Running head: AUDITORY CATEGORICAL-PERCEPTION

Supervised Learning of Auditory Categories:

A Pilot Study

Course: PSYC 494D1/D2 (Psychology Research Project 1)

Term: Winter/Summer 2021

Student Name: Robert Contofalsky

Student ID: 260890115

Supervisor: Dr. Stevan Harnad

Date Submitted: September 16, 2021

Abstract

Learning a category requires detecting the features that distinguish the members from the non-members. After a category has been successfully learned, we sometimes observe a phenomenon referred to as learned Categorical Perception (CP), in which members of different categories are perceived as more different (between-category separation) and members of the same category come to be perceived as more similar (within-category compression). This present study aimed to test whether it is possible to induce CP using sound-sequences that were generated in the lab. Two categories of sounds were generated (Kalaphones and Lakaphones) and were studied with a pairwise similarity judgment task (40 pairs) before and after a supervised categorization training task consisting of 400 stimuli. Subjects would begin and end our study with a similarity rating task which would play pairs of Kalaphones (K) or Lakaphones (L) or both, back-to-back in rapid succession, and subjects were asked to rate how similar or dissimilar these pairs were from each other. The scores they submitted would be used to evaluate their between-category separation, and within-category compression. Once the first similarity judgment task was completed, subjects would start the categorization trial, where they had to identify 400 stimuli as either a K or L, and were given corrective feedback based on whether their answer was right or wrong. We found that our stimuli are learnable and 3 of the 4 the learners who met our 80% correct criterion in our sample show a significant CP effect. The number of correct trials in block 4 was significantly correlated with the size of the CP effect across all subjects, and the same trend (positive but not quite significant) was present when we excluded the 9 lab members from the sample. We also examined our stimuli and subjects' performance for any potential bias (i.e., did a K or L appear in a specific section of our stimuli more often than in others, or did our subjects perform better when a K or L appeared in a specific chunk?). We found no position bias in our

stimuli, and only two instances of performance bias. These pilot results will help us further calibrate our auditory stimuli for future studies.

Key words: categorization, auditory categorical perception, supervised training/learning, chunking

Supervised Learning of Auditory Categories: A Pilot Study

Categorization

Categorization is the action of sorting things into categories based on features that distinguish the categories (Harnad, 2017; Pérez-Gay et al., 2017). What determines which thing belongs to which category is the features that distinguish the categories. These covarying features are what allow us to distinguish the members from the non-members. This process is done by sensorimotor systems (organisms) interacting with their environment. How they interact with their environment depends on what their environment "affords" them in a specific context (Gibson, 1979). For example, a key could be used to open locks, but it would not be used as a frisbee or boomerang, because its features do not afford such usage. Similarly, a key is a category, but this category contains a vast array of members: transponder keys (car keys), padlocks (house keys), etc. One feature that they all have in common and that is lacked by non-keys is that they are objects that can be used to unlock keyholes. Where these objects differ is what kind of keyhole they can unlock (i.e., it would be impossible to use a house key to unlock a car, and vice-versa). Thus, to use a house key to unlock a house would be doing the right thing (using a key to unlock something) with the right kind of thing (using a padlock to unlock the house). That is the essence of categorization: doing the right thing with the right kind of thing (Harnad, 2017).

There are two dominating views in categorization: the classical and the "non-classical" view. The classical view is that all category members share invariant features that determine whether they belong to one category or another, and this distinction is absolute (Harnad, 2017) rather than a matter of degree (Mervis & Rosch, 1981). Some things will belong to one category based on the features they share (covariance), and things that do not belong to that category lack

those features. It is important to keep in mind that categorization and discrimination are not the same thing. Miller (1956) pointed out that discrimination is a relative judgment (i.e., simply comparing similarities and dissimilarities between two or more objects), whereas categorization is an absolute judgment. To categorize something is to identify something in isolation. The alternative/non-classical view is that categories are not absolute and determining category membership becomes more of a matter of degree (McCloskey & Glucksberg, 1978). The rationale behind this view is that the boundaries between categories are "fuzzy", and therefore everything is a member of every category, and what makes some categories more similar to others is their degree of similarity. For this project, the classical view will be assumed.

Category Learning

Categories can be learned by detecting the features that distinguish them through sensorimotor interaction with members and non-members. This applies to categories that are either innate or learned. One example of an innate category in humans would be our ability to perceive colour (Jacobs, 2013). However, most categories are not genetically encoded, and must therefore be learned (Pérez-Gay et al., 2017).

