CrossMark

Task and distribution sampling affect auditory category learning

Casey L. Roark[1] · Lori L. Holt[1]

Abstract

There is substantial evidence that two distinct learning systems are engaged in category learning. One is principally engaged when learning requires selective attention to a single dimension (rule-based), and the other is drawn online by categories requiring integration across two or more dimensions (information-integration). This distinction has largely been drawn from studies of visual categories learned via overt category decisions and explicit feedback. Recent research has extended this model to auditory categories, the nature of which introduces new questions for research. With the present experiment, we addressed the influences of incidental versus overt training and category distribution sampling on learning information-integration and rule-based auditory categories. The results demonstrate that the training task influences category learning, with overt feedback generally outperforming incidental feedback. Additionally, distribution sampling (probabilistic or deterministic) and category type (information-integration or rule-based) both affect how well participants are able to learn. Specifically, rule-based categories are learned equivalently, regardless of distribution sampling, whereas information-integration categories are learned better with deterministic than with probabilistic sampling. The interactions of distribution sampling, category type, and kind of feedback impacted category-learning performance, but these interactions have not yet been integrated into existing category-learning models. These results suggest new dimensions for understanding category learning, inspired by the real-world properties of auditory categories.

Keywords Categorization · Audition · Perceptual learning

Introduction

Everyday decisions depend on well-learned category representations, whereby perceptually discriminable experiences are treated as functionally equivalent. For example, one must be able to categorize an animal encountered on the street as "friendly" or "dangerous" in order to decide whether to approach or avoid it. Speech perception can be considered an example of categorization (Holt & Lotto, 2010), in the sense that perceptually discriminable and acoustically variable utterances come to be mapped to phonetic categories. Speech presents a challenging case of auditory perceptual category learn-

ing because phonetic categories are defined by multiple acoustic dimensions that may not be perceptually separable or easily verbalized and whose distributions are highly overlapping (Hillenbrand, Getty, Clark, & Wheeler, 1995; Holt & Lotto, 2008, 2010; Jongman, Wayland, & Wong, 2000; Lisker, 1986; Vallabha, McClelland, Pons, Werker, & Amano, 2007).

An influential cognitive neuroscience framework of category learning, developed in the visual category-learning literature (Ashby, Alfonso-Reese, Turken, & Waldron, 1998), has recently been applied to auditory and speech category learning (Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014; Maddox, Chandrasekaran, Smayda, & Yi, 2013; Yi, Maddox, Mumford, & Chandrasekaran, 2016). The COmpetition of Verbal and Implicit Systems (COVIS) "dual-systems" model of category learning posits two distinct learning systems mediated by the striatum: an explicit system that involves the frontal cortex and the head of the caudate nucleus, along with an implicit system that recruits the putamen and the tail of the caudate.

✉ Casey L. Roark
croark@andrew.cmu.edu

[1]   Department of Psychology, Carnegie Mellon University and the Center for the Neural Basis of Cognition, Pittsburgh, PA 15213, USA

These dual systems are differentially engaged by distinct distributions of category exemplars (Ashby & Maddox, 2011). Rule-based category distributions—which are thought to engage an explicit, reflective, hypothesis-testing system that relies upon working memory and attention—can be distinguished by a single, simple, verbalizable rule (Ashby et al., 1998). Conversely, information-integration category distributions—proposed to engage an implicit, reflexive system that uses procedural learning—can only be distinguished if information from multiple dimensions is integrated at a predecisional stage (Ashby & Gott, 1988). Because this integration is predecisional, the relationship between information-integration categories is often nonverbalizable. In the dual-systems theory, the distribution structure of category exemplars is thought to be the primary determinant of which of the two category-learning systems drives the motor response.

Research on auditory category learning in the context of the dual-systems model has focused mostly on nonnative speech category learning of Mandarin lexical tones via overt training (Chandrasekaran, Koslov, & Maddox, 2014; Maddox & Chandrasekaran, 2014; Yi et al., 2016). In overt training, participants are aware that they are performing a categorization task, make explicit categorization decisions, and are given explicit feedback about these decisions after each trial. Additionally, with overt training, participants are sometimes explicitly informed of the dimensions on which the stimuli vary. Under these conditions, Mandarin tone speech categories are learned best when participants use a reflexive strategy that relies on the implicit system for learning information-integration categories (Chandrasekaran, Koslov, & Maddox, 2014; Maddox & Chandrasekaran, 2014; Yi et al., 2016). The reasoning is that speech categories—like information-integration categories—are defined by highly variable exemplars signaled by multiple acoustic dimensions in a manner that is difficult to verbalize (Chandrasekaran, Koslov, & Maddox, 2014), and thus are learned better via the implicit, reflexive learning system. In support of this theory, functional neuroimaging has revealed that the patterns of corticostriatal activation during speech category learning are more consistent with involvement of the implicit, reflexive system posited by COVIS (Yi et al., 2016).

These recent results suggest the promise of dual-systems theory for understanding auditory—and in particular, speech—category learning. However, the categorization challenges presented by auditory (and speech) signals are somewhat different from those invited by the visual categories that have served as the principal testing ground for dual-systems theory (Holt & Lotto, 2010). Important questions remain open about how well auditory category learning aligns with the predictions of dual-systems theory. In the present research, we examined the predictions of dual-systems theory in the context of how manipulations of task, distribution sampling, and category type affect auditory category learning.

We tested these questions using novel, artificial nonspeech auditory categories. Although nonspeech categories have not been used as frequently, in the context of examining dual-systems theory, as nonnative speech categories (but see Chandrasekaran, Koslov, & Maddox, 2014), they provide us with the control to precisely define and manipulate category distributions, distribution sampling, and the course of learning as a function of different learning tasks. Thus, in the same way that very simple visual dimensions have been used productively to understand the learning systems available for categorizing more complex objects in the natural world, we employed nonspeech sounds in order to understand the processes available to support learning more complex speech categories. We next describe the rationale for focusing on task, category distribution, and distribution sampling in the present research.

Task

Nearly all studies of visual or auditory category learning from a dual-systems perspective have used an overt training task (Ashby, Maddox, & Bohil, 2002; Chandrasekaran, Yi, & Maddox, 2014; Dunn, Newell, & Kalish, 2012; Ell, Ing, & Maddox, 2009; Maddox, Filoteo, Hejl, & Ing, 2004; Maddox, Love, Glass, & Filoteo, 2008; for a discussion of unsupervised learning, see Ashby, Queller, & Berretty, 1999). Participants are told how many categories exist, they are instructed that the goal is to categorize the stimuli, and they are provided with corrective trial-by-trial feedback on the category decisions. In an exception that used an unsupervised category-training paradigm without overt feedback, Ashby et al. (1999) found that rule-based (RB) categories, distinguished by a single, simple, verbalizable rule, could be learned without feedback, but that information-integration (II) categories, which require information from multiple dimensions to be integrated at a predecisional stage, could not. The researchers concluded that learning II categories is critically dependent on feedback, whereas learning RB categories can occur without feedback. Thus, due to the inability of participants to learn II categories without feedback, the majority of research from the dual-systems literature has utilized supervised-learning tasks with overt feedback.

Yet, recent research in auditory category learning suggests an alternative approach that is neither wholly explicit nor unsupervised. *Incidental learning* occurs without instructions to categorize, overt category decisions, or explicit feedback. Instead, sound categories are learned incidentally by virtue

of their relationship to success in performing a primary task distinct from auditory category learning (Gabay, Dick, Zevin, & Holt, 2015; Lim & Holt, 2011; Lim, Lacerda, & Holt, 2015; Liu & Holt, 2011; Vlahou, Protopapas, & Seitz, 2012; Wade & Holt, 2005). For example, when auditory categories' exemplars are presented in a manner that correlates with where a visual "x" will next appear on the screen in a visual detection task, participants incidentally learn complex auditory categories, including speech, in the course of performing the visual detection task (Gabay et al., 2015; Liu, 2014). Incidental auditory category learning is also apparent across more challenging primary tasks, such as navigating a videogame environment in which sound categories are correlated with aspects of the input that support success in the primary, game navigation, task (Gabay et al., 2015; Lim & Holt, 2011; Lim et al., 2015; Liu & Holt, 2011; Wade & Holt, 2005). Inasmuch as this incidental category learning proceeds even without instructions about the importance of the sounds, knowledge of the existence of auditory categories, overt category decisions, or explicit feedback about categorization, it may better model aspects of category learning in natural environments whereby correlated objects, events, and actions across modalities are available as structure that may guide learning (Gabay et al., 2015; Wade & Holt, 2005).

Although prior studies of category learning in a dual-systems theory framework have relied nearly exclusively on overt training, extensive evidence has demonstrated the importance of task variables in category learning. Investigations of the distinction between II and RB category learning have emphasized the significance of feedback timing (Dunn et al., 2012; Ell et al., 2009; Maddox, Ashby, & Bohil, 2003; Maddox, Ashby, Ing, & Pickering, 2004; Maddox & Ing, 2005; Smith et al., 2014; Worthy, Markman, & Maddox, 2013), amount of feedback and feedback type (Ashby et al., 2002; Ashby & O'Brien, 2007; Ashby et al., 1999; Dunn et al., 2012; Goudbeek, Cutler, & Smits, 2008; Goudbeek, Swingley, & Smits, 2009; Maddox et al., 2008), and changing instructions to participants (Chandrasekaran, Yi, Smayda, & Maddox, 2016; Grimm & Maddox, 2013) However, the question of whether category learning for the II and RB categories differs across overt versus incidental training has yet to been investigated.

We hypothesize that II categories—which are difficult to verbalize and require predecisional integration—may benefit more from incidental training tasks, in which attention is directed toward a primary task and away from decisions about category exemplars. In contrast, consistent with the dual-systems theory, RB categories may benefit more from an overt training task, in which attention can be directed toward the stimuli and features that distinguish the categories. The overt task may encourage more explicit, verbalizable hypothesis testing to support learning the RB categories. Since this kind of explicit strategizing can be detrimental for II category learning

(Grimm & Maddox, 2013), incidental training may be beneficial for learning II categories.

