Attention, Perception, & Psychophysics (2019) 81:958–980
https://doi.org/10.3758/s13414-019-01683-x

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Constraints on learning disjunctive, unidimensional auditory and phonetic categories

Christopher C. Heffner[1,2,3,4] • William J. Idsardi[1,2] • Rochelle S. Newman[1,3]

Published online: 13 February 2019

Abstract

Phonetic categories must be learned, but the processes that allow that learning to unfold are still under debate. The current study investigates constraints on the structure of categories that can be learned and whether these constraints are speech-specific. Category structure constraints are a key difference between theories of category learning, which can roughly be divided into instance-based learning (i.e., exemplar only) and abstractionist learning (i.e., at least partly rule-based or prototype-based) theories. Abstractionist theories can relatively easily accommodate constraints on the structure of categories that can be learned, whereas instance-based theories cannot easily include such constraints. The current study included three groups to investigate these possible constraints as well as their speech specificity: English speakers learning German speech categories, German speakers learning German speech categories, and English speakers learning musical instrument categories, with each group including participants who learned different sets of categories. Both speech groups had greater difficulty learning disjunctive categories (ones that require an "or" statement) than nondisjunctive categories, which suggests that instance-based learning alone is insufficient to explain the learning of the participants learning phonetic categories. This fact was true for both novices (English speakers) and experts (German speakers), which implies that expertise with the materials used cannot explain the patterns observed. However, the same was not true for the musical instrument categories, suggesting a degree of domain-specificity in these constraints that cannot be explained through recourse to expertise alone.

Keywords Category learning · Categorization · Speech perception · Phonetics

Learning a language requires learning phonetic categories. Speech sound tokens vary in their realization from speaker to speaker and from utterance to utterance, making it imperative for listeners to accommodate this variation when understanding speech (Lisker, 1985; McMurray & Jongman, 2011). Despite this variability, listeners readily group speech sounds together using labels that can extend to new instances. This process of categorization has important behavioral consequences. Theories of phonetic learning make different predictions about how these categories are acquired.

In the present set of experiments, two topics of interest are probed. First, we examine the extent to which there are constraints on the types of phonetic categories that are possible to learn. In doing so, we compare instance-based (also known as exemplar-only) theories of phonetic category learning (Hawkins, 2003; Johnson, 2007; Pierrehumbert, 2003) to abstractionist theories, which may take the form of decision-bound (Ashby & Townsend, 1986), prototype (Samuel, 1982), or multiple-system (Chandrasekaran, Koslov, & Maddox, 2014) theories of learning. Although both types of theory have an impressive array of evidence behind them, the focus in this article is on whether the learning process comes with any assumptions about the structure of categories, which the two sets of theories make different predictions about (Ashby & Waldron, 1999). Second, to the extent that such

✉ Christopher C. Heffner
christopher.heffner@uconn.edu

[1] Program in Neuroscience and Cognitive Science, University of Maryland, College Park, MD 20742, USA

[2] Department of Linguistics, University of Maryland, College Park, MD 20742, USA

[3] Department of Speech and Hearing Sciences, University of Maryland, College Park, MD 20742, USA

[4] Present address: Department of Speech, Language, and Hearing Sciences, University of Connecticut, 850 Bolton Rd, U-1085, Storrs, CT 06269, USA

constraints exist, we investigate the domain-specificity of the constraints, comparing phonetic learning with non-speech auditory learning.

We focus on the acquisition of disjunctive categories within a unidimensional stimulus set. Disjunctive categories require "or" statements to describe, as in "A temperature is uncomfortable if it is too hot *or* it is too cold." They are a subset of discontinuous categories, which also include categories that span multiple parts of a category continuum without a different category between those parts of the continuum. These categories exist in a wide variety of real-world contexts. For example, in baseball, a strike is called when the batter hits the ball in foul territory, when the batter swings and fails to hit the ball, or when the batter fails to swing when the ball transverses the strike zone. This is a good example of a disjunctive category in multidimensional stimulus space, as it is challenging to imagine a single dimension along which these three types of actions could be considered continuous. Most speech learning tasks are assumed to include multiple dimensions; say, using patterns in F2 and F3 to characterize the acquisition of the /ɹ/–/l/ distinction by Japanese learners of English (Lotto, Sato, & Diehl, 2004).

For uncomfortable temperatures, on the other hand, the idea of describing these unidimensionally is more plausible. Temperature could be expressed in Celsius or Kelvin, with temperatures above a certain level or below a certain level being labeled under the single category of "uncomfortable." Other examples come from music. In music, identical musical note labels (e.g., "E-flat," "A") are used to describe disjunctive categories spaced throughout the single dimension of pitch. An E-flat is an E-flat no matter which octave it occurs within. Similarly, notes can be perceived as off the musical beat if they occur too fast (if a performer is rushing) or too slow (if a performer is lagging behind), meaning that, across time, there is a span of times perceived as on the beat that are surrounded by notes that are off the beat. Unidimensional, disjunctive categories are seemingly rare in phonetic space. In American English, the category /t/ can be realized as [t] (a voiceless alveolar stop, as in the word *stop*), [t^h] (an aspirated alveolar stop, as in the word *top*), [ ] (an alveolar flap, as in the word *potter*), and even [ ] (a glottal stop, as in the word *button*). Although it is difficult to describe all of these realizations without using a disjunction, it is also likely that these sounds vary across multiple dimensions, not just one. In intonational phonology, pitch accents can be either high (usually annotated H*) or low (L*), but they buttress a set of intermediate fundamental frequency points that are not perceived as pitch accents.

Under one set of theories—here referred to as instance-based models, although often referred to as exemplar or variationist models—listeners do not start with any assumptions about the nature of the categories being learned. Instance-based models see category learning as the result of the encoding of specific instance-to-category-label pairings. Category membership is determined only by the similarity between a new item and previously observed items. Probably the most widely used instance-based theory is the generalized context model (GCM) of Nosofsky (1986). According to the GCM, categorization is essentially a special class of item identification. Categorization requires calculating how closely a new item resembles previously identified ones, using the most similar items to that new item to make a hypothesis about its category label. Indeed, barring an inability to perceptually discriminate individual items, instance-based models can learn almost any category, even ones with very patchy distributions within the stimulus space (Ashby & Alfonso-Reese, 1995; McKinley & Nosofsky, 1995).

One particularly well-cited example of an instance-based theory within speech perception is that of Pierrehumbert (2003). Under Pierrehumbert's (2003) model, speech sound categories are the collection of multiple memorized pairings of individual speech sound tokens (i.e., exemplars) to categories. New items that are fed into the system are simply compared with previously observed ones. The categories that the most similar previous items belong to are compared with one another, and the new item is paired with the category that has the most (and most similar) category connections. The /p/ category, then, is defined by the many specific instances of the /p/ sound that have been encountered on the part of a listener. The model of Pierrehumbert (2003) and its instance-based peers (Johnson, 2007) have explicitly been inspired by instance-based theories of visual category learning, especially the GCM (Nosofsky, 1986) and MINERVA (Hintzman, 1986; Homa, Cross, Cornell, Goldman, & Shwartz, 1973). Under instance-based speech perception theories, even very small phonetic details describing the differences between sounds can be critical, as the recollection of these fine phonetic details may distinguish between categories (Hawkins, 2003; Johnson & Seidl, 2008). This allows for straightforward accommodation of complex aspects of speech perception, especially sensitivity to speaker-specific variation in phonetic cues (Goldinger, 1998; Smith & Hawkins, 2012), as is shown through, for example, speaker-specific studies of perceptual adaptation (Dahan, Drucker, & Scarborough, 2008; Kraljic & Samuel, 2007; Norris, McQueen, & Cutler, 2003).

Abstractionist accounts, on the other hand, can more readily accommodate learner assumptions or prior beliefs about the structure of phonetic categories. Abstractionist accounts include decision-bound theories, prototype theories, and multiple-system theories, all of which include a layer of abstraction above and beyond the level of individual item-to-label mappings. The types of abstraction that are used within each model vary widely. Under decision-bound models, learners determine an abstract ideal boundary in perceptual space to delineate multiple categories. The boundaries need not necessarily be linear, although generally under decision-bound models the boundaries proposed are subject to processing constraints that discourage overly complex boundaries (Ashby & Gott, 1988; Ashby & Townsend, 1986; Maddox, Molis, & Diehl, 2002). For example, a stop in English might be classified as voiced if it has a voice onset time (VOT) of 35 ms or smaller, and voiceless if it falls above that boundary. Prototype models store categories as either a single prototype or set of multiple prototypes (Samuel, 1982). More recent formulations of prototype theories involve each speech category being formed from a mixture of Gaussian distributions (McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010) that abstract away from specific instances (although such mixture models inherently involve disjunctions, as categories are described as falling into one of many possible distributions).

One approach to abstractionist accounts of learning that has been gaining steam has been to propose the use of multiple systems in category learning. Motivated in part by multiple system accounts of memory (Squire, 2009), multiple-system accounts of category learning propose that listeners make use of both instance-based and rule-based systems. Under RULEX (RULes and EXceptions; Nosofsky & Palmeri, 1998; Nosofsky, Palmeri, & McKinley, 1994), learners first attempt to sort items into categories according to simple, linear rules, then attempt successively more complex rules until finally falling back on simple memorization of exceptions. Another dual-system model, COVIS (COmpetition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken, & Waldron, 1998), combines a familiar rule-based system with a second decision-bound system, albeit one that largely replicates instance-based learning. Dual-system models have been proposed at a variety of levels of analysis. The acquisition of morphosyntax (Ullman, 2004, 2016), lexical items (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010), and phonetic categories (Chandrasekaran, Koslov, et al., 2014) have all been approached using multiple-system models that, by and large, rely on one rule-like system and one memorization-like system. This path has much in common with more recent approaches to phonetic category learning that incorporate neurobiological insights (Myers, 2014). It also provides a way to comfortably incorporate the impressive pool of evidence for the idea that listeners are acutely sensitive to fine phonetic detail in speech (Bybee, 2002; Hawkins, 2003; Hay, Nolan, & Drager, 2006; Pierrehumbert, 2002) with findings that seem to require a level of abstraction in phonetic processing (Pajak & Levy, 2014; Pycha, 2009, 2010).

Other approaches to finding a middle ground between instance-based and rule-based models do not rely on multiple systems. SUSTAIN, short for Supervised and Unsupervised STratified Adaptive Incremental Network (Love, Medin, & Gureckis, 2004) is one such example. Like single-system models, SUSTAIN does not explicitly represent two different pathways to learn categories, but the single system that is postulated forms "clusters" of stimuli that have similar category properties, resembling mixture models. When few clusters exist, the model's behavior is said to take the form of a rule-based model. This model behavior might be seen in cases when categories are simple and easy to describe and, thus, when few stimulus clusters need to be posited. Characterizing a pitch as "low" or "high" is a good example of a category learning scenario that would require few clusters. If additional clusters are necessary, though, the model behaves more like an instance-based model, storing new clusters to accommodate the unusual exceptions in a way that resembles instance-based computation. The example of musical notes given earlier is a good example of a category learning scenario that might require many stimulus clusters; each instance of, say, B-flat would give rise to a cluster that would be assigned to the B-flat category.

Although instance-based and abstractionist theories can be described in mathematically interchangeable terms (Ashby & Maddox, 1993; Rosseel, 2002), we focus here on a key difference between these theories: the possibility of constraints on the structure of categories. The relevance of the structure of categories to the speed of learning has been a topic of interest from virtually the beginning of psychological studies of categorization. The classic study performed by Shepard, Hovland, and Jenkins (1961), for example, examined the acquisition of categories of simple geometric objects that varied in their size, shape, and shade. They found that more complex categories (i.e., categories that combined objects of disparate sizes, shapes, and shades) were harder to learn. Under abstractionist theories, these patterns are usually explained in terms of the complexity of the rules or the prototype structures that are necessary to describe the complex categories.

Under instance-based theories, the fact that complex categories are harder to learn results from interstimulus confusability or selective attention across dimensions (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). To explore why, consider the generalized context model (GCM) of Nosofsky (1986). The computational implementation of the GCM is fairly simple. The distance between a current item and previous ones is computed using a Gaussian distance function. This distance function is used to compute the weight that each item has toward categorizing the stimulus into any of the possible categories under consideration. The weighting is dependent on how confusable an item is with its neighbors. The new item is assigned to the category with the greatest summed weight. In this case, neither the exact labels chosen nor whether those labels are repeated across the stimulus space affect categorization, as the category labels themselves are only used as labels for items.