An organism can learn a category through direct sensorimotor input with two forms of learning: unsupervised and supervised learning. Unsupervised learning is learning a category through mere exposure. This type of learning involves an organism passively receiving inputs from its environment and learning to detect the similarities and dissimilarities of the inputs through the correlations among the features of the inputs it encounters. Unsupervised learning is known to be effective (Ell et al., 2012), but its efficiency is dependent on a lot of repetition and on whether the relevant features of the categories are salient enough to be detected without any form of assistance (e.g., the features that make up a car are very different from the ones that

make up a bike, even if both can be categorized as a means of transportation). On the other hand, supervised learning is when you learn a category through trial-and-error and receive corrective feedback based on whether your answer was right or wrong. This type of learning will enable the organism to learn categories that are more complex and nuanced (categories in which only small features distinguish them). This is because the corrective feedback given to the organism helps the brain detect the relevant features of a category more efficiently than through mere exposure (LeCun et al., 2015). An example of this would be trying to identify and distinguish the various species of bees inhabiting the globe. Many of them look alike and have very similar features (they are pollination-based flies that sport similar colours). To learn the difference between a honeybee and a killer-bee, corrective feedback would be necessary, as the features that distinguish them are minute.

It is important to keep in mind that context plays a role in learning as well. Many objects can be used for one purpose but can be used for a completely unrelated one if the object affords such usage. An example of this would be how a table is normally used to place objects on top of it, but if a person would want to sit on it because he/she is tired, the features of the table afford doing so. The feedback from the consequences of doing the right or wrong thing allow organisms to learn to identify which features covary with category membership and to differentiate what to do with the members and non-members. This makes categorization approximate rather than exact; as neither unsupervised nor supervised learning allow an organism to extract all possible features of each input/category. Rather, what happens is that once enough features have been detected to do the right thing with members and non-members, unless there is a change in context, it is unnecessary to seek out more features to detect; the ones that are being used and work are sufficient.

The third (and only indirect) type of learning is through verbal instruction. It is unique to humans and is defined as the ability to acquire new categories through naming and describing them (Blondin-Masse et al., 2013). After enough categories have been "grounded" directly – either learned through observation (unsupervised learning) and/or trial-and-error feedback (supervised learning) – humans can name them. Naming a category means assigning an arbitrary symbol (usually spoken or written) to it: this category-name is one that has been collectively agreed to by the speakers of the language. The English word "bottle" does not have any physical similarities to an actual bottle, nor does the Arabic word "زجاجة" resemble its English counterpart, yet both names (arbitrary symbols) refer to the same category.

As previously stated, the ability to correctly categorize requires selectively abstracting a sufficient number of relevant features to identify the members of a category. Hence naming designates a category because it is "grounded" by having detected its relevant features through sensorimotor experience. As a result, a person can now combine and re-combine category names to create subject/predicate propositions to further define and/or describe categories whose features are a combination of pre-existing categories which have been defined/described exclusively through verbal propositions: "husband = married man"; "bachelor = unmarried man"; "wife = married woman"; "bachelorette = unmarried woman", etc. The recombination of categories into propositions enables humans to learn categories (through verbal instruction) at a much faster rate than through direct sensorimotor interaction with them. As the verbal categories have already been directly "grounded" through sensorimotor experience, humans can teach each other categories without having to take the time to be in direct contact with them. For instance, a person can learn what a Zebra is without having seen one in person; all they have to know is that it is a horse-like creature with white and black stripes marked across its entire body. Once that

description has been understood by the listener, it no longer becomes necessary to physically see one to know what a Zebra is.

Categorical perception

Categorical perception (CP) is the phenomenon that occurs when members of different categories look more dissimilar from one another (between-category separation) and/or members of the same categories look more similar (within-category compression) than one would otherwise expect (Harnad, 1987/2003; Goldstone & Hendrickson, 2010; Notman et al., 2005). Learned CP occurs in some instances (but not all) as a consequence of learning to detect the relevant features of two or more categories. CP is not to be confused with categorization itself, or learning to categorize. Categorization is about the proper sorting of things (doing the right thing with the right kind of thing), and learned CP is a perceptual effect (occurring after category learning) that makes the differences between different categories more obvious (separation) and the similarities between members of the same category more apparent (compression).

Though it may seem trivial to say that different categories look different from one another, the way we perceive categories is not strictly based on how similar they are to one another. If this had been the case, then an organism would be able to learn all categories through unsupervised learning. The capacity to do the right thing with the right kind of thing is not conferred only by innately detected similarities (i.e., the differences we can spot through mere exposure). At times, corrective feedback (supervised learning) may be the only way to learn to detect which features of a specific category are relevant and which features can be ignored.

CP has been studied since the late 1950s. In its infancy, Liberman et. Al., 1957, noticed that speech phonemes segregate perceptually into qualitatively distinct categories, like colors. This

effect went on to be found in non-speech sounds too (Aaltonen et al., 1997; Guenther et al., 1999). CP has also been studied in visual stimuli such as faces and colors (Etcoff & Magee, 1992; Sauter et al., 2011) and other visual stimuli (Andrews et al., 2015). More recently, CP has been found to be inducible by learning (Pérez-Gay et al., 2017).