Category distribution sampling

The categorization challenges presented by speech—and, indeed, most natural categories—almost always involve complex, *probabilistic* category exemplar distributions that overlap in acoustic space (Kuhl et al., 1997; Lotto, Sato, & Diehl, 2004; McMurray & Jongman, 2011; Peterson & Barney, 1952). Yet, many studies of speech and nonspeech auditory category learning have examined learning across nonoverlapping, *deterministic* distributions that are well differentiated in acoustic space and characterized by a small number of exemplars experienced repeatedly across training (Holt & Lotto, 2006; Kluender, Lotto, Holt, & Bloedel, 1998; Kuhl, 1991; Lim & Holt, 2011; Mirman, Holt, & McClelland, 2004; Wade & Holt, 2005). Even when more probabilistic distributions of natural speech productions have been used to study category learning among nonnative listeners (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991; Yi et al., 2016), the impact of distribution sampling on learning has not been a focus of the investigation.

The approaches using probabilistic and deterministic distributions differ on several dimensions. However, each approach is meant to approximate some sampling that is similar to real-world categories, such as speech. Deterministic distributions are highly stylized, their categories do not overlap, and there are relatively few exemplars. Probabilistic distributions are randomly sampled, their categories are often overlapping, and there are many possible exemplars. It is important to understand what, if any, effect sampling from these different kinds of distributions has on category learning. To the extent that each effectively approximates sampling from naturalistic categories, then the approach to distribution sampling should not have an impact on learning. However, it is entirely possible that sparser, nonoverlapping, more stylistically sampled distributions may be learned in a different manner from denser, overlapping, randomly sampled distributions.

This issue has not been investigated thoroughly, even with the artificially constructed visual categories upon which the dual-systems model of categorization is based. To our knowledge, only one study has addressed an issue similar to that of probabilistic versus deterministic category distribution sampling in the visual domain. Ell and Ashby (2006) examined the impact of category overlap on learning. The degree to which the exemplars from the different categories were drawn from overlapping versus entirely distinct regions of stimulus space impacted the learning of visual II categories, but not of visual RB categories. Specifically, when the categories' exemplars

were moderately overlapping across II distributions, participants were able to use optimal II strategies in category learning; however, with too much or too little overlap of the II distributions, participants relied on suboptimal, RB strategies. This indicates that at least some aspects of the sampling distribution—specifically, overlap—may influence the course of learning. Thus, it is important to examine the potential learning differences between carefully sampled deterministic distributions and randomly sampled probabilistic distributions, especially in light of the fact that the auditory category-learning literature has employed them somewhat interchangeably.

In the present study, we manipulate whether the II and RB category distributions were sampled probabilistically or deterministically in acoustic space. Examining the interaction of category distribution type (II vs. RB), task (incidental vs. overt), and distribution sampling (deterministic [nonoverlapping] vs. probabilistic [overlapping]) is important, given the ubiquity of probabilistic, overlapping category distributions in speech and other natural categories, including visual categories (e.g., Nosofsky, Sanders, Meagher, & Douglas, 2018).

In the present study, we investigated learning across highly stylized, deterministic distributions of sound category exemplars like those that have characterized most auditory category-learning studies to date. We also examined learning across categories defined more probabilistically. There is not a large literature to support strong predictions about the effects of these different sampling distributions on learning. Drawing from Ell and Ashby's (2006) results, one might predict that category overlap in deterministic versus probabilistic sampling would impact the learning of II, but not of RB, categories. Participants' learning categories sampled deterministically may be biased toward explicit strategies, thereby impeding II learning and benefiting RB learning. To the extent that explicit strategies influence the learning, there may also be an interaction between task type (incidental, overt) and distribution sampling. As compared to probabilistic distributions learned through incidental training, learning via overt training may be better across deterministic, nonoverlapping category distributions that are easy to learn with verbalizable rules and for which the feedback is perfectly consistent with optimal strategies.

Summary

In the present experiment, we examined the impact of task and category distribution sampling on learning four auditory categories defined by either II or RB stimulus distributions. We trained separate groups of participants with either a traditional, overt categorization paradigm with explicit feedback on every trial, or with an incidental paradigm in which neither the categorization decisions nor feedback were explicit task

demands. Finally, we varied the nature of the distribution sampling in order to examine the influence of the probabilistic or deterministic nature of the category distributions on learning. We directly examined the influences of category type, task, and distribution sampling on auditory category learning using within-training metrics, as well as a common overt labeling task administered posttraining, to assess the generalization of learning to novel category exemplars.

Method

Participants

In all, 166 adults affiliated with Carnegie Mellon University, ages 18–25 years (89 females, 77 males), participated for partial course credit or a small payment ($10). All participants had normal or corrected-to-normal vision and reported normal hearing. A total of eight conditions were tested, which varied by training task, category distribution, and distribution sampling (Table 1). Participants were trained on either an incidental or an overt task, learned to categorize either rule-based or information-integration category distributions, and the distributions were either probabilistic or deterministic in their sampling. An additional five participants were run but were excluded from all analyses because of equipment failure.

Stimuli

The learning challenge differed across four conditions defined by the category input distributions. Each condition had four categories separated by optimal decision boundaries, as is shown in Fig. 1. Participants were trained on either II distributions (Fig. 1b and d) or RB distributions (Fig. 1a and c) that were sampled either deterministically (Fig. 1a and b) or probabilistically (Fig. 1c and d) in acoustic space. For the deterministic distributions, it was possible to define optimal decision boundaries that would classify the exemplars with perfect accuracy, because there was no overlap between the categories. However, no decision boundary would result in 100% accuracy for the probabilistic distributions, because exemplars

Table 1 Number of participants in each condition

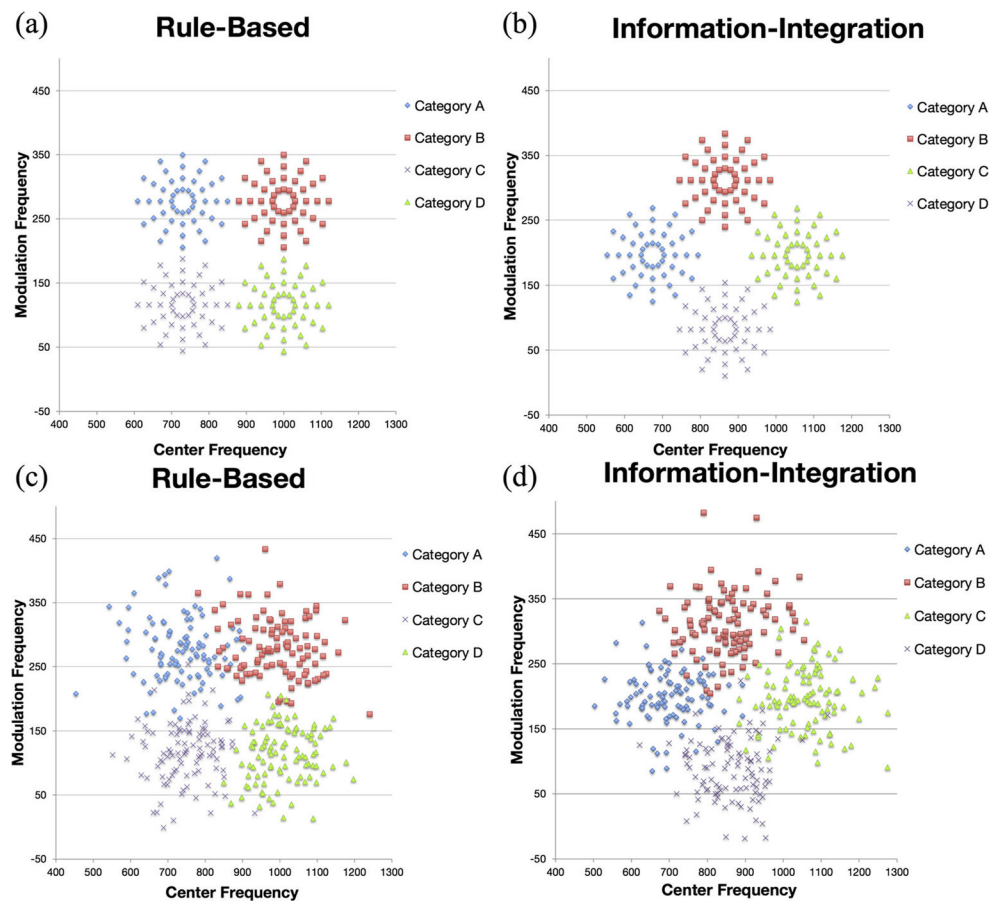| Task	Information-Integration	Rule-Based
Deterministic sampling		
Incidental task	21	20
Overt task	21	21
Probabilistic sampling		
Incidental task	20	21
Overt task	21	21

Fig. 1 Stimulus distributions. (**a**) Rule-based deterministic. (**b**) Information-integration deterministic. (**c**) Rule-based probabilistic. (**d**) Information-integration probabilistic.

could probabilistically belong to more than one category due to category overlap. Optimal performance for the probabilistic RB and II conditions was 90.5% and 90.25%, respectively. We used these moderate levels of overlap to reflect the similar levels of overlap in Ell and Ashby (2006), who found that moderate levels of overlap did not hinder learning for either II or RB categories.