However, when participants can easily discriminate individual tokens, instance-based theories allow for almost limitless flexibility in the end point of learning. As previous authors have testified, “all [tested] exemplar models predict that with enough training, subjects will respond almost optimally in *any* categorization task, no matter how complex” (Ashby & Alfonso-Reese, 1995, p. 227). For the GCM in particular, the model “basically predicts that given enough experience with training exemplars, participants’ response patterns should eventually approximate the underlying category distributions” (McKinley & Nosofsky, 1995, p. 145), with the trajectory of learning only hampered by interinstance discriminability. One way to test this idea is to provide multiple, well-separated pockets of instances. Instance-based theories have a hard time explaining differences that result from category structure when the stimuli fall along a single dimension with easily differentiable items. When instances are differentiable, the structure of the categories being learned should not affect the rate of learning; essentially any assignment of instances to categories should be relatively easy. Instance-based theories predict that an almost limitless number of categories could be learned. When instances are confusable, instance-based models predict that learners should find categories challenging to learn. However, rather than responding with specific patterns, participants should behave approximately according to chance, responding in line with a broad sample of items from the stimulus set.

Previous studies along similar lines have shown mixed patterns. Kingston (2003) examined the ability of English speakers to learn to contrast sets of German vowels that differed in vowel height, rounding, and tenseness. Although the English speakers were affected by exemplar-like effects when learning to categorize by tenseness (e.g., by showing better learning when more vowel contrasts were available in training), no such effects were found for vowel height and rounding, where an abstractionist theory of learning seems to be a better match for the results obtained. Intriguingly, when such patterns were tested in the phonotactic domain, examining learners’ abilities to pick up on regularities of phonological segment coincidence within words, learners did not find more complex categories monotonically more challenging to learn (Moreton, Pater, & Pertsova, 2017). This finding challenges some of the predictions of simple abstractionist theories within the speech learning literature. The results were replicated in the visual domain for participants learning to categorize varieties of cake. However, both studies used classes of segments that varied in a binary (either–or) fashion across multiple dimensions and, in the case of Moreton et al. (2017), relied on matching or mismatching sets of segments across a word rather than single segment categories. Examining multidimensional categories complicates the predictions of both abstractionist and instance-based theories. Further, categories in phonetic space (such as voicing categories) tend not to be discrete, instead relying on the categorization of instances in continuous space.

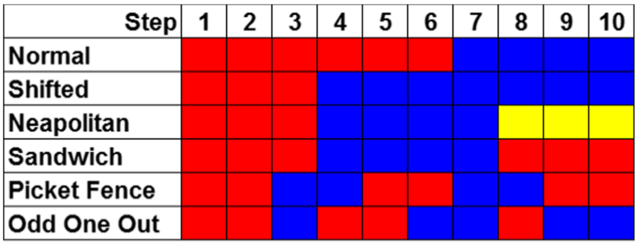
A second question of interest is whether the processes that underpin phonetic learning, and the constraints that might make some aspects of learning more challenging, are shared between speech and nonspeech domains. For most strictly instance-based accounts of phonetic category learning, the process of learning itself is no different from learning any other auditory object. The lack of constraints on the category structures that can be learned should be identical in speech and category learning elsewhere (Port, 2007, 2010). If constraints are uncovered, on the other hand, it is an open question whether these constraints are domain-specific. It could be the case that any constraints on phonetic learning are also found in the auditory modality more generally. Many erstwhile speech-specific properties have been found with other auditory objects (Diehl, Lotto, & Holt, 2004; Holt & Lotto, 2008), and many of the properties of phonetic categories can be explained with recourse to audition-general constraints alone (Diehl, 2000, 2008; Holt, Lotto, & Diehl, 2004). Yet speech must somehow be different from the rest; after all, speech is used as an input to broader language systems, such as syntax and semantics. The sound of jangling keys cannot become a part of a syntactic phrase (Poeppel, Idsardi, & van Wassenhove, 2008). For phonetic learning, the massive amount of experience with speech categories or even innate predispositions might lead learners to make different assumptions about the structure of categories within an unknown phonetic space. Alternatively, dealing with the likely very warped perceptual space in which phonetic categories are learned may require domain-specific dimensional processing. This makes experience a key component of the study of domain-specificity of constraints on phonetic learning.

Many studies examining the acquisition of disjunctive categories outside of the phonetics literature have focused on categories that are disjunctive across *multiple* dimensions. Although this does accurately reflect many kinds of disjunctive categories, this leaves open the question of what learners will do with categories that are disjunctive within a single dimension. Abstractionist and instance-based theories of category learning make predictions for unidimensional categories as well as multidimensional ones. Making predictions within unidimensional category spaces is simpler than doing so for multidimensional categories.

Four studies were used to test claims about the learnability of different category structures. Experiment 1 includes a test of the dimensionality and comparability of speech and nonspeech continua. Experiment 2 includes a set of three subcomponents, assessing either different populations or different stimuli. In Experiment 2a, we examined whether there are constraints on the acquisition of disjunctive speech sound categories, as assessed with English speakers learning categories of German speech sounds. In Experiment 2b, we tested whether these constraints would also apply to German speakers, experts with this dimensional space, who were learning to categorize sounds within the same set of items. And, finally, in Experiment 2c, we determined whether the constraints would also appear for a set of nonspeech sounds, using a synthetic musical instrument continuum. In each case, participants heard items from a 10-point continuum, with different points on the continuum associated with various colored squares (as category labels). Each participant completed one of six different learning conditions, with the categories being learned changing from condition to condition based on which sounds are paired with which squares.

In Experiment 2, six different intersubject conditions were used to probe the influence of category structure on learning (see Fig. 1). Two of them, the Normal and Shifted conditions, were easy to learn under every theory of category learning. Two of the conditions, meanwhile, were predicted to be much more challenging for participants to learn: the Odd One Out and Picket Fence conditions. Both conditions included a large number of disjunctions within the stimulus continuum. Although they both could theoretically be learned, given enough exposure, by a precise instance-based theory, interitem confusability would likely doom an instance-based model in practice. Including both hard and easy conditions allows for the calibration of the relative difficulty of conditions that should be intermediate in difficulty between the two sets.

The key conditions for distinguishing between abstractionist and instance-based learning accounts were the remaining two, the Neapolitan and Sandwich conditions. Both conditions involved two category boundaries along the continua, in the same locations, but the Sandwich condition included a disjunctive category, whereas the Neapolitan condition did not. Here, the theories make divergent predictions. Under instance-based theories of category learning, both categories should have an equivalent difficulty: if the Neapolitan condition is challenging, the Sandwich condition should also be challenging. As previously mentioned, instance-based theories of category learning are very flexible, and the difficulty of categorization depends on interitem similarity. Because the Neapolitan and Sandwich conditions include equally confusable items and equally confusable boundaries, they should be equally easy to learn. No matter where a novel item fell within the speech sound continuum, the distances to adjacent and nonadjacent categories were the same across the conditions. Thus, it should have been equally difficult to identify individual items across the conditions, because the instances being sampled across the two conditions would be approximately identical; the only difference would be in the label of some of the tokens in that sample.



Step	1	2	3	4	5	6	7	8	9	10
Normal	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦
Shifted	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Neapolitan	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟨	🟨	🟨
Sandwich	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟥	🟥	🟥
Picket Fence	🟥	🟥	🟦	🟦	🟥	🟥	🟦	🟦	🟥	🟥
Odd One Out	🟥	🟥	🟦	🟥	🟥	🟦	🟦	🟥	🟦	🟦

Fig. 1 Six conditions used in each component of Experiment 2. Each column shows one of 10 steps in the unidimensional continuum, and each row shows a condition. The cells are colored according to the assignment of step to category in each condition. (Color figure online)

The behavior of abstractionist models, meanwhile, depends on the treatment of the disjunctive red category. Many proponents of dual-system models, for example, have suggested that disjunctive categories may sometimes be learned using the rule-based learning system, rather than the instance-based one (Minda, Desroches, & Church, 2008; Zeithamova & Maddox, 2006), including in speech sound categories (Maddox et al., 2014). However, such ideas have generally been based on multidimensional stimuli, such as visual stimuli that depend on both shape and color, rather than on unidimensional stimuli more like the ones encountered in the present experiment. If both the Sandwich and Neapolitan conditions are processed using identical systems, they should both be equally easy to learn.

Other abstractionist approaches would suggest that the disjunctive, unidimensional category in the Sandwich condition should make it harder to learn than the nondisjunctive ones of the Neapolitan conditions. In the Rational Rules model of concept learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), hypotheses take the form of rules. In learning scenarios that include nondisjunctive categories along a single dimension, these rules are formed from conjunctions or disjunctions of sets that describe parts of a dimension. Participants make responses in line with the small number of hypotheses that they are entertaining at any one point about the categories that they learn, with a small probability of responding incorrectly. Individual items also have the chance of being labeled as an outlier if they belong to a category unexpected by the rules currently under consideration. Simple rules are preferred to more complicated ones due to a strong prior for simple rules. Under the Rational Rules model, participants have strong priors toward simple categories. If listeners find the disjunctive Sandwich condition more difficult than the nondisjunctive Neapolitan condition, this would provide evidence for one of these types of abstractionist theories.

Comparing the subcomponents of Experiment 2 (speech stimuli in Experiments 2a and 2b vs. musical instrument stimuli in Experiment 2c, and English speakers in Experiments 2a and 2c vs. German speakers in Experiment 2b as participants) will allow for a comparison of the effects of expertise to the effects of the materials being used. Both groups of English speakers were equally unfamiliar with the stimuli being used, regardless of whether those stimuli were speech based or instrument based, when compared with the German speakers. Thus, any biases shared by the English speakers but not by the German speakers may reflect the influence of expertise (or the lack thereof) on category learning. Conversely, both the English and German speakers learning phonetic categories were learning sounds taken from the speech domain. This implies that any bias shared by the English and German speakers learning phonetic categories, but not shared by the English speakers learning instrument categories, may reflect the influence of speech-specificity.

Experiment 1: Stimulus properties

Method

Before discussing the acquisition of the auditory categories used in this project, the perceived properties of the stimuli that were being used had to be established. After all, any differences that would be found between the acquisition of phonetic and instrument categories could either be the result of differences in the processing of items inside and outside of language or simply due to differences in the discriminability or dimensionality of the stimuli. Two continua were created through weighted averaging: a speech continuum, used in Experiment 2a and Experiment 2b, and a musical instrument continuum, used in Experiment 2c. It was believed that both continua would be perceived in a unidimensional manner, showing stepwise increases in discriminability between stimuli of successively larger intervals along the continuum. To test this idea, and to determine the extent to which two continua were well matched, participants performed a simple discrimination task to determine the distinctiveness of the stimuli.

Participants

Twenty-seven participants were recruited from Amazon's Mechanical Turk crowdsourcing database. One participant was removed from analysis due to previous experience with German, leaving 26 native English speakers (seven female, 19 male). No participants were old enough that significant high-frequency hearing loss would be expected (M_{age} = 34.7 years, range: 25–47 years). Although participants were asked to use headphones, three participants reported using external or built-in speakers. Despite uncertainty about the precise qualities of the sound equipment that the participants used, previous studies using Mechanical Turk (Buxó-Lugo & Watson, 2016; Slote & Strand, 2016) have generally found Mechanical Turk to be an appropriate venue to run speech perception experiments.

Materials

To create the phonetic stimuli, materials from a previous study (Key, 2014) were used as a starting point. The [x] and [ç] end points of the palatal-to-velar continuum were excised from tokens produced by a native speaker of German, selected from a variety of recordings of [ç] and [x] in nonword frames. The now-isolated tokens, each 95-ms long, were cut at zero crossings, with the longer token cut in size to match the length of the shorter token, and the peak intensities of each file were scaled to an identical 0.9 Pa. The spectral content of these natural tokens was then linearly combined using Praat (Boersma & Weenink, 2001) to create a 10-step continuum, with intermediate points that entailed a linear combination of the acoustic noise that characterizes each fricative. Each intermediate step was therefore a weighted average of the energy found in each end point. The steps were numbered, with Step 1 defined as the most palatal item and Step 10 as the most velar item, with each intermediate number indicating the precise titration of the two end points.