The present investigation

The present research examined whether categories of morse-code-like sound-sequences are learnable, and whether the learning generates CP. This is a calibration study in its pilot phase. Our results were collected to assess whether our sound-sequences were balanced and unbiased or require further adjustment. We compared similarity judgments for pairs of inputs in the same category (20) or in different categories (20), before and after categorization training (400 trials). We tested whether the categories would be learnable with supervised learning and whether successful category learning would produce CP (decreased pairwise similarity between categories and/or greater similarity within). We predicted that the categories would be learnable and that learning would induce CP.

The sounds were a series of 12 bits (dees or dahs) in 3 chunks of 4 bits. One of the 3 chunks contained the relevant feature, the other two chunks were irrelevant. The relevant feature could be in chunk 1, 2 or 3 and its position varied randomly across trials. We tested whether there was any (1) position bias (did the relevant feature appear the same number of times in chunk 1, 2 and 3?) (2) category bias (were there an equal number of Kalaphones and Lakaphones with the relevant feature in chunk 1, 2 and 3?) and (3) performance bias (did the subjects perform equally well when the relevant feature was in chunk 1, 2 and 3?). We predicted that there would be no position, category, or performance bias.

Methods

Subjects and Procedure

This study took place over the summer and included a total N of 35 subjects (46% Female, 54% Male); *M*(age) 30.8, *SD* = 8.17). Of the 35 subjects, 9 were active members in Dr. Harnad's lab (Laboratoire de Cognition et Communication) and the rest were either relatives or friends of the co-authors. Subjects were recruited through text or phone calls and were not compensated for the study (though, their time and effort were greatly appreciated).

This study was conducted virtually and was hosted on a website server. Before taking part in the study, we asked subjects to wear headphones/earphones to participate, and to give themselves at least 90 minutes to complete the task, as it cannot be paused. Furthermore, we required all subjects to use the latest version of Google Chrome, as this was the browser that would run our study from beginning to end with a minimum number of bugs. Measures like these were our best ways to ensure that we could reduce the high variability (variability being that everyone has different computers, internet connection/speeds, headphones, situations at home, etc.,) that is inherent to our study considering that it is not hosted in a controlled laboratory setting.

Once the aforementioned conditions were agreed upon by our subjects, they would be sent a link to the website that hosted our study (taskb.audiocat.net). Before starting, we would debrief them about what the study consists of and what is expected of them. Some of this debriefing was done over the phone and some through text. They were never given hints about what was the feature differentiating the two kinds of stimuli. They were told that the study would be divided in three sections (pairwise similarity task, followed by the categorization task and finishing with a similarity task) and that for the similarity tasks they must differentiate between the two sound-

sequences that will be played back-to-back, and for the categorization trials they will have to select whether a sound-sequence belonged to one category or the other. Once our subjects were ready to participate, they were asked some questions (on the first page of the website) about the equipment they were using (e.g., were their earphones wireless or wired). Afterwards, they read a consent form and accepted it before starting.

Task

The study was split into three phases: an initial pairwise similarity rating task, followed by a learning task of 400 trials, and concluding with a second pairwise similarity rating task. Before beginning the first pairwise similarity task, subjects were asked to complete 6 practice trials that used stimuli that were neither Kalaphones (K) nor Lakaphones (L). This allowed the subjects to become accustomed to the web interface and get a sense of what to expect.

Once practice trials were done, the website started the first phase, the similarity rating. This task had 40 trials where subjects would listen to two 12-bit/3-chunk sound strings played one after the other in rapid succession (with a small pause in between) and were asked to rate how similar they were, on a scale from 0-9, 9 meaning the sounds were completely identical, and 0 being they are completely unalike. The 40 strings were split into four pairs of 10: 10 were K-K pairs, 10 were L-L pairs and the remaining 20 were K-L/L-K pairs. The K-L comparisons were equally split between trials where the K played first and the L played first to ensure that there was no bias caused by hearing one category first.

After the first similarity rating was completed, there were 400 supervised category learning trials, with corrective feedback. Subjects would hear a sound-sequence and could press a key to respond that it was a LAKAPHONE (L) or a KALAPHONE (K). Feedback was given indicating whether their response was correct or incorrect. Using a keyboard instead of a mouse

click (or any other measure) helps standardize the learning method and has been used in previous studies (Pérez-Gay et al., 2017). The learning trials consisted of 200 K strings 200 with L strings. The Ks and Ls were randomly distributed across the 400 trials, and each subject got a different variation of the randomization (i.e., subject 1's randomly distributed K or Ls was not the same as subject 2's distribution, and so on). The 400 learning trials were broken up into 4 blocks of 100 trials each. After each block, subjects were given the opportunity to take an optional 5-minute break before starting the next block.

After subjects finished the category learning trials, they completed the second pairwise similarity judgment task. The 40 pairs were identical, but their order was shuffled with a different randomization.