The two-dimensional acoustic space from which the stimulus exemplars were sampled was defined by two dimensions: center, or carrier, frequency (CF) and modulation frequency (MF). CF can be approximately described as the pitch of the tone, and MF can be described as the warble of the tone. We chose these particular acoustic dimensions because they had been used in an existing auditory category-learning study (Holt & Lotto, 2006), demonstrating that listeners are able to learn categories defined by these acoustic dimensions. Additionally, in this previous study these dimensions were psychoacoustically matched for discriminability across the approximate range used in the present study (Holt & Lotto, 2006). Furthermore, manipulation of sounds across these dimensions creates highly artificial exemplars that participants are unlikely to have heard previously. In the same manner that Gabor patches provide a simple stimulus to manipulate

parametrically in the visual domain, these simple acoustic stimuli provide us with a test bed for auditory category learning. Each 300-ms stimulus was defined by a CF modulated with a depth of 100 Hz at the corresponding MF, with the overall energy root-mean-square-matched across exemplars and all synthesis accomplished using MATLAB R2014a (The MathWorks, Inc., Natick, MA).

The stimulus distributions for the deterministic RB categories were adapted from Holt and Lotto (2006). The stimulus distributions for the deterministic II categories were generated by rotating the RB stimuli counterclockwise in acoustic space by 45 deg. The deterministic categories were highly stylized, in the manner of those in previous auditory studies (Holt & Lotto, 2006; Kluender et al., 1998; Kuhl, 1991). The deterministic categories also did not overlap and had a relatively small number of exemplars per category. For the deterministic categories (Fig. 1a and b), there were 48 exemplars per category plus the centroid of each category. Half of the exemplars were used during the training phase (24 exemplars/category, 96 total stimuli), and half of the exemplars plus the centroid were reserved for the generalization test phase (100 stimuli).

To create the probabilistic stimulus distributions, we defined the underlying distributions to have the same means in

CF×MF acoustic space as each of the deterministic category distributions. We increased the number of exemplars and the variance of the category distributions in order to manipulate the sampling. We then sampled randomly from the defined distribution; the random sampling resulted in means, variances, and covariances that varied somewhat from those defined in the underlying distribution (see the Appendix). The probabilistic distributions (Fig. 1c and d) were created with MATLAB R2012 and had 100 exemplars per category. Half of the exemplars were used during the training phase (50 stimuli/category, 200 total stimuli), and half of the exemplars were reserved for the generalization test phase (200 stimuli). Unlike in the deterministic conditions, not all of the exemplars from the probabilistic distributions were played in the test phase. Rather, they were randomly sampled for each participant, for a total of 100 trials.

Task

Participants were trained on one of two training tasks: incidental or overt (see Fig. 2). After training, all participants were tested on an overt generalization posttest, which included exemplars not experienced during training. By adding an overt generalization posttest, we were able to compare learning between the incidental and overt training tasks and to better

understand learning that extends beyond the individual exemplars experienced in training, to include generalization to novel exemplars consistent with the category distribution.

Training task: Incidental

Research in our laboratory has demonstrated the effectiveness of a simple incidental-learning task, the SMART task, in training listeners to categorize sounds (Gabay et al., 2015; Liu, 2014). This paradigm was adapted as a highly simplified version of a videogame training paradigm that has successfully been used to train speech and nonspeech categories incidentally, without overt training or feedback (e.g., Lim & Holt, 2011; Wade & Holt, 2005). In the SMART task, the primary objective is to rapidly detect the appearance of a visual target in one of four possible screen locations by pressing a key corresponding to that screen location.

Within-trial category–exemplar variability On each trial, five unique exemplars drawn from one of the four sound categories preceded the appearance of the visual target. For the deterministic categories, the sound categories perfectly predicted the location of the upcoming visual target and, consequently, the action required to complete the visual detection task. The overlap of the probabilistic categories made it such that the
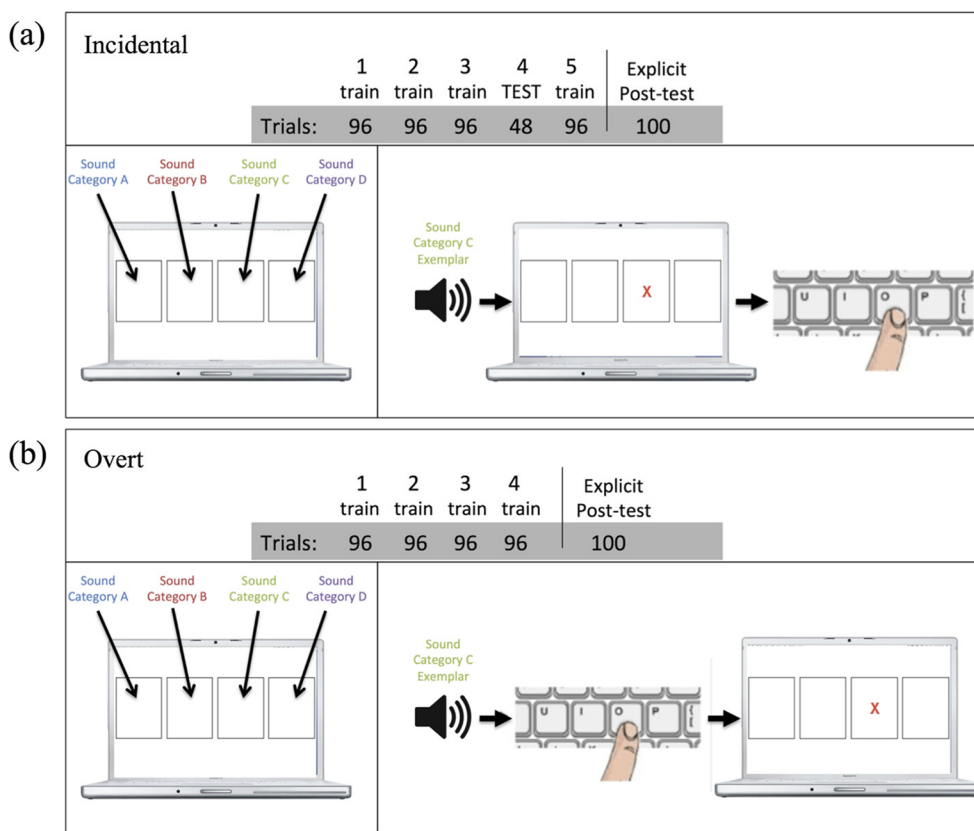

Fig. 2 Outline of the tasks used in the present experiment: (**a**) Incidental training task. (**b**) Overt training task.

exemplars played on a single trial might not be equally representative exemplars of the category based on the optimal boundary between the categories. However, the exemplars played on each trial were always drawn from a single category, so the sound category was predictive of the visual target location. Prior research had demonstrated that participants learn auditory categories incidentally in the SMART task and generalize learning to labeling novel exemplars in an overt, posttraining labeling task (Gabay et al., 2015). The generalization to novel sound category exemplars underscores the important point that this learning is not a simple sound-to-location mapping. The inherent variability in sound category exemplars encourages participants to learn sound *categories* that robustly generalize to guide subsequent responses to novel, unfamiliar exemplars (Gabay et al., 2015).

In the present experiment, we used the SMART task to assess incidental auditory category learning and generalization across deterministic and probabilistic II and RB category distributions, using a covert reaction time (RT) measure during online learning and an overt, posttraining category-labeling task.

Covert reaction time measure of learning Participants were instructed that their task was to indicate where a visual target (a red X) appeared on the screen by responding with the corresponding keyboard button (responses were made with the U, I, O, and P buttons; see Fig. 2). Participants were told that sounds would precede the X, but no mention was made about the relationship between the sounds and the location of the X.

Participants experienced three training blocks (96 trials/ block) in which each of the four sound categories predicted the appearance of the visual target in one specific location. Subsequently, they completed a brief random test block (48 trials), in which the mapping of sound categories to visual target locations was fully random; any sound category exemplar could precede the presentation of the visual target in any position with equal probability. Following the approach of Gabay et al. (2015), this block was shorter so as to avoid extensive exposure to a random mapping that might erode category learning across the training blocks. The final block was one last training block. This block was included in order to reinstate category learning prior to the posttraining generalization posttest.

On each trial, the sound category was chosen pseudorandomly (random shuffle of a fixed number of trials for each category per block). Then, five exemplars were randomly selected from the pool of category exemplars. The 300-ms exemplars were each presented once, with a 50-ms interstimulus silent interval. The final exemplar was followed by a 500-ms silence, after which a red X appeared in the location is associated with the sound category. The trial structure was identical for the random test block, except that the red X appeared in a randomly selected location instead of the location that had been associated

with the sound category in the training blocks. Participants responded, indicating the location of the red X by pressing the associated button as quickly as possible. Reaction times were measured as the time lapsed from the onset of the visual detection target to the press of the response key. After each experimental block, participants were encouraged to rest briefly.

The random test block provided for a covert measure of incidental category learning to be collected online during training. If participants incidentally learned the sound categories in service of guiding their visual detection behavior, then eliminating the consistent relationship between category and location in the random test block should slow responses to the visual target. As was evident in prior studies (Gabay et al., 2015; Liu, 2014), learning should become apparent through an RT cost ($RT_{Block4} - RT_{Block3}$) between Block 4 (random) and Block 3 (consistent). This RT cost would serve as our covert RT measure of learning across incidental-training conditions.

Overt measures of learning and generalization An overt labeling task immediately followed the SMART task. Before the start of the generalization posttest, participants were informed that the location of the X had been associated with the sounds that preceded it in the SMART task and that they should respond with a keypress in order to guess the location where the visual target would be most likely to appear. On each of 100 trials (25 trials/ category), five novel sound category exemplars not experienced in the SMART task were randomly selected from the pool of generalization stimuli and presented with the same timing characteristics as in training. As in training, the generalization posttest had within-trial variability. However, no visual targets appear in this task, thereby providing no feedback. This overt labeling task provided an explicit test of category learning and its generalization to novel exemplars not experienced in training.

Training task: Overt

The overt task modeled the training approach taken in most studies of category learning (Ashby et al., 2002; Yi et al., 2016), while aligning closely with many of the task details of the incidental SMART task (Gabay et al., 2015). In the overt task, participants experienced the same kind of multimodal location-to-sound category mapping as did the incidental (SMART) task participants, but they were explicitly informed of the relationship between the sounds and the visuomotor targets, made explicit categorization decisions, and were given overt corrective feedback following each categorization decision.