To examine the acquisition of rich and acoustically complex nonlinguistic categories, we created a continuum of synthetic musical instrument sounds. This was done using the Wind Instruments Synthesis Toolbox or WIST (Rocamora, López, & Jure, 2009) and Praat (Boersma & Weenink, 2001). The WIST was used to create two 500-ms musical instrument notes, one synthesized from a trumpet template and one synthesized from a trombone template. Both notes were synthesized with identical fundamental frequencies and identical intensity properties; the only thing distinguishing the two notes was their timbre. The instrumental tokens were a great deal longer than the fricative stimuli due to the properties of the WIST. However, such differences would likely only impact the learning of each class of item insofar as the instrument items were differentiable to a different extent from the fricative items.

Next, the notes were spectrally rotated around a 4 kHz midpoint, a type of acoustic manipulation that redistributes information across frequencies in an acoustic signal. This spectral rotation was used to construct synthetic musical instruments that have much of the rich acoustic signature of brass instruments, but without a true connection to the instruments. The trumpet and trombone sounds were low-pass filtered to remove information above 8000 Hz, then spectrally rotated using two channels (split at 4000 Hz) to create two end points for the musical instrument continuum. That is, the intensity and spectral information found in the signal was mirrored around 4000 Hz, with, for example, points of relative prominence at 3500 Hz now being reflected in points of relative prominence at 4500 Hz. The long-term average spectrum of the original sound was then overlaid on the spectrally rotated sounds. This preserves the overall acoustic profile of the original brass sounds while putting a new spin on the relative prominence of different frequencies within the signal. This renders them analogous to the German fricative stimuli: acoustically complex and clearly instrumental, but unfamiliar. The end points of this continuum were labeled the "pettrum" and the "bonetrom," respectively, and were peak scaled to ensure their intensities matched. Next, Praat was used to linearly combine the two end points to make a 10-step continuum. As with the speech stimuli, this was accomplished through use of spectral blending: each point along the continuum represented a linear combination of the two end-point signals. The pettrum end was arbitrarily labeled Step 1, while the bonetrom end was labeled Step 10.

In using unidimensional, flat continua, a level of validity was sacrificed. Flat distributions with perfectly covarying cues are not typical for acoustic categories, particularly ones with only 10 items. Studies of cue trading in phonetics, for example, have shown that many, if not all, phonetic contrasts are signaled with a wide variety of cues, all capable of combining together in many different ways to yield a coherent percept (Repp, 1982). The rich trade-offs between these cues were not available in the present data set. In this case, weighted averaging means that whatever multiple cues that listeners use to perceive the differences between the end points are completely and inextricably correlated. This continuum therefore provides an avenue to measure the perception and acquisition of simple auditory categories, akin to unidimensional voice onset time (VOT) continua used to examine the perception of word-initial voicing. Spectral slices of the midpoint of each stimulus are available in Fig. 2.

Procedure

Participants heard two blocks of trials: one with the speech stimuli, the other with the musical instrument stimuli. The order of the blocks was counterbalanced across participants. Within each trial, participants heard two paired stimuli from one of the continua, back-to-back, with a 500-ms interstimulus interval (ISI). With 10 possible stimuli as both the first and second item, there were 100 possible ordered pairs per continuum. Participants heard all 100 pairs exactly once and were then asked to rank how similar the items within the pair were on a scale from 1 to 9.

Analysis

The similarity judgments for each participant were converted into difference scores, ranging from 0 (*not different*) to 8 (*most different*). These difference scores were used to create a 10 × 10 symmetric data matrix for each participant, with each row and each column being a step within the continuum. These symmetric data matrices were analyzed using the IDIOSCAL (Individual Differences in Orientation Scaling) functionality of the "smacof" package within R (Mair, De Leeuw, Borg, & Groenen, 2016). IDIOSCAL is a generalization of Individual Differences Scaling, INDSCAL (Carroll & Chang, 1970), which has been used extensively in the category learning literature; for example, in determining naïve listeners' parcellation of Mandarin tone categories (Chandrasekaran, Sampath, & Wong, 2010) or to examine the effects of training on categorical perception (Livingston, Andrews, & Harnad, 1998). In INDSCAL and IDIOSCAL, dimensionality analysis requires multiple possible dimensionalities, n. For each dimensionality, an $n \times 10$ matrix is generated, showing the coordinates of each stimulus step in an n-dimensional space. Traditionally, the approach to determine the best number of effective dimensions is to calculate badness-of-fit measures for each n and to look for an "elbow," a point at which additional possible dimensions do not lead to appreciable drops in badness ratings.

Results and discussion

Participants by and large perceived both continua as unidimensional. Table 1 shows averaged difference scores for each pair of items.

Figure 3 shows a scree plot with badness-of-fit values across different possible dimensionalities. Higher stress values indicate larger badness of fit. The lines do not show a clear "elbow"; badness of fit decreases gradually across the possible dimensionalities for both continua. Although the largest numeric difference across dimensionalities occurs between one and two dimensions (0.049 for the fricatives, 0.076 for the instruments), that difference is not particularly large nor much bigger than the next largest difference, between two and three dimensions (0.030 for the fricatives, 0.033 for the instruments). There does not appear to be a reason to reject the unidimensional interpretation of the continua.

To the extent that the stimulus similarity ratings did not conform to a unidimensional distribution, participants generally found the end points of the continuum to be more similar to each other than would be expected given a uniform progression from most to least similar items. This is an interesting contrast to the anchor effects found in other domains, where, for example, studies of intensity discrimination have shown that intensities closer to the end points of a continuum are more easily discriminable than those in the middle (Braida et al., 1984). This can be seen in Fig. 4, which shows the one-dimensional and two-dimensional IDIOSCAL solutions. Although the scree plots of Fig. 3 do not provide conclusive evidence that the unidimensional interpretation of this continuum should be rejected, it is important to note that Dimension 2 in Fig. 4b is scaled to about half the width of Dimension 1. Although the interpretation of absolute differences between the end points of single dimensions is not entirely one to one with estimates of variability explained by that dimension, this indicates that Dimension 2 may be doing important work in the two-dimensional solution obtained in this study.

The dimensions revealed in Fig. 4 are roughly comparable, with slightly lower interitem discriminability on the extreme ends of the continuum for the instrument continuum compared with the fricative continuum. In Fig. 4a, this corresponds to the one-dimensional solution showing bunched-up items on the bonetrom end of the continuum for the instruments to an extent not found in the fricatives. Some portion of these differences between the stimuli may be related to individual differences in the familiarity of the items to the listeners. Although we did not explicitly ask about musical training as a part of our demographic survey, differences in experiences with musical stimuli may have led to differences in the perceived properties of the stimuli.

In Fig. 4b, one possible interpretation of the two dimensions uncovered roughly corresponds to the position in the stimulus along the continuum (Dimension 1) and to whether

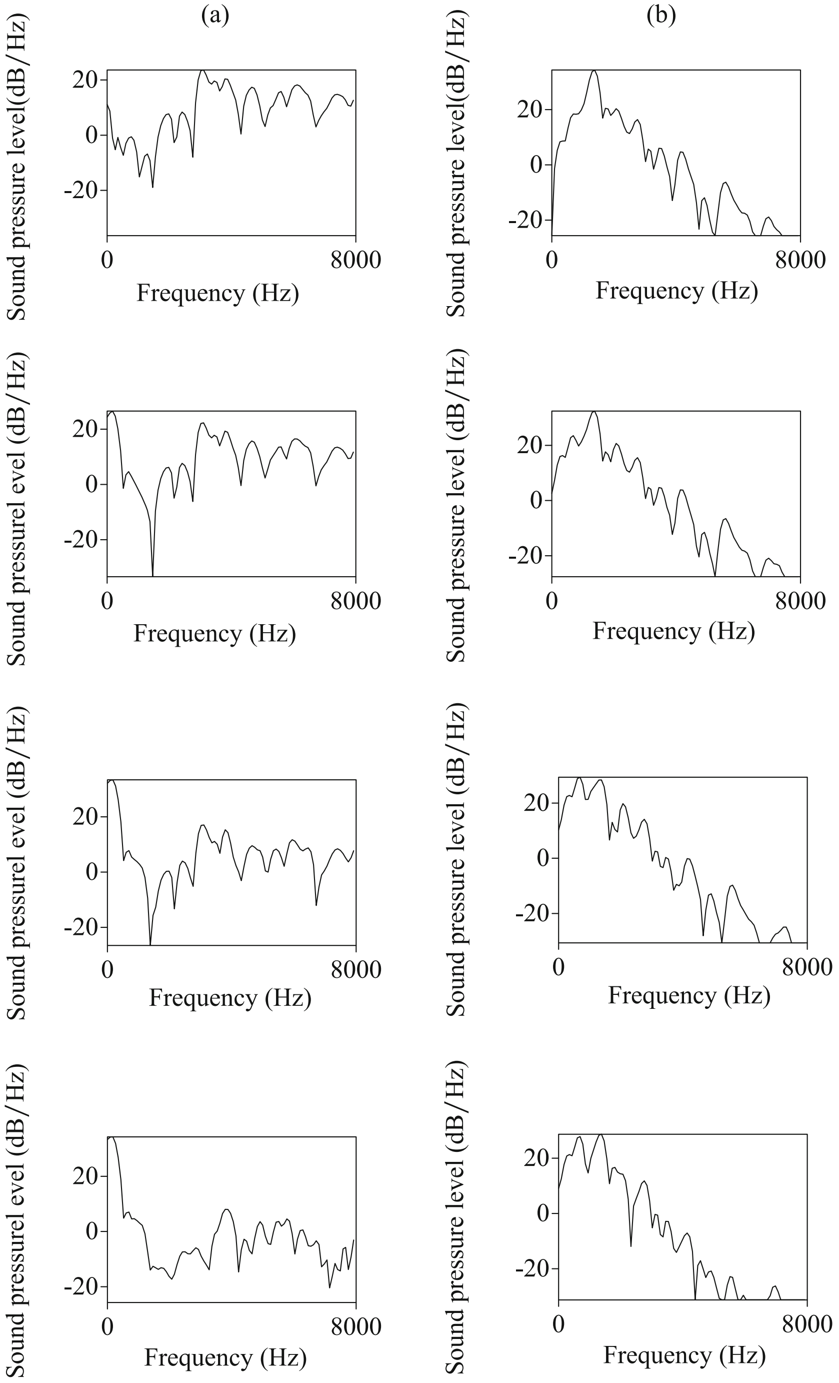


Fig. 2 Spectral slices showing the energy found at various frequencies at the midpoint of Steps 1, 4, 7, and 10 within the fricative continuum (**a**) and instrument continuum (**b**). All displayed spectral slices show a range from 0 to 8000 Hz with a window length of 5 ms

the stimuli are extreme members of the continuum or fall somewhere in the middle (Dimension 2). Although IDIOSCAL is naïve to the true nature of the putative dimensions uncovered, Dimension 1 could be said to show how English-speaking participants sorted the items into categories, whereas Dimension 2 could show the level of certainty that the participants had in that label (with higher values indicating increasing certainty). Regardless of the interpretation of the dimensions, the two-dimensional IDIOSCAL solution shows the “extremely palatal” or “extremely pettrum” items (Steps 1–3) and the “extremely velar” or “extremely bonetrom” items (Steps 8–10) as less distant from each other than one would expect based on stimulus step alone. In general, the items are classified similarly across the sets of stimuli, with roughly equal distances from step to step across the two continua. This suggests that comparing the two continua is appropriate.