Data collection

Data was collected throughout the study in real-time and compiled in an excel file once the subject had completed the task. Real-time data collection was made possible with a TensorFlow library generated by a JavaScript. If subjects did not complete the study for whatever reason, their data would be classified in a folder called "incomplete" and were not used in our analysis.

Generation of stimuli

Our sound-sequences were built using voiced "dee" and "dah" bits and were played at a speed of 70ms each. One sound-sequence consisted of a string of 12 "dee"'s and "dah"'s playing one after the other. The time separating between each bit was 25ms, for a total length of 1140ms. The stimuli were separable into categories based on the presence/absence of a "dee-dee-dah-dah" (Kalaphone) or "dah-dah-dee-dee" (Lakaphone) playing at any location in our sound-sequence

(e.g., XXXX-dah-dah-dee-dee -XXXX, and dee-dee-dah-dah -XXXXXXXX are both Ks; the same would be true for an L). Having our sound-sequence 12-bits long was not an arbitrary decision as having 12 bits allowed us to generate 240 Ks and Ls respectively; 200 for the categorization task and 40 for similarity judgments.

Before launching our pilot study, the stimuli were tested extensively between the team and a couple of volunteers in order to determine whether the categories were learnable. Our initial subjects reported that the stimuli were too confusing and as a result, we had to change the structure of our stimuli to make them easier. The most common response was that the string sounded too holistic and it was virtually impossible to discern which stimuli was a K or L. Consequently, we were forced to split our 12-bit strings into 3, 4-bit chunks. The inspiration for this came from Miller's (1956) observation that subjects could retain a longer bit-string if the bits were combined into larger groups ("chunks") in their working memory instead of trying to memorize them as a single string (e.g., it's easier to remember 456798235 as 456-798-235 as the grouping (chunks) simplifies the demands on short-term memory). Chunking is not exclusive to numbers and language and has been observed in music (Godøy et al., 2010), as well as in the study of acoustics and rhythm (Teng et al., 2018).

These new stimuli were largely similar to the old ones, with a few minor changes that had to be addressed, and tested rigorously. The similarities between our "chunked" strings and "non-chunked" ones are that the length of the individual bits stayed the same (70ms), and we could still keep our Ks and Ls mutually exclusive. (This will be expanded upon shortly.) The most obvious change is that our stimuli are now composed of three chunks of 4-bits (which looks like XXXX-XXXX-XXXX), compared to our non-chunked stimuli. This change in structure added two extra challenges with how we would calibrate our stimuli: (1) how many ms should separate

each bit played within 1 chunk (i.e., how fast should the "dee/dah"'s come after each other; let us call this the "within-chunk" rate) and (2) how long would the delay be between each chunk (i.e., the "between-chunk" rate). After months of testing, it was decided that the within-chunk rate would be 25ms and that the appropriate between-chunk rate would be 300ms. This made our sound-sequences a lot longer, as we went from our original 12-bit sequence lasting a total of 1140ms to 2040ms with the chunked stimuli. As with our old stimuli, only one of the three chunks was the relevant feature (e.g., XXXX-dee-dee-dah-dah-XXXX is a K) and it was positioned randomly and equally across all three chunks, to avoid a position bias.

Measures

Learners

As a criterion for successful learning we used at least 80% correct responses in the last 100 trials, as previously used by Pérez-Gay et al. (2017). With this our subjects could be classified as "Learners" and "Non-Learners".

Results

Learning

Thirty-five subjects successfully completed the category-learning and similarity rating tasks. Only four met our 80% "learner" criterion; the rest were "non-learners". The last block performance of our learners varies from 86% to 98% correct. It is likely that with an extra 200 trials, we may have had more learners; as 3 subjects had 70-75% correct trials in the last block, with another 4 more scoring between 65 and 69% correct. Moreover, it is important to note that all four of the learners are currently in Dr. Harnad's lab, and nine members from the lab participated in our study (Figure 1). It is possible that only the lab-members met the 80%

criterion due to their familiarity with the subject. Removing the nine lab members left an N of 26 but there were still several that were above chance so we went to test whether learning performance was correlated with CP. But first we will describe the measure of CP.