As in the incidental-training task, these participants first experienced three training blocks (96 trials/block) in which sound categories predicted the position of the visual target. However, unlike in the incidental task, participants did not receive a random test block, because such a block could play no role in the covert assessment of learning in this overt feedback version of the

paradigm. In sum, the overt task involved four training blocks followed by the generalization posttest.

For each trial, the sound category was first chosen pseudorandomly (random shuffle of a fixed number of trials for each category per block). As in the incidental task, there was within-trial variability in the overt task. Five category exemplars were then randomly selected from the pool of exemplars for that category. The 300-ms exemplars were sequentially presented at the onset of the trial, with 50-ms silent intervals. Participants pressed a button (U, I, O, or P) to indicate which visual location they believed was associated with the sound. A 500-ms silence followed the response, after which a red X appeared in the visual location associated with the sound category presented on that trial, as feedback about the category identity.

After each block, participants were encouraged to rest briefly. Button presses were considered correct if they corresponded to the correct visual location mapped to the trial's sound category, providing a measure of accuracy across blocks.

Overt generalization posttest Immediately after the last block of training, participants engaged in an overt categorization test with within-trial variability that was identical to that in the generalization posttest described for incidental training.

Results

We describe the results separately for the incidental and overt training conditions because some measures were task-specific. The incidental training task provided covert online measures of learning via RTs (Fig. 3), whereas the overt task did not. For the overt task, the relevant behavior was accuracy across

training blocks (Fig. 4). Across the incidental and overt conditions, a common posttraining overt labeling task was included in order to assess the generalization of learning (Fig. 5).

Training task: Incidental

Reaction time filtering We filtered the RTs in order to include only trials on which participants were accurate in responding to the X on the screen and for which the RTs were less than 1,500 ms and greater than 100 ms. A total of 3.77% of the trials were excluded across all conditions (2.74% of trials were excluded for II–probabilistic, 3.74% for RB–probabilistic, 4.46% for II–deterministic, and 4.11% for RB–deterministic).

Covert reaction time measure of learning The covert measure of learning, RT cost, provided an online measure of incidental category learning. We predicted that eliminating the consistent relationship between the sound category and the upcoming location of the visual target established across Blocks 1–3 would slow the RTs to detecting the visual target in Block 4, as compared to Block 3, as expressed by a positive RT cost ($RT_{Block4} - RT_{Block3}$). Since exemplars varied on a trial-by-trial basis, this would be indicative of sound category learning.

Figure 3 shows the average RTs for each condition. Following the approach of prior research (Gabay et al., 2015), we first examined the RT cost by conducting paired-sample t tests comparing the Block 4 and Block 3 RTs for each condition. There were significant RT costs, indicative of incidental auditory category learning, for both the deterministic II [$M = 33.7$ ms, $t(20) = 3.39$, $p = .003$, Cohen's $d = 0.56$] and RB [$M = 22.3$ ms, $t(19) = 2.79$, $p = .012$, Cohen's $d = 0.38$] distributions. For the probabilistic distributions, only the RB condition resulted in a significant RT cost indicative of incidental auditory category learning [$M = 31.2$ ms, $t(20) = 3.39$, $p = .003$, Cohen's $d = 0.45$]. The RT cost
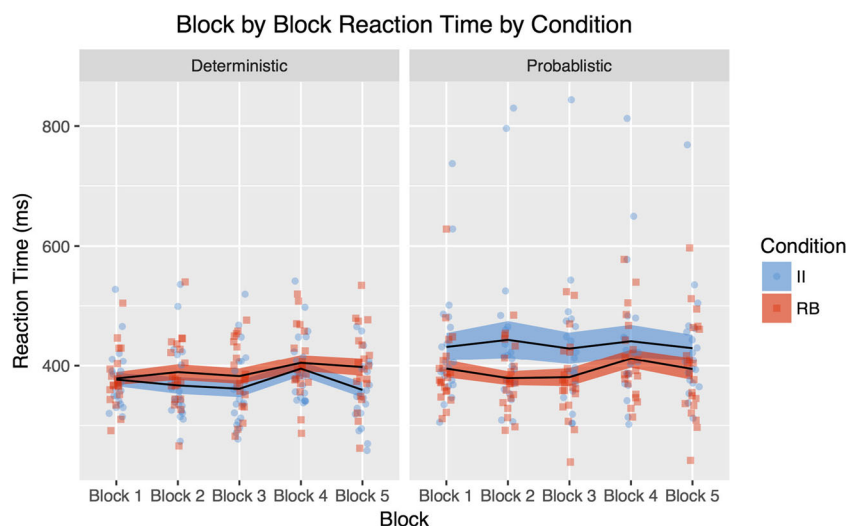

Fig. 3 Average RTs during the incidental training task. Ribbon error bars represent the standard errors of the means. Individual points represent the individual participants' averages. The participants in the II condition are shown as circles, and the participants in the RB condition are shown as squares.
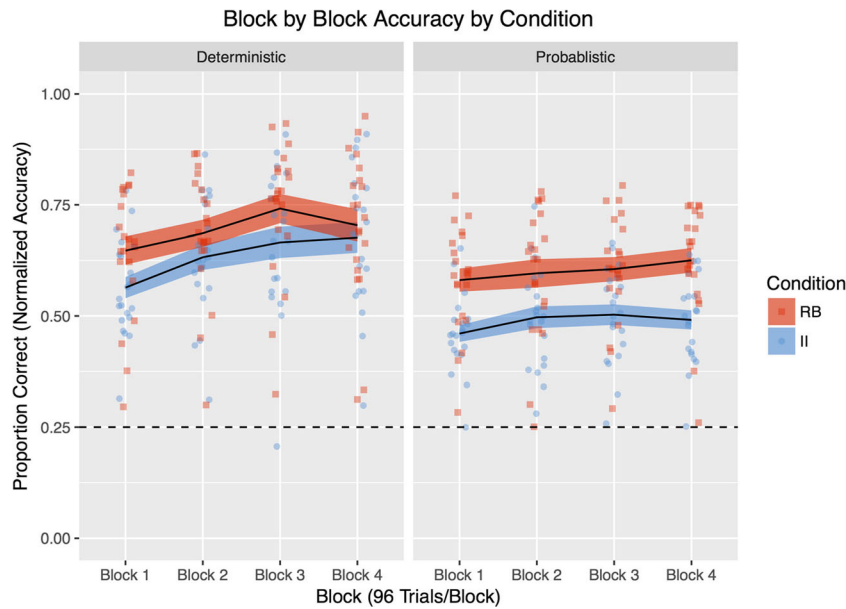
Block by Block Accuracy by Condition

Fig. 4 Block-by-block performance during the overt training task. The dotted line denotes chance performance (25%). Ribbon error bars represent the standard errors of the means. Individual points represent the individual participants' averages. The performance for participants in the II condition is shown by circles, and performance for the RB condition is shown by squares.

for the probabilistic II condition was not significant [$M = 12.3$ ms, $t(19) = 1.10$, $p = .29$, Cohen's $d = 0.10$]. Thus, with the exception of the probabilistic II condition, each group exhibited significant incidental auditory category learning, as indexed by the covert, online RT cost measure.

We next asked whether the magnitude of the RT cost varied as a function of the learning challenges presented by the different conditions. It did not. A 2 (Category Distribution) × 2 (Distribution Sampling) between-subjects analysis of variance (ANOVA) revealed that the magnitude of the RT cost did not

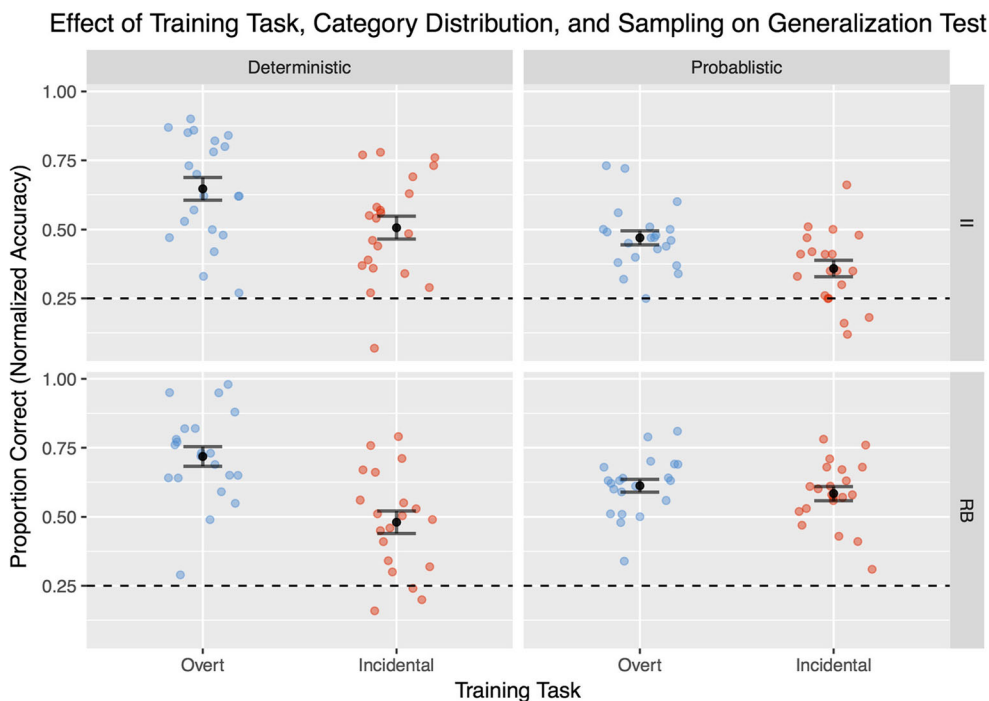## Effect of Training Task, Category Distribution, and Sampling on Generalization Test

Fig. 5 Generalization test performance for all conditions. The dotted lines denote chance performance (25%). Error bars represent the standard errors of the means. Individual points represent the individual participants' averages.

depend on either category distribution [II vs. RB; $F(1, 78) = 0.15$, $p = .70$, $\eta_p^2 = .002$] or sampling [deterministic vs. probabilistic; $F(1, 78) = 0.42$, $p = .52$, $\eta_p^2 = .005$], and there was no interaction [$F(1, 78) = 2.48$, $p = .12$, $\eta_p^2 = .031$]. Although all except the probabilistic II condition demonstrated evidence of incidental category learning according to the RT cost measure, the magnitude of the RT cost was not dependent on either the category distribution or sampling.