Table 1 Average difference score for each pair of stimuli, averaged across participants and orders, for the fricative and instrument items

			Pair Member 2									
			1	2	3	4	5	6	7	8	9	10
Fricatives	Pair Member 1	1	0.25 (0.51)									
		2	0.93 (1.03)	0.36 (0.54)								
		3	1.3 (1.12)	0.84 (0.9)	0.39 (0.73)							
		4	1.83 (0.95)	1.5 (0.83)	1.05 (1.28)	0.43 (0.47)						
		5	2.92 (1.14)	2.5 (1.24)	1.88 (1.04)	1.1 (0.91)	0.47 (0.56)					
		6	4 (1.52)	3.74 (1.35)	2.94 (1.42)	2.19 (1.19)	1.61 (1.17)	0.62 (0.81)				
		7	5.14 (1.57)	4.43 (1.62)	4.28 (1.58)	3.51 (1.47)	2.67 (1.67)	1.66 (1.27)	0.48 (0.52)			
		8	5.93 (1.64)	5.76 (1.53)	5.12 (1.72)	4.75 (1.59)	3.74 (1.44)	2.83 (1.39)	1.26 (1.04)	0.34 (0.48)		
		9	6.2 (1.59)	6.13 (1.53)	5.8 (1.62)	5.32 (1.61)	4.6 (1.67)	3.71 (1.42)	2.61 (1.54)	1.4 (1.09)	0.3 (0.42)	
		10	6.07 (1.73)	6.13 (1.62)	5.79 (1.69)	5.18 (1.66)	4.89 (1.67)	4.05 (1.82)	2.91 (1.42)	1.33 (0.94)	0.81 (1.24)	0.31 (0.47)
Instruments	Pair Member 1	1	0.31 (0.35)									
		2	0.87 (1.04)	0.41 (0.58)								
		3	1.39 (1.19)	1.21 (0.76)	0.52 (0.68)							
		4	2.2 (1.1)	1.9 (1.11)	1.32 (0.86)	0.53 (0.6)						
		5	3.55 (1.27)	3.14 (1.26)	2.73 (1.19)	1.73 (0.93)	0.61 (0.71)					
		6	4.81 (1.73)	4.73 (1.6)	3.73 (1.52)	3.05 (1.35)	1.76 (1.11)	0.46 (0.48)				
		7	5.25 (1.87)	4.97 (1.67)	4.58 (1.73)	3.68 (1.34)	2.52 (1.33)	1.29 (0.83)	0.59 (0.68)			
		8	5.33 (1.9)	5.3 (2.04)	4.97 (1.86)	4.28 (1.54)	3.27 (1.46)	2.07 (1.08)	1.08 (0.77)	0.44 (0.43)		
		9	5.42 (1.91)	5.43 (1.67)	4.65 (1.67)	4.59 (1.58)	3.7 (1.4)	2.62 (1.2)	1.62 (1.03)	1.19 (1.28)	0.37 (0.39)	
		10	5.15 (2.08)	5.19 (1.75)	4.99 (1.74)	4.49 (1.64)	3.69 (1.66)	2.89 (1.22)	2.04 (1.4)	1.17 (1.15)	0.97 (1.08)	0.4 (0.39)

Note. Larger numbers are pairs that were judged to be less similar. The numbers in parentheses within each cell are standard deviations

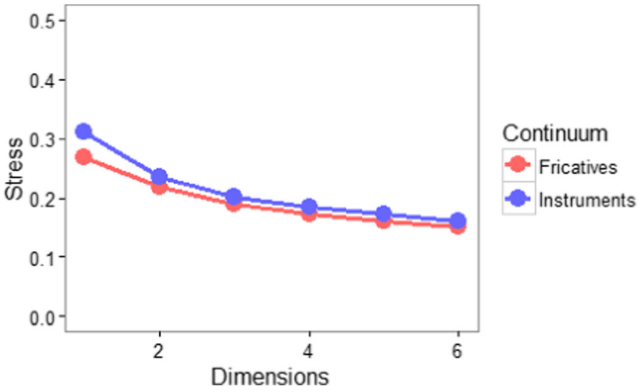


Fig. 3 Badness-of-fit values in Experiment 1, across different numbers of considered dimensions (x-axis) for each continuum (color). Higher stress values indicate a worse model fit. (Color figure online)

Experiment 2: Auditory category learning

Method

With the relevant characteristics of the stimuli established, three experiments were carried out. In Experiment 2a, American English speakers were trained to learn perceptual categories of German fricatives. In Experiment 2b, native German speakers were trained on these same German fricative categories. And in Experiment 2c, native English speakers were trained on categories of instrument sounds. Because these studies often involve parallel analyses, they are discussed in tandem.

Participants

Experiment 2a Sixty-eight participants were recruited at the University of Maryland, College Park. Participants were compensated either with class credit in introductory linguistics or hearing and speech sciences courses, or with financial compensation. Data points were excluded from three participants who had accrued more than incidental exposure to the German language, either through formal training or by living in a German-speaking country for at least a month; from one participant who was missing a demographics sheet; from six who were out of the target age range; and from one whose data file was corrupted. The participants remaining (n = 57) came from a typical undergraduate population (M_{age} = 20.2 years, range: 18–27 years, 34 female, 17 male, six not stated). All participants self-reported normal hearing and no history of speech or language disorder. Many participants had studied languages with a voiceless velar fricative in their phonological inventories (e.g., Spanish), but none of these languages had both velar and palatal fricatives.

Experiment 2b Sixty-three participants were recruited at the University of Tübingen to perform this experiment. Participants were given €5 as payment for their participation in the task. They were recruited from linguistics-related LISTSERVs on campus or from previous participation in experiments within the linguistics department at the University of Tübingen. Two participants were excluded due to technical issues during the experiment, leaving a total of 61 participants. Of the 61 participants remaining—all young adults, between

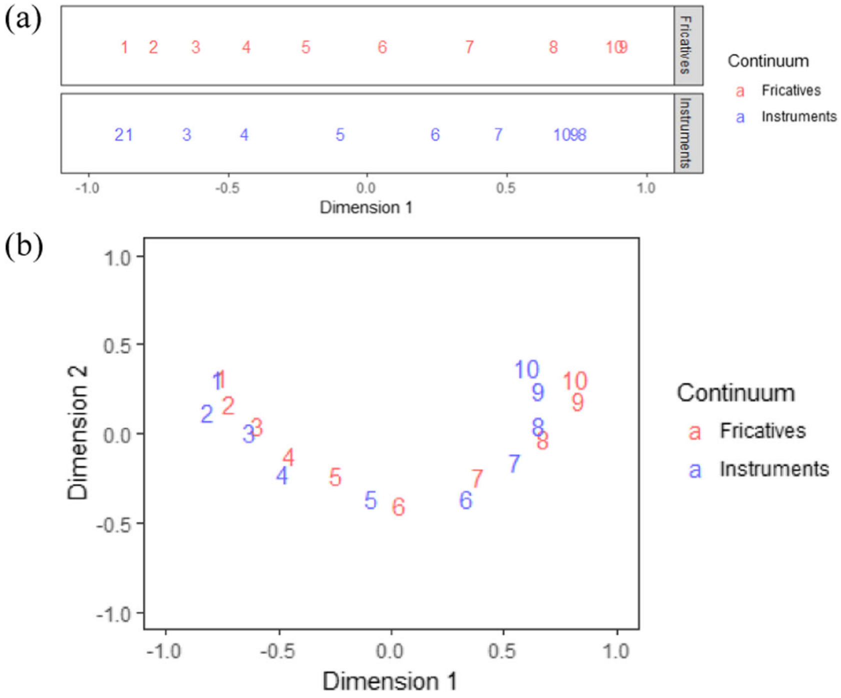


Fig. 4 Two-dimensional IDIOSCAL solution, Experiment 1, for the dimensions in the (**a**) one-category and (**b**) two-category fits. (Color figure online)

the ages of 19 and 34 years (M_{age} = 23.7 years)—nine were male, 51 were female, and one did not indicate a gender. All participants gave informed consent, which was conducted in German. Experiments were performed in line with German ethics standards, which do not require explicit ethics panel review for language-related experiments.

Experiment 2c Sixty-three participants were enrolled in the experiment. Of those, eight participants were excluded from further analysis: one because of a missing demographics survey, three due to technical errors, and four due to a failure to follow directions (as indicated either by pressing every single key simultaneously on every trial or 10 or more trials without a timely response). That left 55 participants with analyzable data (27 female, 28 male). All participants were at least 18 years of age (M_{age} = 20.5 years, range: 18–29 years) and had no history of hearing impairment. Participants were recruited from the University of Maryland, College Park, community for either course credit or paid compensation.

Materials

All stimuli were identical to the ones used in Experiment 1. Experiments 2a and 2b used the fricative stimuli, and Experiment 2c used the instrument stimuli.

Procedure

For all experiments, a time-to-criterion paradigm was used to explore learning using these items. Participants were first given brief instructions, telling them that they would hear speech sounds and that they would be asked to pair them with colored squares using the keyboard in front of them. They then heard a single sound chosen at random from the relevant 10-step continuum. Other than the stimuli used, all other procedures were identical across the three subcomponents of the present study. This sound was presented simultaneously with a bank of three possible responses: a blue square, a yellow square, and a red square, presented in a single visual row on the screen. Note that the order of these squares did not reflect anything about the correct responses depicted in Fig. 1. Participants were given 5 seconds to pair the sound that they just heard with one of the three squares using one of three buttons on keyboard. They then received feedback about their selection in line with the condition they had been assigned to, as described below, which appeared 250 ms after the participant selected a square and stayed on the screen for 1 second. The feedback took the form of a yellow *X* if the participant responded incorrectly or a green check mark if the participant responded correctly. The feedback was followed by a 500-ms interstimulus interval (ISI).

The order of trials was randomized in blocks of 10 steps each, such that participants heard all 10 steps every 10 trials with no predictable order within the block. Participants heard trials until one of two conditions was met: either 450 trials elapsed, or when the participant responded correctly to 90% of the last quasiblock (the last 10 unique items), which could span portions of the last two successive blocks. This meant that participants had to correctly respond to the most recent appearance of 9 out of 10 items on the continuum to complete the experiment early. There were six conditions, assigned on a between-participant basis, with participant numbers in each condition approximately balanced. These conditions differed in which responses were considered correct on each trial. They are outlined in Fig. 1, in the Introduction. For example, the correct answer for Step 8 was yellow in the Neapolitan condition, red in the Sandwich condition, and blue in the Picket Fence condition. Note that item-color associations were not counterbalanced across participants. All three possible responses were available for all conditions, including the five conditions in which only two responses were correct.

The conditions differed in the numbers of possible categories and the number and composition of items assigned to each category. In the Normal condition, items were assigned to categories on the basis of the phonetic categorization preferences of English-speaking and German-speaking listeners from Key (2014), with a single boundary between Items 6 and 7. In the Shifted condition, the category boundary was moved to between Items 3 and 4. Almost every theory of category learning would predict that both of these conditions should be easy to learn if listeners start with a clean slate in learning the categories. Alternatively, because prior work using these stimuli (Key, 2014) suggests that the Normal condition represents listeners' initial subdivision biases for the fricative stimuli, it might suggest that this boundary location represents a natural acoustic discontinuity that could easily be latched onto (Diehl, 2000; Holt et al., 2004). If so, it may prove easier to learn than the Shifted condition, although both should be easily learnable.

In the Neapolitan condition, the category boundary of the Shifted condition was preserved, while a third category was added between the Stimulus Steps 7 and 8. In the Sandwich condition, the yellow stimuli from the Neapolitan condition were coded as red, thus making the red category disjunctive (including Items 1–3 and 8–10). Both these conditions involved two "cuts" along both continua. In the Neapolitan condition, there are then three categories along the continuum: red, blue, and yellow. In the Sandwich condition, meanwhile, there are just two categories, with the categories alternating along the continuum: red, blue, and red. In both cases, the boundaries are in the same places. As such, the interstimulus similarity is kept constant across conditions, as the yellow items in the Neapolitan condition are exactly as confusable with the adjacent blue ones as the second set of red items are in the Sandwich condition; the only difference is in which items are assigned to which categories.

In the Picket Fence condition, the assignment of items to categories went back and forth across the continuum, with Items 1, 2, 5, 6, 9, and 10 assigned to red, and Items 3, 4, 7, and 8 assigned to blue. Finally, in the Odd One Out condition, a boundary was placed between the red and blue categories between Items 5 and 6 (near where the boundary was in the Normal condition), but with a single item on either side (Items 3 and 8) being assigned to the category on the other side of the boundary (blue and red, respectively). The complexity in these last two conditions means that they should be hard to learn under any category learning theory—abstractionist theories because of the complexity of the category structure for both categories, and instance-based theories because the items should be relatively confusable across category boundaries.

