation Graduate Research Fellowship (DGE1745016, DGE2140739). We thank Christi Gomez and Erin Smith for their assistance in conducting the studies. The Mandarin speech recordings used in creating the stimuli were drawn from Dr. Ran Liu's dissertation research. Research materials and supplemental information can be found online at https://osf.io/bnk9e/.

References

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management, 41*(2), 521–543. https://doi.org/10.1177/0149206314560412

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baese-Berk, M. M., Chandrasekaran, B., & Roark, C. L. (2022). The nature of non-native speech sound representations. *Journal of the Acoustical Society of America, 152*(5), 3025–3034.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*(03), 387–414. https://doi.org/10.1017/S0272263105050175

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*(3), 392–402. https://doi.org/10.3758/s13421-012-0272-7

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research, 74*(4), 245–248. https://doi.org/10.1080/00220671.1981.10885317

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977–985. https://doi.org/10.3758/BF03206911

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language, 126*. https://doi.org/10.1016/j.jml.2022.104352. Article 104352.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition, 41*(5), 671–682. https://doi.org/10.3758/s13421-012-0291-4

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology, 5*(1), 936. https://doi.org/10.3389/fpsyg.2014.00936

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*(1), 281–288. https://doi.org/10.3758/s13423-014-0676-4

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Carvalho, P. F., & Goldstone, R. L. (2021). The most efficient sequence of study depends on the type of test. *Applied Cognitive Psychology, 35*(1), 82–97. https://doi.org/10.1002/acp.3740

Chandrasekaran, B., Yi, H.-G., Smayda, K. E., & Maddox, W. T. (2016). Effect of explicit dimension instruction on speech category learning. *Attention, Perception & Psychophysics, 78*(2), 566–582. https://doi.org/10.3758/s13414-015-0999-x

Chao, Y. R. (1965). *A grammar of spoken Chinese*. Berkeley: Univ of California Press.

Corvin, S., Fauchon, C., Peyron, R., Reby, D., & Mathevon, N. (2022). Adults learn to identify pain in babies' cries. *Current Biology, 32*(15), R824–R825. https://doi.org/10.1016/j.cub.2022.06.076

Dick, F. K., Lehet, M. I., Callaghan, M. F., Keller, T. A., Sereno, M. I., & Holt, L. L. (2017). Extensive Tonotopic mapping across auditory cortex is recapitulated by spectrally directed attention and systematically related to cortical Myeloarchitecture. *The Journal of Neuroscience, 37*(50), 12187–12201. https://doi.org/10.1523/JNEUROSCI.1436-17.2017

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 475–485. https://doi.org/10.1016/j.jarmac.2017.07.005

Estes, K. G., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology, 51*(11), 1517–1528. https://doi.org/10.1037/a0039725

Estes, W. K., & Burke, C. J. (1953). A theory of stimulus variability in learning. *Psychological Review, 60*(4), 276–286. https://doi.org/10.1037/h0055775

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America, 142*(5). https://doi.org/10.1121/1.5009688. EL448–EL454.

Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, & Psychophysics, 82*(4), 2049–2065. https://doi.org/10.3758/s13414-019-01925-y

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance, 41*(4), 1124–1138. https://doi.org/10.1037/xhp0000073

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development, 11*, 66–79. https://doi.org/10.1080/15475441.2014.895249

Gandour, J. (1983). Tone perception in far eastern languages. *Journal of Phonetics, 11*(2), 149–175.

Goldwater, M. B., Hilton, C., & Davis, T. H. (2022). Developing an educational neuroscience of category learning. *Mind, Brain, and Education, 16*(2), 167–182. https://doi.org/10.1111/mbe.12306

Helmer, L. M. L., Weijenberg, R. A. F., de Vries, R., Achterberg, W. P., Lautenbacher, S., Sampson, E. L., & Lobbezoo, F. (2020). Crying out in pain—A systematic review into the validity of vocalization as an indicator for pain. *European Journal of Pain (London, England), 24*(9), 1703–1715. https://doi.org/10.1002/ejp.1623

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Ho, A. T. (1976). The acoustic variation of mandarin tones. *Phonetica, 33*(5), 353–367.

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hearing Research, 366*, 50–64. https://doi.org/10.1016/j.heares.2018.06.014

Howie, J. M. (1976). *Acoustical studies of mandarin vowels and tones* (Vol. 18). London: Cambridge University Press.

Hugenberg, K., & Sacco, D. (2008). Social categorization and stereotyping: How social categorization biases person perception and face memory. *Social and Personality Psychology Compass, 2*, 1052–1072. https://doi.org/10.1111/j.1751-9004.2008.00090.x

Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate /r–l/ on the basis of F3 onset frequency? *Bilingualism: Language and Cognition, 15*(2), 255–274. https://doi.org/10.1017/S1366728911000447

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/−/l/ to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/−/θ/ contrast by francophones. *Perception & Psychophysics, 40*(4), 205–215. https://doi.org/10.3758/BF03211500

JASP Team. (2022). *JASP (0.16.3)*..