Average reaction times We also compared the average RTs across all training blocks as a function of condition. The participants learning deterministic category distributions were marginally faster ($M = 381$ ms, $SE = 11.4$) to respond to the visual targets than were the participants learning probabilistic category distributions ($M = 413$ ms, $SE = 11.4$) [$F(1, 78) = 4.06$, $p = .047$, $\eta_p^2 = .049$]. It appears that the more highly overlapping probabilistic category distributions slowed visual target detection somewhat, relative to simpler, more coherent deterministic category distributions. One possibility is that participants might be sensitive to the deterministic versus probabilistic structure of the category input distributions. However, group differences cannot be ruled out in this between-subjects design. We observed no effect of category distribution (II, RB) on the average RT [$F(1, 78) = 0.56$, $p = .46$, $\eta_p^2 = .007$], and only a marginal interaction between category distribution and sampling [$F(1, 78) = 3.63$, $p = .060$, $\eta_p^2 = .045$].

Note that we did not have any a priori predictions that the conditions would differ in average RTs. Thus, examining average RTs served as a manipulation check to make sure that the different conditions did not differ in RTs. However, one participant in the probabilistic II condition was consistently slower than the other participants. We also ran the analyses excluding this participant, who was more than three standard deviations above the mean on four of the five blocks. No other participant was more than three standard deviations above the mean on any individual block. Examining the RT data without this participant produced largely the same results, except in terms of average RT. When we excluded the outlier from the probabilistic II condition, the effect of sampling distribution on average RT disappeared, such that there were no significant differences between the probabilistic and deterministic average RTs [probabilistic $M = 404$ ms, $SE = 9.6$; deterministic $M = 381$ ms, $SE = 11.4$; $F(1, 77) = 2.94$, $p = .090$, $\eta_p^2 = .037$]. Similar to the results including the outlier, the effect of category distribution and interaction were not significant after excluding the outlier from the probabilistic II condition [category distribution, $F(1, 77) = .042$, $p = .84$, $\eta_p^2 = .001$; interaction, $F(1, 77) = 2.52$, $p = .12$, $\eta_p^2 = .032$]. After excluding the outlier in the probabilistic II condition, there were no differences between the groups in average RTs, indicating that any differences in learning were not tied to differences in RT.

Training task: Overt

Normalization Recall that an optimal observer would achieve 100% accuracy in the deterministic conditions, but only 90.25% or 90.5% accuracy in the probabilistic conditions. To account for this difference, we first computed normalized accuracy values for the data from the probabilistic learning conditions as (normalized accuracy = raw accuracy/optimal accuracy), with optimal accuracy = .9025 and .905 (for probabilistic II and RB, respectively). All comparisons were conducted with these normalized accuracy values. We note that none of the qualitative patterns of results changed as a result of normalization; it simply provided for equitable cross-condition comparisons.

Accuracy across blocks In the overt training task, (normalized) accuracy across blocks was the principal measure of category learning; Fig. 4 plots these results. Examining performance across blocks with a $2 \times 2 \times 4$ repeated measures ANOVA [Category Distribution (II, RB) × Sampling (deterministic, probabilistic) × Block], we found that independent of category distribution or sampling, participants generally improved with training across blocks [$F(2.6, 209.2)^1 = 15.42$, $p < .001$, $\eta_p^2 = .162$]. Participants in the deterministic condition learned more across blocks than did participants in the probabilistic condition [$F(2.6, 209.2) = 3.25$, $p = .029$, $\eta_p^2 = .039$]. Performance across blocks was not impacted by the interaction of category distribution (II, RB) and sampling [$F(2.6, 209.2) = 1.40$, $p = .25$, $\eta_p^2 = .017$], and there was no advantage for learning II versus RB category distributions across training [$F(2.6, 209.2) = 0.54$, $p = .63$, $\eta_p^2 = .007$]. Generally, this may indicate that sampling—whether the categories were sampled deterministically or probabilistically—was the main driver of the difference in improvement across conditions, rather than category distribution or the interaction between category distribution and sampling. Deterministic auditory categories were learned more readily than probabilistic auditory categories. The overall patterns of learning were differentiated by the distribution sampling, not by category distribution.

In examining overall accuracy, rather than performance across blocks, we found that participants learning the RB categories had higher average accuracy than did participants learning the II categories [$F(1, 80) = 11.58$, $p = .001$, $\eta_p^2 = .13$]. Deterministic category input distributions were learned more easily than probabilistic category input distributions [$F(1, 80) = 21.87$, $p < .001$, $\eta_p^2 = .22$]. There was no interaction between category distribution and sampling [$F(1, 80) = 1.12$, $p = .30$, $\eta_p^2 = .014$], indicating that the RB conditions were learned better than the II conditions for both deterministic and probabilistic category distributions. Moreover, the deterministic distributions were learned better than the probabilistic distributions for both II and RB

---

[1] Huynh–Feldt corrected because Mauchly's test of sphericity was significant, $p < .001$.

categories. These results support the prediction that overt training should benefit RB category learning. We also predicted that RB might not be affected by the sampling distribution. In contrast to our predictions, the sampling distribution affected learning for both II and RB categories, such that deterministic sampling led to better overall category learning than did probabilistic sampling.

We also note that learning was quite rapid. Significant learning was evident in the first 96-trial block of training in each of the conditions [chance = 25%; deterministic II, $t(20) = 13.15$, $p < .001$, $M = 56.4\%$, Cohen's $d = 5.88$; deterministic RB, $t(20) = 12.62$, $p < .001$, $M = 64.7\%$, Cohen's $d = 5.64$; probabilistic II, $t(20) = 12.41$, $p < .001$, $M = 51.0\%$, Cohen's $d = 4.97$; probabilistic RB, $t(20) = 13.57$, $p < .001$, $M = 64.2\%$, Cohen's $d = 5.66$].

Generalization posttest: Incidental and overt training

Both incidental and overt training resulted in auditory category learning, as assessed by the task-specific measures during training. Here, we examine the posttraining measure of generalization of category learning, assessed using a common task across the groups trained incidentally and overtly. Performance during the generalization posttest is shown in Fig. 5.

Normalization As for the overt training task, we normalized the generalization test accuracies. We computed normalized accuracy values for data from the probabilistic learning conditions as (normalized accuracy = raw accuracy/optimal accuracy), with optimal accuracy = .9025 and .905 (for probabilistic II and RB, respectively). We did not compute normalized accuracies for the deterministic conditions because the optimal accuracy was 100%. We note that none of the qualitative patterns of results changed as a result of the normalization; it simply provided for fair cross-condition comparisons.

To compare learning in all conditions, we ran a 2 × 2 × 2 ANOVA on Training Task (Incidental vs. Overt) × Category Distribution (II vs. RB) × Sampling (deterministic vs. probabilistic). This allowed us to determine the aspects of the training task and/or stimulus components that drove the differences among conditions. The COVIS model predicts that the main driver of differences in performance would be category distribution, because the II and RB categories are learned by distinct neural systems. Additionally, we examined the impacts of training task and distribution sampling on performance. We predicted that learning differences would depend on the category distribution, but also on the training task and sampling. In comparing the generalization test performance, we found a marginally significant three-way interaction among training task, category distribution, and sampling [$F(1, 158) = 3.75$, $p = .055$, $\eta_{\mathrm{p}}^2 = .023$]. To understand the causes of this marginal interaction, we looked more closely at the two-way interactions.

We predicted that performance on the II and RB category distributions would depend on training task, such that overt training would better support learning RB categories and incidental training would support learning the II categories. We did not find support for this hypothesis; the interaction was not significant [$F(1, 158) = 0.002$, $p = .96$, $\eta_{\mathrm{p}}^2 = .000$]. Instead, we found significantly better generalization of category learning for overt than for incidental training, irrespective of whether the categories were RB or II. Ignoring distribution sampling, for both II and RB, overt training resulted in significantly greater generalization of category learning than did incidental training [RB, $t(81) = 3.78$, $p < .001$, Cohen's $d = 0.84$; II, $t(81) = 3.41$, $p = .001$, Cohen's $d = 0.76$].

We also predicted that the incidental and overt training tasks might have different effects on probabilistic and deterministic category learning. We predicted that overt training would lead to better performance for deterministic than for probabilistic categories, but that incidental training would lead to better performance for probabilistic than for deterministic categories. We found that performance for the probabilistic and deterministic distributions did depend on training task [$F(1, 158) = 5.18$, $p = .024$, $\eta_{\mathrm{p}}^2 = .032$]. In line with our predictions, overt training led to better generalization of category learning for deterministic category than for probabilistic distributions [$t(82) = 2.37$, $p = .020$, Cohen's $d = 0.52$]. Ignoring category type for overt training, the deterministic conditions had an average test accuracy of 68.2%, and the probabilistic conditions had an average test accuracy of 54.1%. However, incidental training did not result in significant differences in generalization across learning deterministic and probabilistic distributions [$t(80) = 0.75$, $p = .46$, Cohen's $d = 0.17$]. Ignoring category type for incidental training, the deterministic conditions had an average test accuracy of 49.4%, and the probabilistic conditions had an average test accuracy of 47.46%. These findings cannot be accounted for directly by the differences in difficulty between the deterministic and probabilistic distributions, because we used normalized accuracies in these analyses. In line with our predictions, overt training led to better learning for deterministic than for probabilistic categories. However, we predicted that probabilistic categories would be learned better during incidental training, but instead we found that deterministic and probabilistic categories were learned equivalently during incidental training.