Analysis

Most analyses for category-learning studies include a metric of the proportion of trials correct over time, averaged across blocks. Such experiments are based on multiple blocks, perhaps spaced across many sessions, where participants never quite reach an optimal learning strategy (Nosofsky, 1986), which was not the case here. Many category learning experiments also make use of mixed modeling to examine whether learners are improving over time within the experiment. Given that the primary interest in this article is whether participant learning would differ between conditions, the usual course of action would be to look for an interaction between trial number and condition using mixed models (Chandrasekaran, Yi, & Maddox, 2014; Scharinger, Henry, & Obleser, 2013). However, that approach is not ideal for the present methodology, where participants were cut off after a certain criterion point.

Instead, we used survival analysis. A survival analysis is meant to model how long it takes for a specific event to occur (e.g., death, recovery) across multiple groups. Under a survival analysis, the main dependent measure is the number of trials needed until the objects of study reach the criterion assigned to them. In epidemiology, where such analyses are common, this criterion might be something related to a health outcome, such as mortality or remission. In the present experiment, this criterion is getting 9 out of the 10 steps along the continuum correct the most recent time that criterion was reached. Participants can be censored if they do not reach criterion, allowing their data to be retained in the survival model without being logged as reaching criterion at the end of the study. The variation in the number of trials needed to reach criterion is used to create a hazard function, which is a model of the likelihood that a participant will reach criterion on each trial. These functions always increase over time; they differ in how quickly they increase, since, in some conditions, participants are faster to reach the criterion than others. These functions can then be compared across conditions.

Multiple statistical analyses were performed to analyze the time-to-criterion data. The Cox proportional hazards model (Lin & Wei, 1989) can allow for a global assessment of the significance of the influence of condition on time to criterion, as well as coefficients comparing each condition with a baseline value. The Cox proportional habits model also allows for outlier detection by way of deviance residuals, which pick out times to criterion that are unusually fast or slow relative to the expected values that could be calculated from the rest of the data set (Therneau, Grambsch, & Fleming, 1990). In line with best practice in survival analyses, Cox models, however, do not allow for a pairwise comparison of all of the conditions in the current experiments; the log-rank test (Mantel, 1966; Peto & Peto, 1972) can be used to compare hazard functions across conditions in a pairwise fashion to examine differences in the time to criterion. The R packages “survival” and “survminer” were used to analyze and graph the survival curve models (Kassambara, Kosinski, Biecek, & Fabian, 2018; Therneau, 2015).

Results and discussion

The results are presented in Fig. 5. As can be seen from the graph, there is a stark difference between the first three conditions and the last three conditions for the English speakers learning German fricatives (Experiment 2a). Participants generally found the Normal, Shifted, and Neapolitan conditions much easier than the Sandwich, Picket Fence, and Odd One Out conditions. Nobody failed to learn in any of the first set of conditions, for example. In general, the German-speaking participants of Experiment 2b were just as likely to struggle with disjunctions as were English speakers, despite their massively larger exposure to the speech sounds being learned over the course of the experiment. Some of the results for English speakers learning instrument categories (Experiment 2c) resemble those found for phonetic categories: the Normal and Shifted conditions are easy to master, whereas the Picket Fence and Odd One Out categories are difficult. However, a more detailed examination of the results indicates that the distinction between the Neapolitan and Sandwich conditions shrank for the nonspeech categories.

Figure 6 shows the proportion of participants who reached criterion across time for each condition and each experiment. From this figure, it appears that participants in the Normal and Shifted conditions were fastest to reach criterion in all studies, with the vast majority reaching 9 of 10 steps correct within just 50 trials. The Odd One Out and Picket Fence conditions were always much harder, with few participants ever reaching criterion. Meanwhile, the Neapolitan and Sandwich conditions differed from study to study in their difficulty. For Experiments 2a and 2b, the Sandwich condition was learned more slowly than the Neapolitan condition, whereas for Experiment 2c, there was little difference between the two conditions.

A detailed exploration of the results was undertaken using survival analyses. First, Cox proportional hazard tests were used to determine whether there was a global effect of condition on participant times to criterion. The global likelihood ratio test showed that the conditions differed from each other significantly in every individual study: in Experiment 2a, $\chi^2(5) = 63.3$, $p < .001$; in Experiment 2b, $\chi^2(5) = 50.3$, $p <$

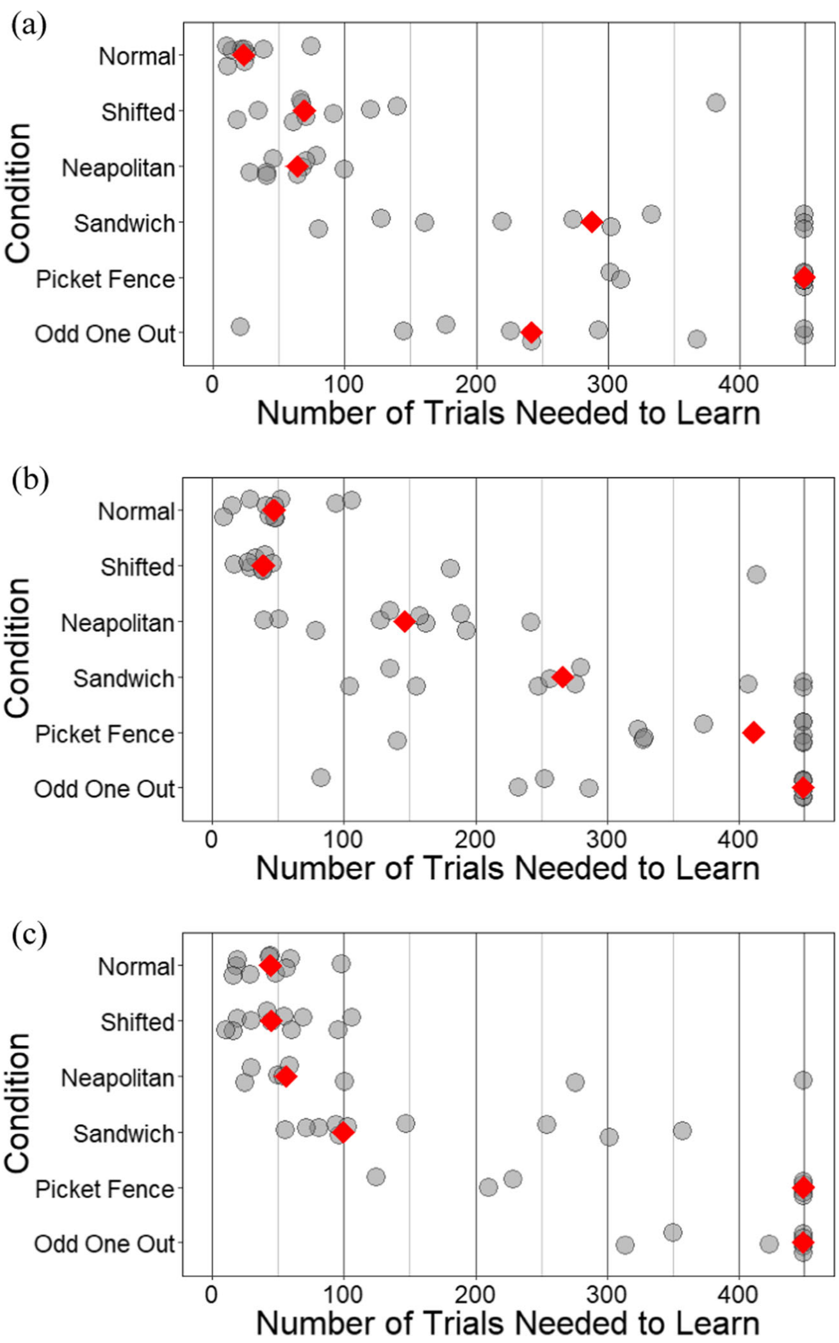


Fig. 5 Results for English speakers hearing German speech sounds (**a**), German speakers hearing German speech sounds (**b**), and English speakers hearing instrument sounds (**c**). Each row represents a single condition with individual participants shown as single circles. Horizontal displacement along the graph shows the time to criterion (90%: 9 of 10 steps) for each participant, with participants located at the vertical line at far right being participants who failed to learn within 450 trials. Points are jittered to better display the number of individual participants at each location. Red diamonds show the median time to criterion for each condition, including participants who did not reach criterion within 450 trials. (Color figure online)

.001; and in Experiment 2c, $\chi^2(5) = 50.1$, $p < .001$. Thus, participants differed in their speed to reach criterion in each experiment based on the condition.

Inspection of deviance residuals flagged a handful of values in each experiment as possible outliers, which were operationally defined in the present study as values more than two standard units from the expected value. These included four participants in Experiment 2a (Participant A0024, who took 384 trials to learn in the Shifted condition; Participant A0050, who took 20 trials to learn in the Shifted condition; Participant A0022, who took 76 trials to learn in the Normal condition; and Participant A0020, who took 22 trials to learn in the Odd One Out condition), two participants in Experiment 2b (Participant B0051, who took 10 trials to learn in the Normal condition, and

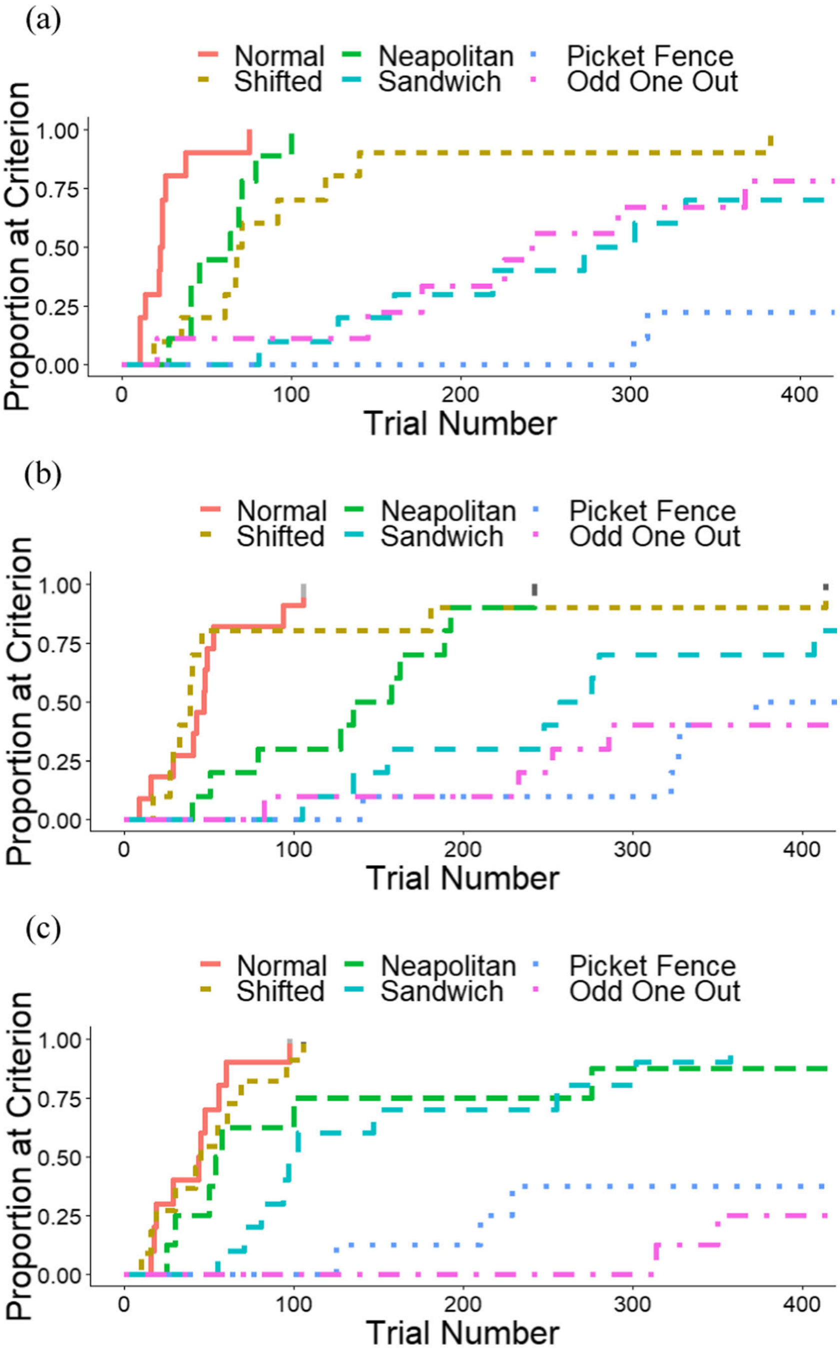


Fig. 6 Proportion of participants who reached criterion over time for each condition in Experiment 2, with different lines showing conditions, for Experiments 2a (**a**), 2b (**b**), and 2c (**c**). (Color figure online)

Participant B0041, who took 415 trials to learn in the Shifted condition), and two participants in Experiment 2c (Participant C0015, who took 11 trials to learn in the Shifted condition, and Participant C0016, who did not learn within 450 trials in the Neapolitan condition). The following results are primarily reported with outliers included, although special note is made where excluding the outliers would affect significance.