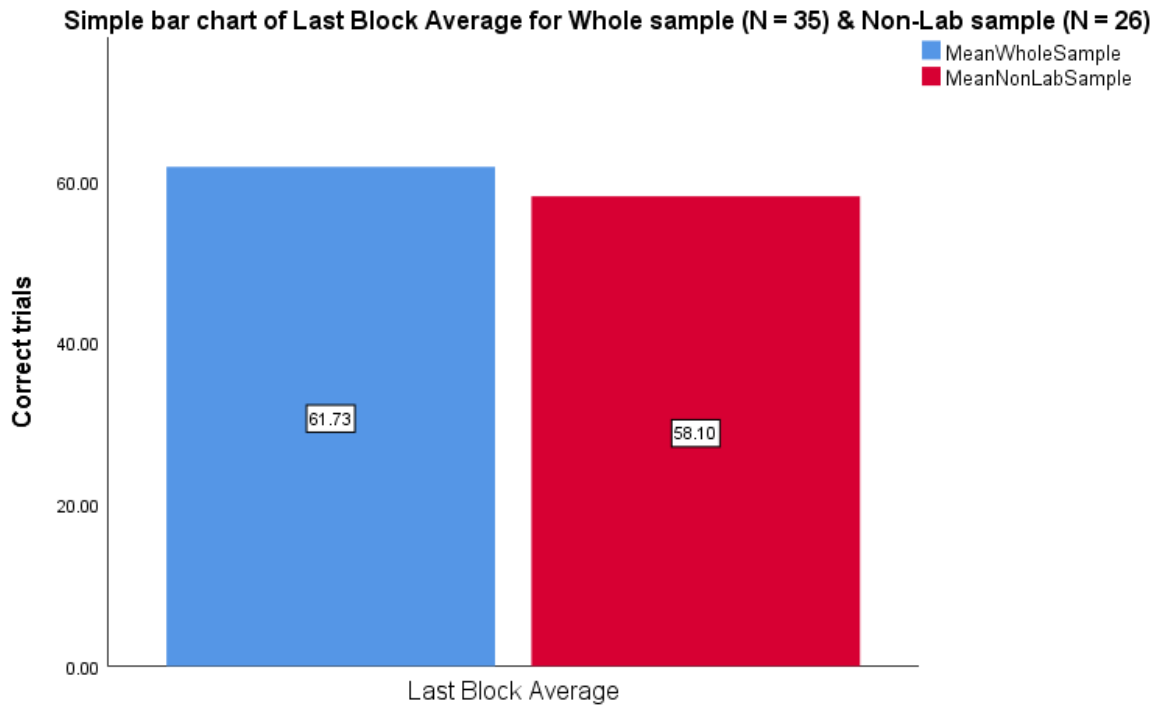

Figure 1. Mean number of correct trials in the last block for the whole sample (*N* = 35; blue bar) and for the non-lab member sample (*N* = 26; red bar). The average correct trials on the last block for the whole sample (blue bar) is *M* = 61.73, *SE* = .023, and for the non-lab member sample (red bar) is *M* = 58.1, *SE* = .019.

Similarity Judgments

To test for CP (between-category separation and within-category compression), we calculated the average similarity ratings. These were calculated for both between-category (B) and within-category (W) pairs, before (pre) and after (post) learning. This created four variables: "Bpre", "Bpost", "Wpre", and "Wpost". Further calculations were done to measure the change in

similarity judgments for both the between-category (diffB = Bpost – Bpre), and within-category (diffW = Wpost – Wpre) pairs. A between-category separation would be indicated by a positive "diffB" value, as the members of different categories are perceived as more different after category learning. Within-category compression would result in a negative "diffW" value, as the members of the same category are perceived as more similar after category learning. To combine separation and compression into a single CP score, we used "Global CP" (GCP): GCP = diffB - diffW. If diffB is positive, and diffW is negative, GCP increases. If diffB is negative (i.e., between-category compression instead of separation), or diffW is positive (i.e., between-category compression instead of compression), GCP decreases.

Many correlational analyses were done to test whether learning produces CP. The first was the correlation between the number of correct trials in block 4 and GCP. The correlation was positive and significant, $r = .599$, $N = 35$, $p = .001$. The greater the learning, the greater the GCP.