Our third prediction was that deterministic and probabilistic sampling might not affect the performance of participants learning RB categories, but that deterministic sampling would be learned better than probabilistic sampling for II categories. We found an interaction between category distributions (II, RB) and the distribution sampling for the generalization of category learning [$F(1, 158) = 13.31$, $p < .001$, $\eta_{\mathrm{p}}^2 = .078$]. In support of our prediction, generalization of RB categories did not differ for the deterministic and probabilistic distributions [$t(64.1) = 1.56$, $p = .13$, Cohen's $d = 0.39$, corrected for inequality of variances]. We also note that this was different from what we found during overt training, in which deterministic categories were learned better than probabilistic ones. After either incidental or overt training,

the generalization of RB categories did not differ between deterministic and probabilistic distributions. Also in support of our prediction, for II categories the deterministic distributions resulted in significantly higher generalization accuracy than did the probabilistic distributions [$t(76.3) = 3.00$, $p = .004$, Cohen's $d = 0.69$, corrected for inequality of variances]. The ability to generalize RB categories is not affected by differences in the sampling distributions, but generalization of II categories is worse with probabilistic than with deterministic distributions.

Though overall generalization accuracy can give us some clues as to what participants were able to learn about these categories through overt or incidental training, this does not allow for a full understanding of the category representations that participants learned. To gain a better understanding of these representations, we constructed confusion matrices for each condition, learning II (Fig. 6) or RB (Fig. 7) categories. These confusion matrices demonstrate participants' response behavior in the generalization test on the basis of the actual category that was presented to them. For correct responses, the actual category on a trial (columns) and the categories of participants' responses (rows) converge (Figs. 6 and 7 on the positive diagonal). For incorrect responses, we can observe a clear pattern of confusion among multiple categories or a random confusion across all categories. Confusion matrices allow us to quantify similarities and

differences among categories on the basis of categorization errors during the generalization test.

The pattern of results in the confusion matrices for II categories demonstrates a tendency for the participants in all four conditions to respond in a way that groups categories A and B together and groups categories C and D together (Fig. 6). The confusable categories are not easily distinguished by either dimension used to construct the categories. Instead, this particular pattern of responses is consistent with responses being informed by integration along the positive correlation between the two dimensions. Note that this pattern of responses was similar across II conditions, despite the quantitatively different levels of overall performance.

The pattern of results in the confusion matrices for RB categories demonstrates a different tendency (Fig. 7). The participants in the RB overt deterministic condition were the most consistent in their responses across the four categories. These participants did not demonstrate a clear pattern of confusion among any of the categories in the generalization test, which may have stemmed from their higher accuracy in the generalization test. The participants in the other three RB conditions demonstrated varying levels of confusion between categories B and C. Participants demonstrated clear response patterns that distinguished categories A and D from the other categories, and there
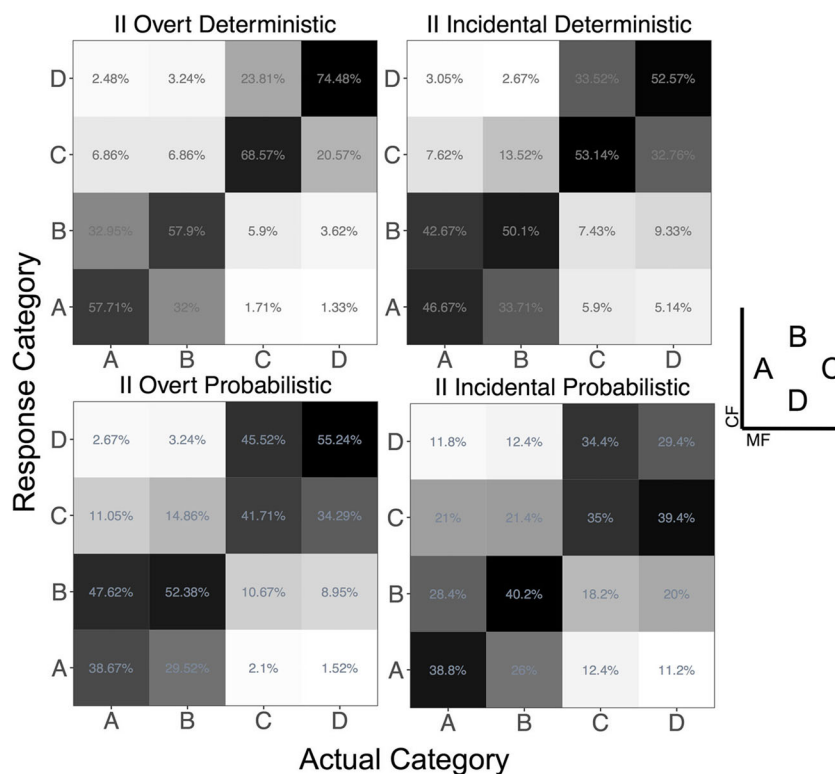

Fig. 6 Confusion matrices for information-integration conditions in the generalization test. Each column represents the actual category identity of the exemplars played on a trial, and each row represents the category response that participants made. The shading gradient and percentages within each cell represent how frequently participants gave a particular response for each category. The cells in each column sum to 100%. To the right is a schematic diagram of the information-integration category structures (also shown in Fig. 2).
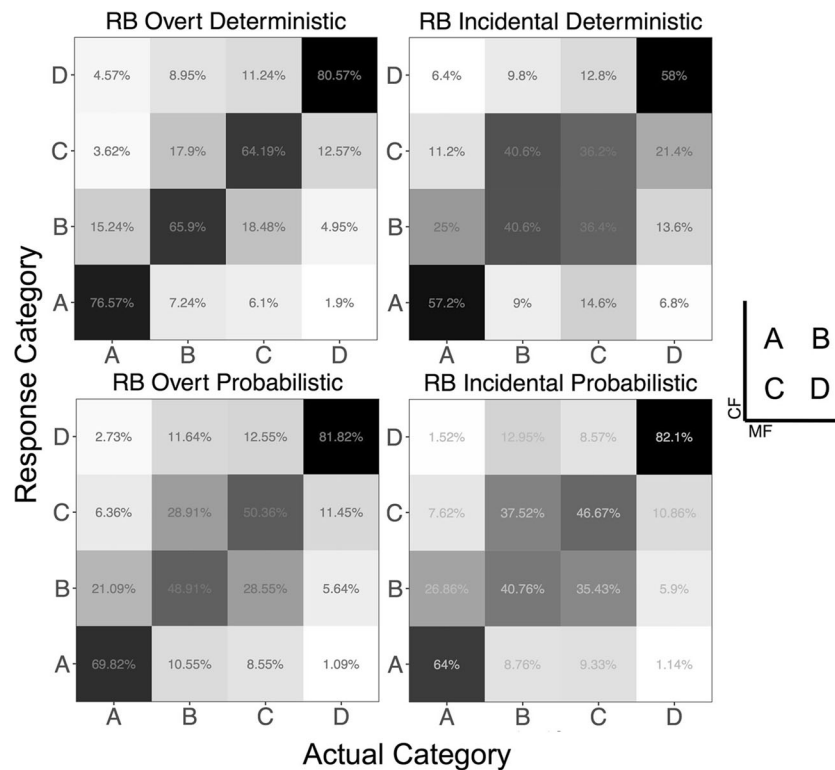
Fig. 7 Confusion matrices for rule-based conditions in the generalization test. Each column represents the actual category identity of the exemplars played on a trial, and each row represents the category response that participants made. The shading gradient and percentages within each cell represent how frequently participants gave a particular response for each category. The cells in each column sum to 100%. To the right is a schematic diagram of the rule-based category structures (also shown in Fig. 2).

was some confusion between categories B and C. Categories B and C differed on both stimulus dimensions used to construct the categories, just as the distinct categories of A and D had. In these three RB conditions, participants did not appear to be using simple unidimensional rules to separate the four categories into two groups. Instead, there is a more complex pattern of responses that even includes confusion of two categories that differ on both acoustic dimensions. The pattern of responses—and particularly of the errors—in the generalization test provides us some information about how participants represented the categories and the relations between the categories as a function of task and distribution sampling.

General discussion

We examined learning and generalization of auditory categories across incidental and overt training tasks, likewise assessing the influence of probabilistic versus deterministic sampling of category distributions defined by a simple rule or requiring integration across dimensions. To our knowledge, this study is the first to compare dual-systems category learning across II and RB stimulus distributions in an incidental training task and an overt training task, as well as the first to systematically examine the effect of the sampling distribution

of the categories and the interaction with training task type. We aimed to understand the extent to which category learning generalizes to novel exemplars, since generalization is a central characteristic of categorization. This further served to provide a common measure across the incidental and overt training tasks, which we will focus on in discussing the results.

Incidental versus overt training

The results demonstrate that artificial nonspeech auditory categories can be learned incidentally under conditions in which participants to not overtly make categorization decisions and are not informed that categories of sound relate to the primary (visual detection) task. Participants were engaged in a simple visual detection task and were not told that the sounds were important or related to the task, that the sounds were drawn from different categories, or that the sounds would later be central in an overt generalization task. The incidental category learning was apparent in overt labeling of novel generalization sounds at posttest, which requires a transfer of incidentally acquired category knowledge to an explicit category-labeling task. Across conditions, incidental training led to successful generalization of category learning across both II and RB stimulus distributions, defined both deterministically and probabilistically. This tells us that explicit awareness of the

relevance of the feedback, or even the goal of the task to learn and generalize category knowledge, is not necessary for category learning. This is notable because prior studies have almost exclusively examined learning with training tasks that involve explicit feedback following each overt categorization decision (Ashby et al., 2002; Chandrasekaran, Yi, & Maddox, 2014; Dunn et al., 2012; Ell et al., 2009; Maddox, Filoteo, et al., 2004; Maddox et al., 2008; for a discussion about unsupervised learning, see Ashby et al., 1999).