Finally, pairwise log-rank tests were performed to compare the hazard functions for each pair of conditions. These tests were Bonferroni corrected to account for multiple comparisons. Results of the log-rank tests are shown in Table 2 on an experiment-by-experiment basis.

The pairwise comparisons and the data graphed indicate that there were approximately three tiers of difficulty across the six conditions in Experiment 2a, although there was some overlap between the second and the third tiers. The Normal condition was the easiest to learn for English speakers learning German fricative categories; the hazard ratio for the Normal condition did not overlap with any of the other conditions. The Neapolitan and Shifted conditions were the next easiest;

Table 2 Pairwise log-rank test significance levels for Experiments 2a, 2b, and 2c

	Normal	Shifted	Neapolitan	Sandwich	Picket Fence
Exp. 2a					
Shifted	.04	–	–	–	–
Neapolitan	.007	1	–	–	–
Sandwich	<.001	.049	<.001	–	–
Picket Fence	<.001	<.001	<.001	.42	–
Odd One Out	.002	.20	.005	1	.14
Exp. 2b					
Shifted	1	–	–	–	–
Neapolitan	.008	1	–	–	–
Sandwich	<.001	.11	.04	–	–
Picket Fence	<.001	.004	<.001	1	–
Odd One Out	<.001	.002	.002	1	1
Exp. 2c					
Shifted	1	–	–	–	–
Neapolitan	.95	1	–	–	–
Sandwich	.006	.03	1	–	–
Picket Fence	<.001	<.001	.18	.06	–
Odd One Out	<.001	<.001	.05	.002	1

although they were both harder than the Normal condition, they did not significantly differ from each other. The hardest group was the Sandwich, Picket Fence, and Odd One Out conditions, none of which differed from each other. The only potential overlap between these distributions was in the Shifted and Odd One Out conditions, which did not reach significance after Bonferroni correction for multiple comparisons, although this difference was significant after removing outliers.

Like the English speakers, the German speakers learning German categories (Experiment 2b) found the Sandwich, Picket Fence, and Odd One Out conditions quite hard; they were not significantly different from each other in the time to criterion. Also like the English speakers, the German speakers in the Sandwich condition were generally slower to reach criterion than the German speakers in the Neapolitan conditions were. The divergences in the speaker groups were largely peripheral. Most notably, for the German speakers, but not for the English speakers, the Normal and Shifted conditions were not different from each other in the expected likelihood to reach criterion. There was also no significant difference between the Shifted condition and the Sandwich condition in their hazard functions, perhaps reflecting the influence of the single participant in the Shifted condition who took more than 400 trials to complete the experiment. The difference disappeared after removing the outliers of Experiment 2b, and, interestingly, was replaced instead with a significant difference between the Shifted and Neapolitan conditions, with the Neapolitan condition also being harder than the Shifted condition. Indeed, comparing Fig. 6a with Fig. 6b suggests that native German speakers found the Neapolitan condition more challenging relative to the English speakers, perhaps due to their experience with encountering only two categories within this particular phonetic continuum.

For the English speakers learning instrument categories (Experiment 2c), on the other hand, the results are relatively complicated. Some patterns emerge. Unlike in Experiment 2a, the Normal and Shifted conditions are no different from each other; this is unsurprising, given that there should be no reason for the musical instrument categories to be biased toward any specific partition of the stimulus space. The Picket Fence and Odd One Out conditions are both very difficult. And, most importantly, the Neapolitan and Sandwich conditions are also not significantly different from one another. The Sandwich condition was learned as quickly as the Neapolitan condition, although it is not clear whether the Neapolitan condition was easily differentiable from any of the other conditions in the present data set. The Sandwich condition was learned more slowly than the Normal and Shifted conditions, but more quickly than the Picket Fence condition. The Neapolitan condition, meanwhile, was not learned at a significantly different pace from *any* of the other conditions, perhaps due to the unusual distribution of times to learn in this condition—most participants reached criterion quickly, but one did not learn within 450 trials. Removing the outliers of Experiment 2c leads to a significant difference between the Neapolitan condition and the Picket Fence and Odd One Out conditions in their times to criterion. Regardless of whether the outliers are removed, though, the likelihood that a participant would complete the experiment on any single trial was not different between the two key conditions.

The behavior of participants in the Sandwich condition seems particularly instructive, since abstractionist and instance-based theories of category learning make different predictions even about the learning shown by participants who are not successful learners. Instance-based theories predict that unsuccessful learners should be approximately at chance for many of the stimuli, while abstractionist accounts predict that some participants may operate on the basis of (inaccurate) rules or abstractions. Figure 7 shows the responses for all the participants in the Sandwich condition within the last 25% of trials in the experiment within each participant group.

If participants were basing their responses only on the similarity of adjacent items, the ends of the continuum should be reddish, and the middle stimulus steps should be bluish, with purple (reflecting both blue and red responses for a certain point along the continuum) being likely around stimulus boundaries. This is what is seen for Participants A0006, A0038, B0011, C1014, and C1020, for example. Yet there are also participants with quite different patterns of responses. Participant A0013 had an almost linear grading from uniformly red on the velar end of the continuum to uniformly blue on the palatal end of the continuum. In many ways, the results are strikingly similar for German speakers to those of English speakers when learning German fricative categories. Only one large difference was present between Experiments 2a and 2b. Namely, most participants who did settle on the correct answer in Experiment 2a seemed to do so by anchoring the red category to the palatal end of the continuum, whereas

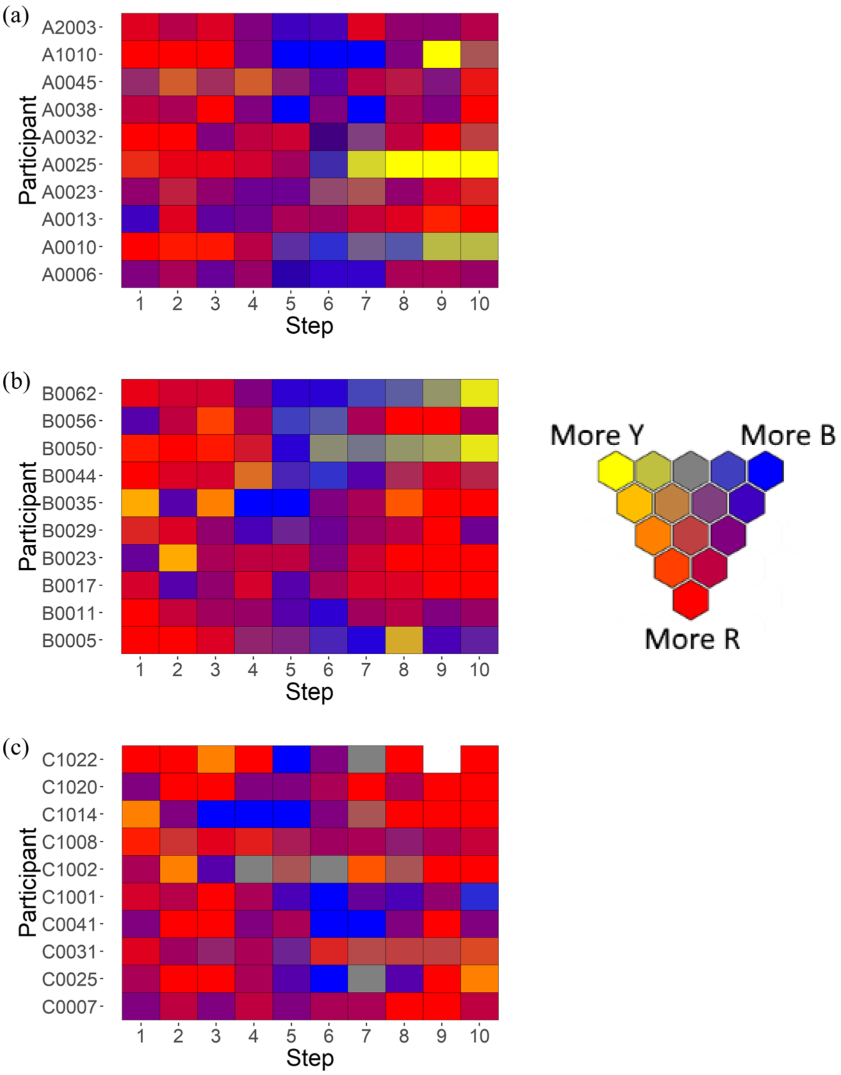


Fig. 7 Experiment 2a (**a**), 2b (**b**), and 2c (**c**) results for the Sandwich condition during the last 25% of trials; the legend on the right shows sample combinations of response patterns. Each row is an individual participant; each column is a step. Cells are colored in line with the proportion of responses pairing the step in question with a particular colored square (see legend). Pure red, blue, and yellow colors show near-universal responses of one type for a step, whereas colors like purple or gray represent a mix of responses. The white cell for Participant C1022 indicates that that participant did not have any responses for Stimulus Step 9 in the final 25% of trials. (Color figure online)

participants in Experiment 2b were just as likely to start from the velar end as from the palatal end.

Most strikingly, Participants A0010, A0025, A1010, B0050, and B0062—of whom all but A1010 did not learn at all within 450 trials—continued pressing "yellow" for the velar end of the continuum until almost the very end. They were so certain that there must be three putatively rule-based categories within the continuum that they were giving a yellow response through to the end of the experiment, even though they were always told that such a response was wrong! This was not merely a by-product of the fact that participants in the Sandwich condition had to learn to ignore a possible response. Participants in the Normal and Shifted conditions had to do the same thing, but had little problem ignoring the yellow square. The responses of participant A0025, in particular, show a clear categorical separation: Steps 1–5 as red, 7–10 as yellow, and 6 as blue, with some noise in responses. Many participants do not seem to be using similarity alone to categorize the fricative stimuli. This pattern of responses was also found in the Picket Fence conditions, where some participants showed a random pattern of responses across the continuum but others showed rule-like behavior.

Again, qualitatively different patterns emerged for the participants of Experiment 2c. Participants in the Sandwich condition were only very rarely using the yellow button to respond, unlike for the fricatives, and most of the participants seemed to have little trouble positing a red category that was on both ends of the continuum. Errors in the Picket Fence condition were more evenly scattered across the continuum. Meanwhile, participants in the Odd One Out condition across all experiments almost uniformly ignored the exception items on both sides of the continuum.

General discussion

We set out to test the learnability of auditory and auditory-phonetic categories. In Experiment 2, we trained native English speakers to categorize tokens from a German voiceless fricative continuum, native German speakers to learn categories with the same German fricatives, and native English speakers to learn categories of unfamiliar musical instrument sounds. We found that conditions that included easy-to-describe binary categories (the Normal and Shifted conditions), which should be easy under just about every theory of category learning, were indeed easy for all participants learning each set of items. Conditions that included complex categories (the Odd One Out and Picket Fence conditions) were challenging to learn for all participants and each set of items. The most interesting comparison was between the Sandwich and Neapolitan conditions, where the only difference between the conditions was in the assignment of one end of the stimulus continuum being learned. The Sandwich condition, on the other hand, was challenging for both English speakers and German speakers when learning categories in German fricative space, yet was no more difficult than the Neapolitan condition for English speakers learning musical instrument sound categories.

These patterns of results—in particular, the distinction between the Neapolitan and Sandwich conditions for the fricative stimuli—are challenging to explain using solely instance-based models of category learning. Strictly instance-based learning theories, such as GCM (Nosofsky, 1986), predict that categories with equally differentiable items should be equally easy to learn essentially no matter their structure. But this is not what was observed; instead, the Sandwich condition, which included a disjunctive category structure, was learned more slowly than the Neapolitan condition, which did not include one.