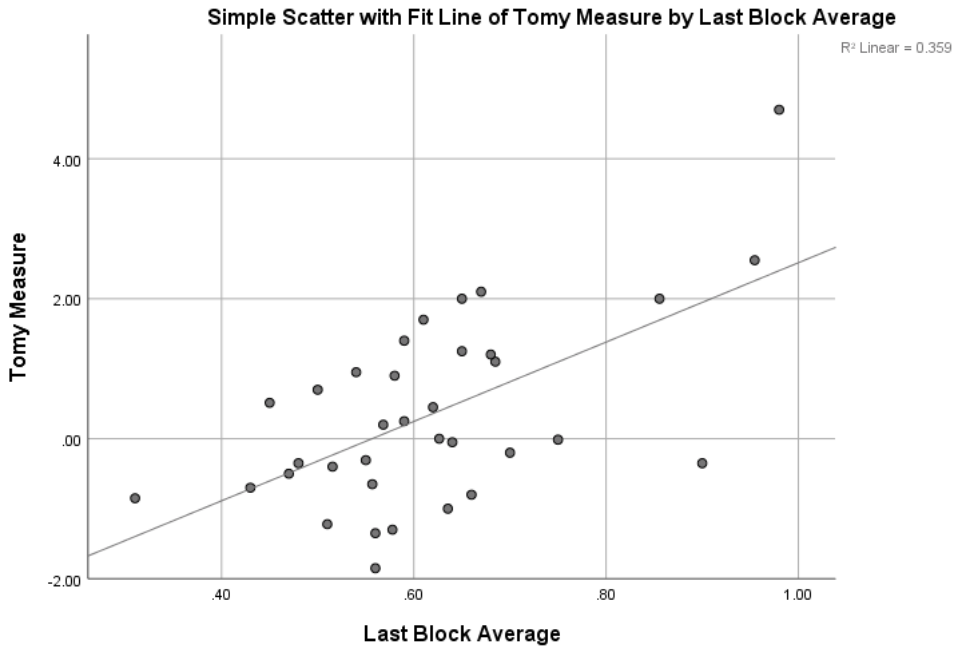

Figure 2. Correlation between percent correct in block 4 and GCP for all 35 subjects (r = .599, N = 35, p = .001).

Because all the learners (N = 4) (80% correct or higher) came from Dr. Harnad's lab, we removed all 9 lab-members to see whether there was still a positive correlation between percent correct and GCP. The correlation was still positive, but not significant, r = .398, N = 26, p = .062

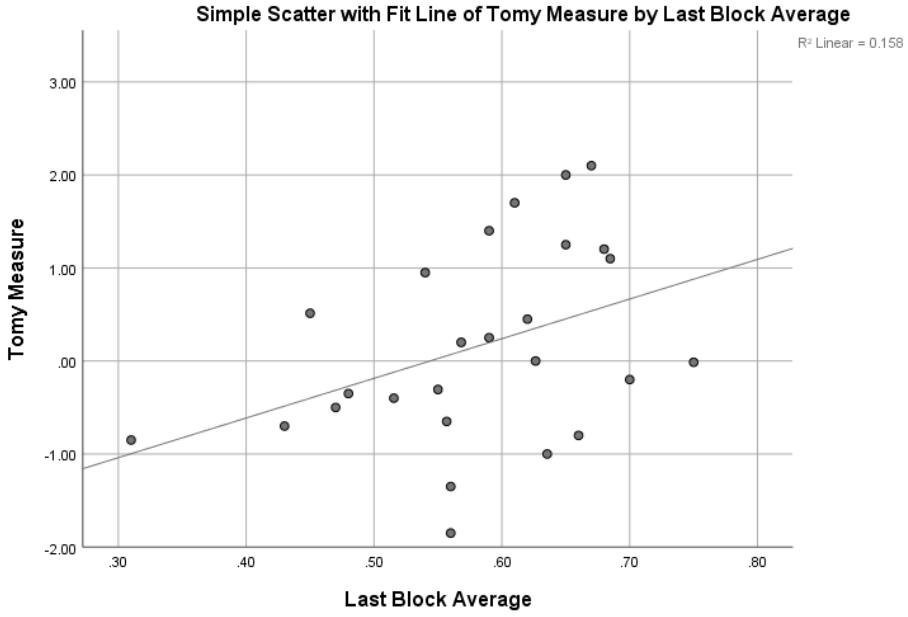

Figure 3. Correlation between percent correct in block 4 and GCP for the 9 non-members of the lab (r = .398, N = 26, p = 0.62 NS).

Of the four learners who attained 80% correct in isolation, one did not show a CP effect. This subject's final B/W scores show almost no difference in similarity ratings before and after learning. Figure 4 & 5)

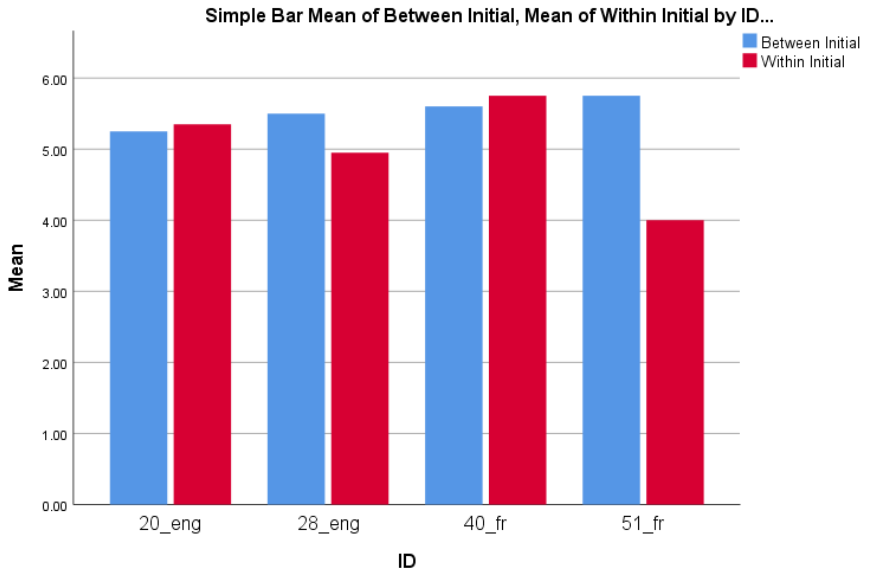

Figure 4. Similarity ratings between (blue bar) and within (red bar) categories before the 400 training trials for the four Ss who eventually reached 80% correct by the end of the training trials. Similarity was about the same for between and within categories. Average similarity judgment scores for between (blue bar) were $M = 5.5$, $SE = .10$, and for within (red bar) was $M = 5$, $SE = .37$.