The COVIS model emphasizes the importance of feedback in driving learning, particularly in the case of learning II stimulus distributions. In this context, it may seem surprising that there was such robust incidental learning of II stimulus distributions. However, although the incidental training paradigm does not utilize feedback in the traditional manner of overt training tasks, it should not be considered to lack feedback entirely. The consistent correlation of category exemplars with the location of visual targets presents a situation in which auditory categorization supports predictions regarding the primary visual detection task. These predictions are either correct or incorrect, as indicated by the ultimate appearance of the visual target. In this way, categorization is incidentally associated with outcomes via the primary visual detection task. We have argued previously that this form of feedback may relate more closely to how sound categories are used in the world; they allow listeners to use variable sensory input to make predictions that support behavior in the larger environment, which sometimes leads to positive outcomes (Gabay et al., 2015; Lim & Holt, 2011). The present results demonstrate that this alternative, less overt, form of feedback is sufficient to support category acquisition across both RB and II stimulus distributions when they are sampled either probabilistically or deterministically. Even for II distributions, which COVIS posits rely more heavily on feedback, neither overt awareness of the category-learning task nor explicit feedback appears to be necessary for category learning.

This result has important implications for theory. On the basis of the prior literature on visual and auditory category learning and the COVIS model, we predicted that categories defined by II stimulus distributions would be learned better via incidental than via overt training and, conversely, that categories defined by RB stimulus distributions would be learned better under overt than under incidental training. Specifically, since the incidental task was speeded visual detection and not auditory categorization, it directed attention away from overt categorization decisions. Thus, we hypothesized that learning the II stimulus distributions would benefit from incidental training because overt reasoning is thought to hinder II learning (Ashby & Maddox, 2011). The data did not support this prediction; there was no interaction of training task and category stimulus distribution. Both RB and II stimulus distributions were learned better in the overt, relative to the incidental, training task.

Overt training led to better performance than incidental training, regardless of category type. One factor possibly contributing to this finding was that the incidental training task involved a brief block in which the relationship between sound category and visual location was randomized (in order to covertly assess learning online). This short block may have been enough to differentiate the incidental training condition from overt training in its influence on generalization performance. Another possible explanation for the overt training advantage was that the simple visual detection of the SMART incidental training task may not be fully tapping into the procedural learning system that best learns II categories. Therefore, caution is warranted in concluding that learning via overt training is necessarily always superior to learning via incidental training.

Category distribution sampling

The sampling distributions defining the categories impacted learning and generalization performance. This finding is critical, because many speech and nonspeech auditory category-learning studies have used highly stylized, deterministic distributions, whereas natural categories, including speech, are defined by more variable and probabilistic distributions. We predicted, on the basis of the visual category-learning results of Ell and Ashby (2006), that sampling might affect the learning of II, but not of RB, categories. Our generalization test results were consistent with Ell and Ashby's findings that overlap affected the learning of II but not of RB categories. We found poorer category generalization accuracy for II stimulus distributions defined probabilistically than for distributions defined deterministically. In contrast, category generalization accuracy was equivalent across the probabilistic and deterministic RB stimulus distributions. Although our results are consistent with the general premise from Ell and Ashby—that RB category learning is unaffected by differences in overlap, but that II category learning is affected—our finding that generalization was better for deterministic than for probabilistic distributions is inconsistent with their findings. Ell and Ashby found that moderately overlapping categories, such as our probabilistic distributions, led to better II learning than did categories that did not overlap, such as our deterministic distributions. Of course, the sampling manipulation in the present study involved more than just overlap, which may have accounted for the differences between our study and Ell and Ashby's. Additionally, this difference may have been driven by the stimuli themselves. It is possible that simple, verbalizable visual dimensions may be used differently by participants during learning than are the auditory dimensions used in the present study. Further research

will be needed to disentangle the effects of overlap or sampling distribution on auditory II category learning.

Our results provide further evidence of the applicability of COVIS to auditory category learning and the instantiation of the multiple-systems theory for auditory category learning. Although Ell and Ashby (2006) did not test the generalization of learning to novel category exemplars, this finding is in accord with their conclusion that category overlap, one of the differences between our sampling distributions, affects II but not RB category learning.

In future research, it will be necessary to disentangle the potentially interacting effects of the factors defining deterministic and probabilistic category distributions, including overlap, number of exemplars, and stylistically sampled versus randomly sampled distributions. The deterministic distributions that mirror those used in many nonspeech and speech category-learning studies have fewer exemplars and less exemplar overlap, both between and within categories, than did our probabilistic distributions, which were meant to more closely approximate natural category distributions. The differences in learning that were explained by the sampling distributions underscores the significance of this factor in category learning. If our goal is to understand natural category learning, whether visual or auditory, it will be critical to closely approximate the natural structure of those categories in future experimental studies.

These results caution that reliance on simple, carefully designed deterministic input distributions may not capture the learning challenges involved in acquiring speech categories, characterized by highly overlapping distributions across complex and multidimensional input dimensions (Hillenbrand et al., 1995; Swingley, 2009). If we are to generalize the conclusions about II categories and the mechanisms that are used to learn them, we must also carefully consider differences in the distributions that can define different existing real-world speech categories (see Wanrooij & Boersma, 2013, for a similar argument about frequency distributional learning).

Interaction of sampling and task

We predicted that sampling distribution might also interact with training task, such that incidental training might be better across probabilistic distributions, whereas overt training might be better across deterministic distributions. In line with our predictions, overt training led to better generalization of category learning for deterministic than for probabilistic stimulus distributions. However, in contrast to our predictions, incidental training resulted in equivalent generalization of category learning across deterministic and probabilistic stimulus distributions. This interacted with the type of stimulus distribution sampling, as well. For II stimulus distributions, the learning

advantage of overt training over incidental training held for both probabilistic and deterministic distributions. For RB stimulus distributions with deterministic sampling, there was an overt-training advantage. However, this advantage was not apparent for probabilistic RB stimulus distributions. This highlights that important differences in category learning occur with different distribution sampling, training, and category types.

For probabilistic distributions, participants receive information about category boundaries that is inherently less consistent than the feedback given for deterministic distributions. All category exemplars in the deterministic distributions fell perfectly within the hypothetical boundaries within acoustic space defining the respective categories. For the probabilistic categories, a minority of exemplars from each category crossed these hypothetical boundaries, leading to category overlap. This meant that the category-consistent feedback (incidental or overt) that was available in training was not as well-aligned with exemplar similarity in the probabilistic as in the deterministic conditions. The ambiguous nature of the alignment of the feedback signal with acoustic similarity may lead to less clear category representations, especially around category boundaries. Thus, this would lead to a specific benefit of deterministic over probabilistic distributions in the overt task when information was available to explicitly process feedback and incorporate it in future category decisions. According to the hypothesis that feedback given in a deterministic manner depends more on explicit memory systems (Seger & Cincotta, 2005), the poorer alignment of exemplar similarity and feedback that is associated with probabilistic distributions might be less impactful during learning in the incidental task if it draws from learning via more implicit procedural-learning systems. The nature of the category distributions and the complexity of the category sampling are important aspects to consider, because they can greatly impact learning outcomes.

Within-trial variability

A key difference between the present study and previous studies investigating the dual systems of category learning is that we used within-trial variability in our training and testing paradigms. On each trial, participants heard five unique exemplars from within a single category. In typical dual-systems experiments, whether visual or auditory, participants encounter a single exemplar on each trial. This methodology allows experimenters to model the decision-bound strategy response on the basis of how a participant responds to each exemplar.

Previous research with auditory category learning, including speech, has demonstrated an overall benefit in

generalization performance following training with high within-category variability (Bradlow et al., 1997; Iverson, Hazan, & Bannister, 2005; Liu, 2014; Logan et al., 1991). This appears to be particularly potent when within-category variability is aligned with trial-level feedback. Using the same incidental training task as in the present study, Gabay et al. (2015) found superior learning when participants experienced category exemplar variability within a trial, and therefore when such variability was tightly coupled with task-driven predictions and feedback. Participants who experienced the same overall exemplar variability across individual trials in the experiment learned less. However, whereas within-trial variability is likely to have promoted learning and generalization in the present study, it also precluded the use of decision-bound modeling to assess individual participant response strategies during learning. Current iterations of decision-bound models map an individual's decision boundary on the basis of the location of a single exemplar in the stimulus space, given their response. In future work, it will be useful to build decision-bound models that can incorporate within-trial variability.

Strategy use during category learning across within-trial exemplar variability remains an open question for future research. Among many possible strategies, for example, it could be that our participants used only one exemplar out of the five that they experienced on a trial to make their decisions, or that the average similarity space of exemplars experienced within a trial influences decisions. Since distributional sampling had an influence on learning in the present research, it will be informative to direct future research toward understanding how trial-level distributional statistics and longer-term distributional statistics that must be accumulated across an experiment interact to influence category learning.

To take a step in this direction, we examined the patterns of responses in the generalization test in order to obtain a broad sense of participant strategies. Participants' category confusions across the generalization test provide a window through which to approximate the kinds of representations learned. The confusion matrices make clear that similar overall performance in the generalization test can be arrived at via distinct paths. For the II categories, the pattern of confusability implies that listeners tended to group categories in a way that suggests integration across the dimensions, particularly in a positive-going direction. Intriguingly, this same pattern may be evident in the confusion matrices for RB categories. Rather than confusing RB categories distinguished by a single dimension in the stimulus space, listeners tended to make errors consistent with dimension integration across the positive-going dimension correlation. The apparent pattern of reliance on a positive-going integration strategy is consistent with recent results demonstrating a learning advantage for categories defined by a positive-going, as compared to a negative-going, slope in this same stimulus space (Roark & Holt, 2018).