A subtler point of note relates to the performance of participants who failed to learn the pairings of speech sounds to colored squares. According to the GCM, there are essentially two possible outcomes for participants in conditions such as the Picket Fence condition: learning or guessing. Participants will learn the categories under consideration if they can discriminate the items on the continuum. If they cannot discriminate the individual items, however, participants will use items taken from multiple categories to determine category membership, choosing a category roughly in proportion to the items sampled from each category. Even the participants who failed entirely did not show this pattern. Instead, many participants in the three challenging conditions failed in ways that differed *systematically* from the input they were given. Unsuccessful learners seemed to be imposing some amount of structure on the input they were being given. Again, models that rely only on instance-based learning would find it challenging to accommodate this level of systematicity, particularly the behavior of the participants in the Sandwich condition who doggedly continued using the yellow response button even when such a response was never rewarded.

Explaining the differences between the Sandwich and Neapolitan conditions can be accommodated under many abstractionist theories of learning. Of course, the amount of abstraction necessary, and the precise mechanisms by which that abstraction occurs, are up for debate. However, not every abstractionist theory can predict the patterns observed here. Many abstractionist theories of category learning permit disjunctive rules (Minda et al., 2008; Zeithamova & Maddox, 2006), and, in fact, some studies of disjunctive phonetic learning have claimed that it is optimal for participants to use abstract rules to learn these categories (Maddox et al., 2014).

Instead, it seems that the abstractionist theories supported by the present data set should have a bias toward *simpler* categories, at least for speech items. These theories could, theoretically, involve very small changes to strictly instance-based conceptualizations of categorization. For example, a simple prototype theory of category learning, with a single prototype per category, could easily accommodate some aspects of this pattern of results, as the prototypes from both the blue and the red categories in the Sandwich condition would be centered in the exact midpoint of the continuum. Many modern formulations of prototype theories would still suggest that learning such categories would be hard, as the proximity of the category prototypes would make the categories highly unstable (Toscano & McMurray, 2010).

Under other abstractionist theories, learners may have a metacategorical bias against disjunctive categories. Recall the rational rules model of concept learning (Goodman et al., 2008), where hypotheses take the form of rules that are composed of conjunctions or disjunctions of sets when nondisjunctive categories along a single dimension are present. Rational rules suggests that learners have priors toward simple categories, priors that seem to be respected in the results of the phonetic components of Experiment 2. The RULes and EXceptions model (RULEX) provides another avenue of exploration (Nosofsky & Palmeri, 1998; Nosofsky, Palmeri, et al., 1994). In RULEX, categories are formed through a multistage process. First, learners try to identify simple rules to characterize the categories being taught to them. If those rules are categorically (or close-to-categorically) successful at characterizing the stimulus space, the rules are kept unaltered. If they are entirely unsuccessful, the learners try instead to learn more complex rules (i.e., multidimensional ones) that involve interactions between categories. And if the rules are moderately successful—say, successful 75% of the time—learners memorize exceptions to the rule, with the number and specificity of the exemplars depending on the learner's memory constraints. Again, this build-up from simple to more complex categories could explain the difference between the Neapolitan and Sandwich categories found for the speech stimuli.

A model like SUSTAIN (Love et al., 2004), meanwhile, where categories are treated as the sum of various "clusters" in the stimulus space, also shows some promise. SUSTAIN is said to attempt to learn simple category mappings before switching to more complex ones. Similar items tend to form clusters that pattern together. So, plausibly, having category learning situations in which dissimilar items formed clusters that patterned together (e.g., the Sandwich condition) would be more challenging to learn than when only similar items worked together (e.g., the Neapolitan condition). However, SUSTAIN must be modified to incorporate the partially supervised nature of the present experiment, where learners received incomplete feedback about their responses. That is, learners were informed their choice of color/category was wrong, but not which of the two alternative answers was correct. Meanwhile, multiple-system theories (Chandrasekaran, Yi, et al., 2014; Maddox & Chandrasekaran, 2014) predict that learners start with a bias toward simpler, rule-based categories, but fall back on instance-based learning if rule-based learning proves insufficient to learn the categories in question. However, like some of their abstractionist peers, the success of multiple-system theories in understanding the acquisition of disjunctive categories depends on whether the rule-based learning system comes with a bias against disjunctive categories, something that is underspecified in most accounts.

Besides the implications of this work for theories of category learning, this work also can shed light on the role of expertise and domain-specificity on category learning. The English and German groups showed far more similarities than differences when learning the German fricative categories. Like English-speaking participants, German-speaking participants found disjunctive phonetic categories harder to learn than nondisjunctive categories. The differences between English and German speakers learning fricative categories were relatively minor, constrained primarily to the Normal and Shifted conditions. The English speakers found the Normal condition easier than the Shifted condition, whereas the German speakers found the two conditions equally easy to learn.

Why should the English speakers show a bias in learning that is not present for the German speakers, who have much more expertise with a category boundary at exactly the location of the boundary in the Normal condition (Key, 2014)? Although computational descriptions of the possible effects of expertise are myriad (Palmeri, Wong, & Gauthier, 2004), many focus on the discriminability of items that one has expertise with. In particular, many conceptions of expertise involve sharpening the representations of items, whether through improving the quality of the representations themselves (Nosofsky, 1987) or through decreasing perceptual noise throughout a stimulus space (Goldstone, Lippa, & Shiffrin, 2001). If German speakers' representations of the speech sounds along this continuum, or of the space itself, are less subject to noise, this could help explain why, for them, the Shifted condition was no harder to learn than the Normal condition. English speakers, meanwhile, might still be subject to simple perceptually driven biases that later training might make it possible to overcome (Diehl, 2000). Given the crucial role of discriminability for both abstractionist and instance-based theories of learning, these differences in discriminability could be used to explain observed discrepancies between English and German speakers in their learning. Interestingly, this is rather different than many accounts of categorical perception in speech contexts, which emphasize decreases in the discriminability of items within categories (Goldstone & Hendrickson, 2010).

Regarding domain-specificity, meanwhile, the results are more striking. The main finding of the studies with fricative categories was not replicated for the instrument categories. In the musical instrument continuum, there was no difference in the learning of the disjunctive and nondisjunctive categories, even though, in Experiment 1, it was shown that the two continua were well matched. The difficulty of the Sandwich condition for the fricatives cannot be explained by recourse to expertise alone, given the similarity of the results of English speakers (phonetic "novices") to those of German speakers (phonetic "experts"). These experiments provide a useful future target for studies of domain-specificity in phonetic category learning: disjunctive, unidimensional categories. Interestingly, these sorts of disjunctive rules seem to describe some other categories: Recall that musical notes are identically categorized across a single pitch space. Why should phonetic categories work any differently? And how might listeners learn cases where categories could be said to be disjunctive—say, the many realizations of the phoneme /t/ (Pitt, Dilley, & Tat, 2011)? The answers to such questions require more study, but pilot data indicates that the antidisjunctivity bias also extends to consonant length categories.

Interestingly, this bias was present even though participants were not always confident that the fricative stimuli were speechlike, as might be expected given the fact that they were fricatives presented in isolation. We did not ask participants explicitly what they thought the sounds they were hearing were, but we did ask participants if they noticed any patterns in the experiment. Of the participants who reported a pattern that involved characterizing the sounds used in Experiments 2a, approximately one third mentioned something explicitly speechlike (e.g., "s," "ts," or "ch"), many of the others mentioned something that could be characteristic of speech or nonspeech (usually "high pitch" vs. "low pitch"), and a few mentioned some other possible sound (coughing sounds and phlegm were popular descriptions of the velar fricatives). The German speakers of Experiment 2b, meanwhile, almost never mentioned something unambiguously speechlike in their description of the patterns in the experiment, although most seemed to make the connection when the pattern was explicitly pointed out during the debriefing.

At the same time, the bias against disjunctive categories is just that, a bias. A domain-specific bias is quite challenging to accommodate under almost any theory of learning, primarily because these theories were initially formulated within a single domain (usually the visual domain). Indeed, most instance-based learning theories borrowed into language made the specific and strong prediction that learning processes in the phonetic domain are *identical* to those found outside of it (Port, 2007, 2010), which, if these results reflect domain-specificity, cannot be supported. Many participants were able to successfully learn to pair sounds to disjunctive categories eventually. If the difference in results reflects differences between speech and nonspeech domains in terms of category learning, this suggests a more nuanced picture of domain-specificity that allows for very similar learning processes inside and outside of language, but the possibility of different constraints for the two domains. It might be the case, for example, that the rule-based category learning system has different properties for phonetic and nonphonetic stimuli such that it is constrained to positing nondisjunctive categories for speech in a way that it is not required to do for nonspeech stimuli. In other words, the "overhypotheses" of phonetic learning would be different from the "overhypotheses" of nonphonetic learning, in the terminology of hierarchical Bayesian models (Kemp, Perfors, & Tenenbaum, 2007).

Learners may be biased in this way because of the relative frequency of disjunctive categories in language versus disjunctive categories in music. Disjunctive categories are rarely present in language, barring occasional counterexamples, as with allophones of /t/. Music, however, contains at least one highly salient example of disjunctive categories—namely, the musical pitch categories described in the Introduction. On the other hand, the fact that participants only infrequently categorized the speech tokens as being speechlike also suggests the possibility that the differences between the speech and nonspeech tokens could instead be explained by idiosyncratic properties of the stimuli. Teasing apart domain-specificity from stimulus-specific processing requires the study of additional speech and nonspeech contrasts.

An alternative explanation of the differences between the speech and nonspeech stimuli relates to the perceived dimensionality of the continua used in the present experiment. Experiment 1 showed that the speech and nonspeech stimuli used were roughly comparable. However, they were not identical, and, in some ways, it appears that the nonspeech stimuli were perceived to be less unidimensional than the speech stimuli. Could this difference in perceived dimensionality have led to the differences in the patterns observed across the two sets of stimuli?

On the face of it, the idea is plausible. Although the differences in the reported scales of similarity between the items do not seem large, quantifying those differences poses a challenge. The predictions of all theories of category learning depend strongly on the dimensions that make up stimulus categories. Consider, again, the example of categories within temperature. If one considers how to categorize temperatures as "uncomfortable" along a linear continuum of temperatures (with a scale from *very cold* to *very hot*), uncomfortable temperatures are disjunctive. If, on the other hand, the dimension of similarity of temperatures is reframed in terms of a deviation from the most perfectly comfortable temperature (i.e., the scale is now from *completely comfortable* to *completely uncomfortable*), uncomfortable temperatures now form a nondisjunctive category.

For our stimuli, if the categories are perceived in a multidimensional fashion, it may be possible to obtain the pattern of results seen here even under a single-system, instance-based theory. Doing so would require an interesting set of suppositions. Under theories such as the GCM (Nosofsky, 1986), changing the perceived importance of an individual dimension for categorization takes time. Consider a situation in which the two-dimensional solution of Fig. 4b is the one that matches the true dimensionality of these stimuli. The two dimensions could be labeled "stimulus step" (*x*-axis) and "certainty" (*y*-axis), given that the latter dimension seems to sort extreme items from less extreme items along the stimulus continuum. If participants come into the category learning scenario with a strong weight on one dimension over the other, the cost to switch between dimensional weights to a more optimal configuration might explain differences in learning times between conditions. For example, a learner with a strong weight on certainty but a weak one on stimulus step would thrive in the Sandwich condition, as characterizing the stimuli by certainty alone is an excellent strategy for learning a distinction between the red and blue categories; the members of each category are very similar to each other in terms of certainty. Learners with the opposite dimensional weights would take time to switch to a more optimal configuration, making their learning slower.

This story is complicated by the patterns of similarity in the speech and instrumental continua. In the instrumental continuum, the end points were rated as more similar to one another than in the speech continuum. This means that the stimulus step dimension would be a worse determinant of category membership for the nonspeech items than the speech items. Thus, if learners start with a bias toward weighting the stimulus step dimension highly, they should find the instrumental categories more challenging than the fricative categories, as the instrumental categories would be more confusable. But, if anything, the opposite effect appeared; the Sandwich condition is easier for the nonspeech stimuli than the speech stimuli. If, on the other hand, learners start with a bias toward attending to the certainty dimension, there should be no particular reason for the Sandwich condition to be harder than the Neapolitan condition for either stimulus group—and, in fact, the Sandwich condition may in some ways be *easier* than the Neapolitan condition, as the yellow and red stimuli would then be confusable in the Neapolitan condition. Such a story would require an interaction between dimensional bias and stimulus group, with participants in Experiment 2c being biased toward weighting stimulus step strongly and participants in Experiments 2a and 2b weighting each dimension roughly equally. This is possible, but requires more study, perhaps including tests of discriminability or similarity before and after training. Alternatively, the speech continuum may be unidimensional while the nonspeech continuum is multidimensional. However, again, this would lead to the prediction that the nonspeech continuum would be harder than the speech one in general, which does not seem to be the case.