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Koutseff, A., Reby, D., Martin, O., Levrero, F., Patural, H., & Mathevon, N. (2018). The acoustic space of pain: Cries as indicators of distress recovering dynamics in pre-verbal infants. *Bioacoustics, 27*(4), 313–325. https://doi.org/10.1080/09524622.2017.1344931

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Knight, S., Hazan, V., & Mcgettigan, C. (2019). The effects of high variability training on voice identity learning. *The Journal of the Acoustical Society of America, 146*, 3053–3054. https://doi.org/10.1121/1.5137589

Leong, C. X. R., Price, J. M., Pitchford, N. J., & Heuven, W. J. B. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLoS One, 13*(10). https://doi.org/10.1371/journal.pone.0204888. Article e0204888.

Lim, S., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science, 35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Liu, R. (2014). *Category learning supporting non-native speech perception: Investigating issues of variability, generalization, and transfer*. Unpublished doctoral dissertation. Carnegie Mellon University.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America, 94*(3 Pt 1), 1242–1255. https://doi.org/10.1121/1.408177

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89*(2), 874–886. https://doi.org/10.1121/1.1894649

Maddox, W. T., Koslov, S., Yi, H.-G., & Chandrasekaran, B. (2016). Performance pressure enhances speech learning. *Applied PsychoLinguistics, 37*(6), 1369–1396. https://doi.org/10.1017/S0142716415000600

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience, 2*(2), 89–108. https://doi.org/10.3758/CABN.2.2.89

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/−/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiology & Behavior, 77*(4), 657–662. https://doi.org/10.1016/S0031-9384(02)00916-2

McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1417–1432. https://doi.org/10.1037/a0032184

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review, 1*(2), 250–254. https://doi.org/10.3758/BF03200776

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Munsinger, H., & Kessen, W. (1966). Preference and recall of stimulus variability. *Journal of Experimental Psychology, 72*, 311–312. https://doi.org/10.1037/h0023457

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition, 155*, 23–29. https://doi.org/10.1016/j.cognition.2016.06.007

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science, 27*(2), 129–135. https://doi.org/10.1177/0963721417740954

Nosofsky, R. M., Slaughter, C., & McDaniel, M. A. (2019). Learning hierarchically organized science categories: Simultaneous instruction at the high and subtype levels. *Cognitive Research: Principles and Implications, 4*. https://doi.org/10.1186/s41235-019-0200-5. Article 48.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461–472. https://doi.org/10.1121/1.3593366

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3, Pt 1), 353–363. https://doi.org/10.1037/h0025953

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction, 23*, 98–114. https://doi.org/10.1016/j.learninstruc.2012.07.003

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences, 26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Reetzke, R., Xie, Z., Llanos, F., & Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology, 28*(9), 1419–1427.e4. https://doi.org/10.1016/j.cub.2018.03.026

Retter, T. L., Jiang, F., Webster, M. A., & Rossion, B. (2020). All-or-none face categorization in the human brain. *NeuroImage, 213*. https://doi.org/10.1016/j.neuroimage.2020.116685. Article 116685.

Roark, C. L., & Holt, L. L. (2019). Perceptual dimensions influence auditory category learning. *Attention, Perception, & Psychophysics, 81*(4), 912–926. https://doi.org/10.3758/s13414-019-01688-6

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The representational glue for incidental category learning is alignment with task-relevant behavior. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 48*(6), 769–784. https://doi.org/10.1037/xlm0001078

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24*(3), 355–367. https://doi.org/10.1007/s10648-012-9201-3

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy, 15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.01318. article 1318.

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/−/l/. *Journal of Phonetics, 66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106*(2), 833–870. https://doi.org/10.1016/j.cognition.2007.05.002

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*(6), 837–848. https://doi.org/10.1002/acp.1598

Waite, S., Grigorian, A., Alexander, R. G., Macknik, S. L., Carrasco, M., Heeger, D. J., & Martinez-Conde, S. (2019). Analysis of perceptual expertise in radiology – Current knowledge and a new perspective. *Frontiers in Human Neuroscience, 13*. https://doi.org/10.3389/fnhum.2019.00213. Article 213.

Wiener, S., Chan, M. K. M., & Ito, K. (2020). Do explicit instruction and high variability phonetic training improve nonnative Speakers' mandarin tone productions? *The Modern Language Journal, 104*(1), 152–168. https://doi.org/10.1111/modl.12619

Wiener, S., Murphy, T. K., Goel, A., Christel, M. G., & Holt, L. L. (2019). Incidental learning of non-speech auditory analogs scaffolds second language learners' perception and production of mandarin lexical tones. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic sciences* (pp. 1699–1703). Australasian Speech Science and Technology Association Inc.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34*(2), 387–398. https://doi.org/10.3758/BF03193416

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*(1), 16–27. https://doi.org/10.3758/s13421-012-0238-9

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction, 22*(3), 215–221. https://doi.org/10.1016/j.learninstruc.2011.11.002