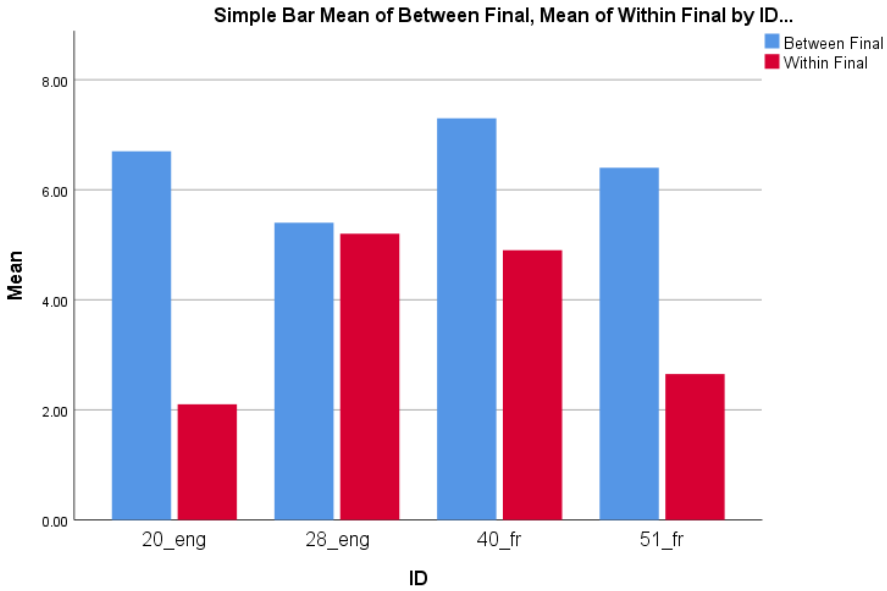

Figure 5. Similarity ratings between (blue bar) and within (red bar) categories after the 400 training trials for the four Ss who eventually reached 80% correct by the end of the training trials. Three of the subjects show CP, one (28_eng) does not. Average similarity judgment scores for between (blue bar) were $M = 6.5$, $SE = .39$, and for within (red bar) was $M = 3.7$, $SE = .78$.

Supplemental Analyses

Position Bias

After completing our analysis of categorization performance and its relation to CP effects, we analyzed whether our stimuli or the subjects' performance showed any chunk biases. We checked for (1) position bias in the stimuli by calculating whether there was any inequality in the distribution of stimuli in which the distinguishing feature appeared in chunk C1, C2 and C3, (it should be 1/3 for each chunk). We also checked for (2) category bias by calculating whether there was any inequality in the distribution of K stimuli and L stimuli for C1, C2 and C3 (it should be 1/2 for each chunk). (1) and (2) would be stimulus biases. We also checked for (3) performance bias by calculating whether there was any inequality in the distribution of the proportion of correct responses for all the stimuli when the distinguishing feature was in chunk 1, 2 or 3 (it should be 1/3 for each chunk). (We also checked (3) separately for K stimuli and L stimuli biases). The Chi-Square test was used to test departures from predicted proportions.

For the two potential stimulus biases ((1) position bias and (2) category bias), no Chi-Squares were statistically significant, hence our Ks and Ls were equally distributed in C1/2/3 for all 400 trials, and in each block.

There were two instances of subject bias. The first was that for the whole sample (N = 35), subjects performed significantly better on C1 compared to C2, X^2 (504, N = 35) = 566.25, p = .028 (Figure 6).