Both the results suggesting a bias against complex category formation and the notion that these results may be domain-specific provide an interesting contrast to the results reported by Moreton et al. (2017), who showed almost the opposite pattern. That is, learners of complex phonotactic patterns were not always slower or worse learners than learners of simpler-to-describe patterns in the study of Moreton et al., and those patterns were identical across phonotactic categories and visual categories. This is unexpected under many of the abstractionist category learning theories previously cited, including rational rules (Goodman et al., 2008) and RULEX (Nosofsky & Palmeri, 1998), where categories that rely on complex rules are almost universally nonpreferred. The results described presently are on the face of it challenging to reconcile with those of Moreton et al. However, Moreton et al. (2017) studied phonotactic categories, which contrast with phonetic categories on many dimensions. Primary among these is the fact that the phonotactic constraints studied by Moreton et al. (2017) included several binary dimensions (e.g., ± voiced for consonants, ± high for vowels), whereas the present experiments concentrated on a single continuous dimension. The categories being learned were also distributed over time, as the constraints were enforced on the co-occurrence of nearby segments rather than on a single category. Further, the nonlinguistic category learning task that was chosen in Moreton et al. (2017) was in the visual domain rather than the auditory domain. It is not clear how those methodological differences would map onto the differences between the results described here and those of Moreton et al. (2017), but dimensionality in particular is known to strongly influence categorization.

To conclude, an investigation of auditory category learning uncovered evidence, using speech category learning on the part of English-speaking and German-speaking participant groups, of a bias against disjunctive categories in category learning. This bias was not present for musical instrument categories. These studies have two major implications. First, instance-based theories, in which constraints militating against certain types of category structures are challenging to include, may require some revision to accommodate these findings. And, second, these biases appear to be domain-specific, appearing in the speech domain but not in a relatively similar set of nonspeech items.

Acknowledgments This work was supported by a National Science Foundation Graduate Research Fellowship award and a University of Maryland, College Park Graduate School Flagship Fellowship, to C.C.H., as well as NSF IGERT Grant 0801465 (PI: C. Phillips), an NSF Linguistics Doctoral Dissertation Research Improvement grant awarded to W.J.I. (co-PI: CCH), BCS 1650791, the Maryland Language Science Center, and the University of Maryland–University of Tübingen International Interdisciplinary Research and Teaching Collaboration. Data from this experiment will be available online, in line with the Open Science Framework, for participants who consented to sharing data. We thank Peter Deaville, Stephen DeVilbiss, Scott Kaplowitz, Priyanka Konanur, Zoe Schlueter, and the members of the UMD Language Development Lab for their help in running English-speaking participants for this project. We thank Andrea Weber's lab at Eberhard Karls University, Tübingen, for their space and support for the German-speaking participants, particularly Sara Beck, Ann-Kathrin Grohe, Lisa Kienzle, and Sarah Schwarz. Michael Key generously allowed us to use his German fricative stimuli for this study.

Many people have provided discussion and insightful comments during the development and publication process in this study. Among others, we would like to thank Ann Bradlow, Al Braun, Catherine Carr, Bharath Chandrasekaran, Jeff Chrabaszcz, Karthik Durvasula, Naomi Feldman, Drew Hendrickson, Lori Holt, Ellen Lau, Todd Maddox, Holger Mitterer, Emily Myers, Chris Neufeld, Rob Nosofsky, Janet Pierrehumbert, Eva Reinisch, Kirsten Smayda, and Joe Toscano. Jared Linck gave immensely useful statistical support. Portions of this research were presented at the eighty-eighth annual Meeting of the Linguistics Society of America (LSA) in Minneapolis, Minnesota, travel to which was financed by the Department of Linguistics at the University of Maryland, College Park, at the sixth annual Michigan State Undergraduate Linguistics Conference (MSULC) in East Lansing, Michigan, travel to which was funded by the Department of Linguistics and Germanic, Slavic, Asian, and African Languages of Michigan State University, and at the Mini-Workshop on Phonetic Processing and Learning at Eberhard Karls University, Tübingen in Tübingen, Baden-Württemberg, Germany.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

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*(3), 442–481. 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*(1), 33–53. https://doi.org/10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400. https://doi.org/10.1006/jmps.1993.1023

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363–378. https://doi.org/10.3758/BF03210826

Boersma, P., & Weenink, D. (2001). Praat: Doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity perception. XIII. Perceptual anchor model of context-coding. *Journal of the Acoustical Society of America*, *76*(3), 722–731. https://doi.org/10.1121/1.391258

Buxó-Lugo, A., & Watson, D. G. (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of Memory and Language*, *90*, 1–13. https://doi.org/10.1016/j.jml.2016.03.001

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, *24*(2), 215–221. https://doi.org/10.1017/S0272263102002061

Carroll, J. D., & Chang, J.-J. J. (1970). Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283–319. https://doi.org/10.1007/BF02310791

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

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *Journal of the Acoustical Society of America*, *128*(1), 456–465. https://doi.org/10.1121/1.3445785

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488–495. https://doi.org/10.3758/s13423-013-0501-5

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*(3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

Diehl, R. L. (2000). Searching for an auditory description of vowel categories. *Phonetica*, *57*(2/4), 267–274. https://doi.org/10.1159/000028479

Diehl, R. L. (2008). Acoustic and auditory phonetics: The adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *363*(1493), 965–978. https://doi.org/10.1098/rstb.2007.2153

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.026

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*, 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154. https://doi.org/10.1080/03640210701802071

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3–4), 373–405. https://doi.org/10.1016/j.wocn.2003.09.006

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review*, *23*(3), 351–379. https://doi.org/10.1515/TLR.2006.014

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*. https://doi.org/10.1037/0033-295X.93.4.411

Holt, L. L., & Lotto, A. J. (2008). Speech perception wthin an auditory cognitive science framework. *Current Directions in Psychological Science*, *17*(1), 42–46. https://doi.org/10.1111/j.1467-8721.2008.00545.x

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *Journal of the Acoustical Society of America*, *116*(3), 1763–1773. https://doi.org/10.1121/1.1778838

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116–122. https://doi.org/10.1037/h0035772

Johnson, E. K., & Seidl, A. (2008). Clause segmentation by 6-month-old infants: A crosslinguistic perspective. *Infancy*, *13*(5), 440–455. https://doi.org/10.1080/15250000802329321

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M. J. Sole, P. Speeter Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology: In honor of John Ohala* (pp. 25–40). New York, NY: Oxford University Press.

Kassambara, A., Kosinski, M., Biecek, P., & Fabian, S. (2018). survminer: Drawing survival curves using 'ggplot2' [Computer software]. Retrieved from https://cran.r-project.org/web/packages/survminer/index.html

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x

Key, M. (2014). Positive expectation in the processing of allophones. *Journal of the Acoustical Society of America*, *135*(6), EL350–EL356. https://doi.org/10.1121/1.4879669

Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, *46*(2/3), 295–349. https://doi.org/10.1177/00238309030460020201

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Lin, D. Y., & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, *84*(408), 1074–1078.

Lindsay, S., & Gaskell, M. G. (2010). A complementary systems account of word learning in L1 and L2. *Language Learning*, *60*(Suppl. 2), 45–63.

Lisker, L. (1985). The pursuit of invariance in speech signals. *Journal of the Acoustical Society of America*, *77*(3), 1199–1202. https://doi.org/10.1121/1.392185

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 732–753.

Lotto, A. J., Sato, M., & Diehl, R. 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* (pp. C181–C186). Cambridge, MA: MIT Press.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Maddox, W. T., & Chandrasekaran, B. (2014). Tests of a dual-system model of speech category learning. *Bilingualism: Language and Cognition*, *17*(4), 709–728. https://doi.org/10.1017/S1366728913000783

Maddox, W. T., Chandrasekaran, B., Smayda, K., Yi, H.-G., Koslov, S., & Beevers, C. G. (2014). Elevated depressive symptoms enhance reflexive but not reflective auditory category learning. *Cortex*, *58*, 186–198. https://doi.org/10.1016/j.cortex.2014.06.013

Maddox, W. T., Molis, M. R., & Diehl, R. L. (2002). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics*, *64*(4), 584–597. https://doi.org/10.3758/BF03194728

Mair, P., De Leeuw, J., Borg, I., & Groenen, P. J. F. (2016). smacof: Multidimensional scaling [Computer software]. Retrieved from https://cran.r-project.org/web/packages/smacof/index.html

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, *50*(3), 163–170.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128–148. https://doi.org/10.1037/0096-1523.21.1.128

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. https://doi.org/10.1111/j.1467-7687.2009.00822.x

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*(2), 219–246. https://doi.org/10.1037/a0022325

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1518–1533. https://doi.org/10.1037/a0013355

Moreton, E., Pater, J., & Pertsova, K. (2017). Phonological concept learning. *Cognitive Science*, *41*(1), 4–69. https://doi.org/10.1111/cogs.12319

Myers, E. B. (2014). Emergence of category-level sensitivities in non-native speech sound learning. *Frontiers in Neuroscience*, *8*, 238. https://doi.org/10.3389/fnins.2014.00238

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of intergral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. https://doi.org/10.1037/0033-295X.101.1.53

Pajak, B., & Levy, R. (2014). The role of abstraction in non-native speech perception. *Journal of Phonetics*, *46*(1), 147–160. https://doi.org/10.1016/j.wocn.2014.07.001

Palmeri, T. J., Wong, A. C.-N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences*, *8*(8), 378–386. https://doi.org/10.1016/j.tics.2004.06.001

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, *135*(2), 185–207.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101–139). Berlin, Germany: Mouton de Gruyter.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2/3), 115–154. https://doi.org/10.1177/00238309030460020501

Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics*, *39*(3), 304–311. https://doi.org/10.1016/j.wocn.2010.07.004

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1493), 1071–1086. https://doi.org/10.1098/rstb.2007.2160

Port, R. F. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, *25*(2), 145–172. https://doi.org/10.1016/j.newideapsych.2007.02.001

Port, R. F. (2010). Rich memory and distributed phonology. *Language Sciences*, *32*(1), 43–55. https://doi.org/10.1016/j.langsci.2009.06.001

Pycha, A. (2009). Lengthened affricates as a test case for the phoneticsphonology interface. *Journal of the International Phonetic Association*, *39*(1), 1–31. https://doi.org/10.1017/S0025100308003666

Pycha, A. (2010). A test case for the phonetics-phonology interface: Gemination restrictions in Hungarian. *Phonology*, *27*(1), 119–152. https://doi.org/10.1017/S0952675710000059

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110. https://doi.org/10.1037/0033-2909.92.1.81

Rocamora, M., López, E., & Jure, L. (2009, September). *Wind instruments synthesis toolbox for generation of music audio signals with labeled partials*. Paptper presented at the 12th Brazilian Symposium on Computer Music, Recife, Brazil.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210. https://doi.org/10.1006/jmps.2001.1379

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, *31*(4), 307–314. https://doi.org/10.3758/BF03202653

Scharinger, M., Henry, M. J., & Obleser, J. (2013). Prior experience with negative spectral correlations promotes information integration during auditory category learning. *Memory & Cognition*, *41*(5), 752–768. https://doi.org/10.3758/s13421-013-0294-9

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. https://doi.org/10.1037/h0093825

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, *48*(2), 553–566. https://doi.org/10.3758/s13428-015-0599-7

Smith, R., & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics*, *40*(2), 213–233. https://doi.org/10.1016/j.wocn.2011.11.003

Squire, L. R. (2009). Memory and brain systems: 1969–2009. *Journal of Neuroscience*, *29*(41), 12711–12716. https://doi.org/10.1523/JNEUROSCI.3575-09.2009

Therneau, T. M. (2015). survival: A package for survival analysis in S [Computer software]. Retrieved from https://cran.r-project.org/package=survival

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x.Cue

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1/2), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Ullman, M. T. (2016). The declarative/procedural model: A neurobiological model of language learning, knowledge, and use. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 953–968). Amsterdam, Netherlands: Elsevier.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*(2), 387–398.