Implications

Although dual-systems theory has been largely developed in the context of empirical data regarding visual category learning, recent work has very successfully applied it to auditory and speech categorization and yielded important insights (Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014; Maddox & Chandrasekaran, 2014). Because the categorization challenges presented by auditory (and speech) signals are somewhat different from those of visual categories (Holt & Lotto, 2010), this also presents the opportunity to examine the first principles of the model in greater detail through the lens of auditory category learning. We view the present research as a necessary bridge between auditory category-learning research that has focused on the representations acquired in category learning and the highly influential COVIS approach that is beginning to influence auditory category-learning research. Our results highlight that small differences in task demands result in quite different patterns of learning that interact with the sampling of category exemplars in acoustic space. Overt categorization decisions and explicit awareness of the category-learning task were not necessary for learning II or RB categories. In the present work, the most effective training approach involved overt training across deterministic category distributions. Since the majority of studies informing the theoretical development have relied on just such category-learning challenges, it will be important to consider that laboratory-based studies may tend to overestimate the ease of category learning under more natural conditions that involve probabilistically defined categories learned across incidental conditions. This is true for both auditory and visual studies. Although we have used a specific pair of acoustic dimensions here, future work should examine the effects of these aspects of the training stimuli on learning with other acoustic dimensions and visual dimensions. The dual-systems approach has not yet investigated the effect of these aspects of distributions and training tasks with either auditory or visual dimensions. Next-generation models of category learning will need to consider the nature of the complexity and overlap of sampling distributions, along with their interaction under more incidental learning situations, in order to better characterize how the system reacts to real-world category-learning challenges.

Appendix: Category distribution means, variances, and covariances

Deterministic category distribution information

Category	Mean (CF, MF)	Variance (CF, MF)	Covariance
II: Category A	(674, 197)	(3,377.5, 1,213.3)	0
II: Category B	(865, 312)	(3,377.5, 1,213.3)	0
II: Category C	(1,056, 197)	(3,377.5, 1,213.3)	0
II: Category D	(865, 82)	(3,377.5, 1,213.3)	0
RB: Category A	(730, 278)	(3,377.5, 1,213.3)	0
RB: Category B	(1,000, 278)	(3,377.5, 1,213.3)	0
RB: Category C	(730, 116)	(3,377.5, 1,213.3)	0
RB: Category D	(1,000, 116)	(3,377.5, 1,213.3)	0

Probabilistic category distribution information

Category	Mean (CF, MF)	Variance (CF, MF)	Covariance
II: Category A	(701.1, 199.4)	(6,306.0, 1,629.8)	− 15.5
II: Category B	(853.1, 310.7)	(7,032.6, 2,281.8)	494.9
II: Category C	(1,059.0, 196.5)	(6,677.8, 2,207.2)	− 36.7
II: Category D	(863.1, 86.8)	(6,156.6, 2,123.3)	119.8
RB: Category A	(735.8, 277.1)	(7,198.3, 2,600.3)	− 201.7
RB: Category B	(922.4, 279.8)	(7,496.4, 2,082.0)	− 804.6
RB: Category C	(749.0, 116.7)	(5,081.2, 2,251.4)	156.3
RB: Category D	(1,007.8, 114.3)	(5,102.4, 1,891.0)	367.6

The probabilistic category distributions were created by defining a two-dimensional Gaussian distribution with the same means and increased variances relative to the deterministic category distributions. Then, 100 random samples from that underlying distribution were taken to form the probabilistic category distributions. Thus, the means, variances, and covariances between categories in the probabilistic category distributions vary relative to the underlying multidimensional Gaussian distributions.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481. doi:https://doi.org/10.1037/0033-295X.105.3.442

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53. doi: https://doi.org/10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0 *Annals of the New York Academy of Sciences*, *1224*, 147–161. doi: https://doi.org/10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*, 666–677. doi:https://doi.org/10.3758/BF03196423

Ashby, F. G., & O'Brien, J. B. (2007). The effects of positive versus negative feedback on information-integration category learning. *Perception & Psychophysics*, *69*, 865–878. doi:https://doi.org/10.3758/BF03193923

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*, 1178–1199. doi:https://doi.org/10.3758/BF03207622

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299–2310. https://doi.org/10.1121/1.418276

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology*, 5, 825:1–17. doi:https://doi.org/10.3389/fpsyg.2014.00825

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, 21, 488–495. doi:https://doi.org/10.3758/s13423-013-0501-5

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

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 840–859. doi:https://doi.org/10.1037/a0027867

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, 68, 1013–1026. doi:https://doi.org/10.3758/BF03193362

Ell, S. W., Ing, A. D., & Maddox, W. T. (2009). Criterial noise effects on rule-based category learning: The impact of delayed feedback. *Attention, Perception, & Psychophysics*, 71, 1263–1275. doi:https://doi.org/10.3758/APP.71.6.1263

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

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, 50, 109–125. doi:https://doi.org/10.1016/j.specom.2007.07.003

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1913–1933. doi:https://doi.org/10.1037/a0015781

Grimm, L. R., & Maddox, W. T. (2013). Differential impact of relevant and irrelevant dimension primes on rule-based and information-integration category learning. *Acta Psychologica*, 144, 530–537. doi:https://doi.org/10.1016/j.actpsy.2013.09.005

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099–3111. doi:https://doi.org/10.1121/1.411872

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, 119, 3059–3071. doi:https://doi.org/10.1121/1.2188377

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science*, 17, 42–46. doi:https://doi.org/10.1111/j.1467-8721.2008.00545.x

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, 72, 1218–1227. doi:https://doi.org/10.3758/APP.72.5.1218

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. *Journal of the Acoustical Society of America*, 118, 3267–3278. doi:https://doi.org/10.1121/1.2062307

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108, 1252–1263. doi:https://doi.org/10.1121/1.1288413

Kluender, K. R., Lotto, A. J., Holt, L. L., & Bloedel, S. L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *Journal of the Acoustical Society of America*, 104, 3568–3582. doi:https://doi.org/10.1121/1.423939

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50, 93–107. doi:https://doi.org/10.3758/BF03212211

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277,* 684–686. https://doi.org/10.1126/science.277.5326.684

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

Lim, S., Lacerda, F., & Holt, L. L. (2015). Discovering functional units in continuous speech. *Journal of Experimental Psychology: Human Perception and Performance*, 41, 1139–1152. doi:https://doi.org/10.1037/xhp0000067

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, 29, 3–11. doi:https://doi.org/10.1177/002383098602900102

Liu, R. (2014). Investigating learning, generalization, and transfer of perceptual representations supporting non-native speech perception (Doctoral dissertation), Carnegie Mellon University, Pittsburgh, PA.

Liu, R., & Holt, L. L. (2011). Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. *Journal of Cognitive Neuroscience*, 23, 683–698. doi:https://doi.org/10.1162/jocn.2009.21392

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. *Journal of the Acoustical Society of America*, 94, 1242–1255. doi: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. *Journal of the Acoustical Society of America*, 89, 874–886. doi:10.1121/1.1894649

Lotto, A. J., Sato, M., & Diehl, Y. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* (pp. C-181–C-186). Cambridge, MA: MIT, Department of Linguistics.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 650–662. doi:https://doi.org/10.1037/0278-7393.29.4.650

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, 32, 582–591. doi:https://doi.org/10.3758/BF03195849

Maddox, W. T., & Chandrasekaran, B. (2014). Tests of a dual-systems model of speech category learning. *Bilingualism: Language and Cognition*, 17, 709–728. doi:https://doi.org/10.1017/S1366728913000783

Maddox, W. T., Chandrasekaran, B., Smayda, K., & Yi, H.-G. (2013). Dual systems of speech category learning across the lifespan. *Psychology and Aging*, 28, 1042–1056. doi:https://doi.org/10.1037/a0034969

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 227–245. doi:https://doi.org/10.1037/0278-7393.30.1.227

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 100–107. doi: https://doi.org/10.1037/0278-7393.31.1.100

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition*, *108*, 578–589. doi:https://doi.org/10.1016/j.cognition.2008.03.010

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219–246. doi:https://doi.org/10.1037/a0022325

Mirman, D., Holt, L. L., & McClelland, J. L. (2004). Categorization and discrimination of nonspeech sounds: differences between steady-state and rapidly changing acoustic cues. *Journal of the Acoustical Society of America*, *116*, 1198–1207. doi:https://doi.org/10.1121/1.1766020

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*, 530–556. doi:https://doi.org/10.3758/s13428-017-0884-8

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184. doi:https://doi.org/10.1121/1.1906875

Roark, C. L., & Holt, L. L. (2018). *Auditory, not acoustic, dimensions impact auditory category learning*. Manuscript submitted for publication.

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience*, *25*, 2941–2951. doi:https://doi.org/10.1523/JNEUROSCI.3401-04.2005

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*, 447–457. doi:https://doi.org/10.1177/0956797613509112

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*, 3617–3632. doi:https://doi.org/10.1098/rstb.2009.0107

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*, 13273–13278. doi:https://doi.org/10.1073/pnas.0705369104

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*, 363–381. doi:https://doi.org/10.1037/a0025014

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America*, *118*, 2618–2633. doi:https://doi.org/10.1121/1.2011156

Wanrooij, K., & Boersma, P. (2013). Distributional training of speech sounds can be done with continuous distributions. *Journal of the Acoustical Society of America*, *133*, EL398–EL404.

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition*, *81*, 283–293. doi:https://doi.org/10.1016/j.bandc.2012.11.006

Yi, H.-G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, *26*, 1409–1420. doi:https://doi.org/10.1093/cercor/bhu236