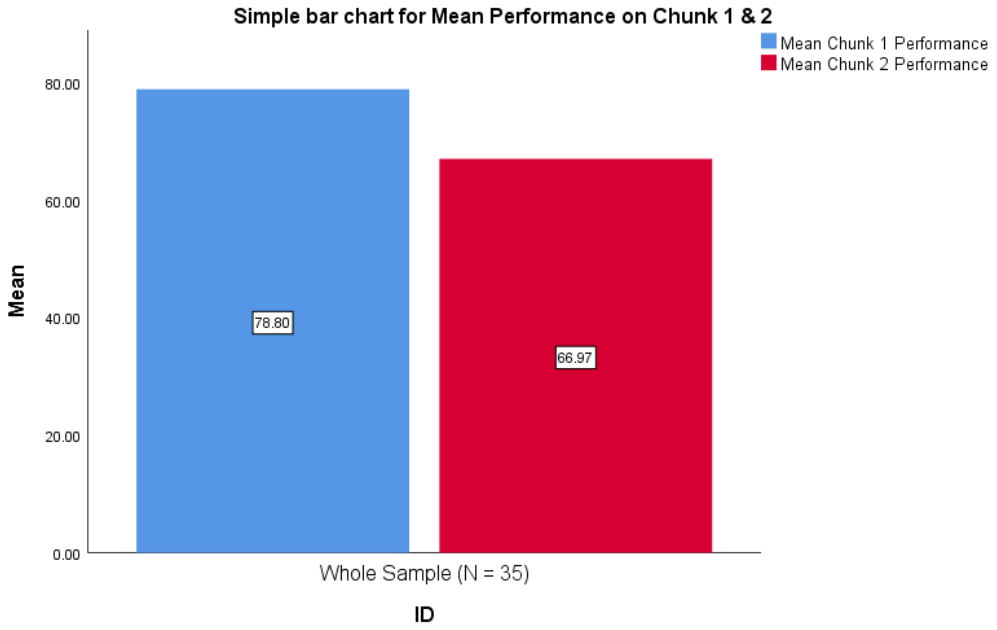

Figure 6. Performance scores for the whole sample (N = 35) for Chunk 1 (blue bar) & 2 (red bar) across 400 trials. The mean number of correct trials for Chunk 1 was M = 78.80, SE = 2.95, and for Chunk 2 was M = 66.97, SE = 1.76.

The second performance bias we found was when we evaluated the sample without lab-members (N = 26), our subjects performed better when K's distinguishing feature was on C2 (C2K), than when L's distinguishing feature was on C2 (C2L), X^2 (144, N = 26) = 185.727, p = .011 (Figure 7).

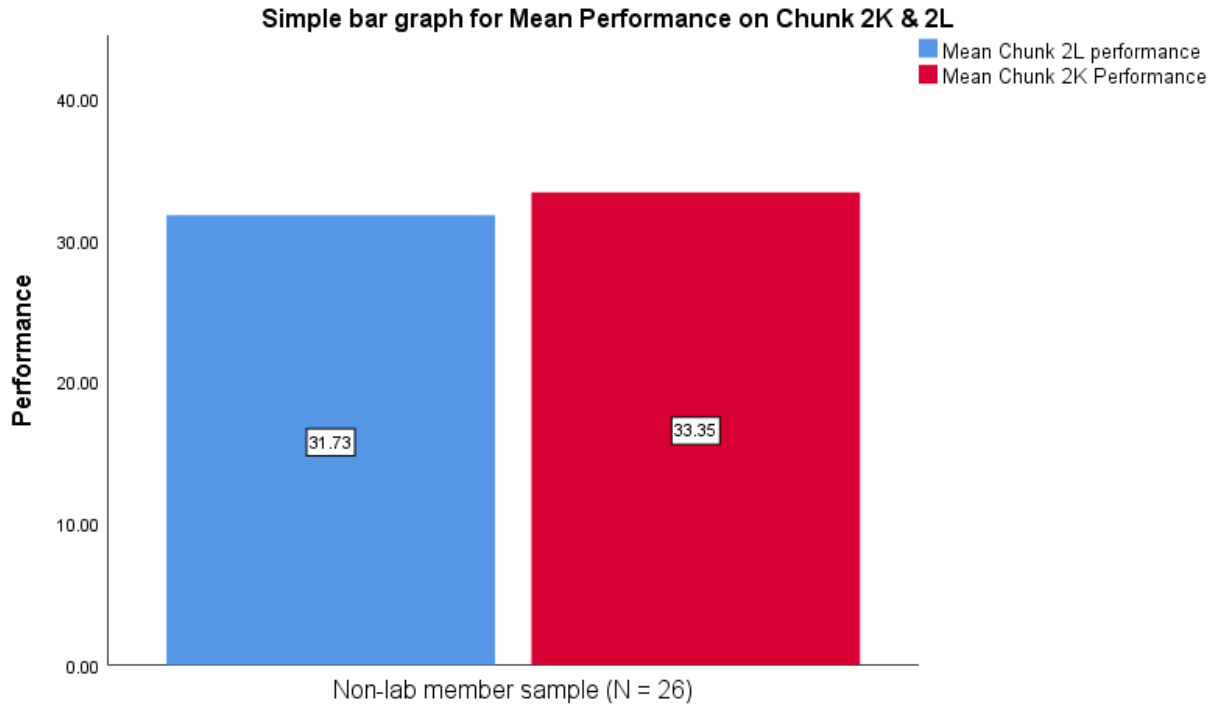

Figure 7. Performance scores for the non-lab members sample (*N* = 26) for Chunk 2K (red bar) and Chunk 2L (blue bar) across 400 trials. The mean number of correct trials for Chunk 2K is *M* = 33.35, *SE* = .93. and for Chunk 2L is *M* = 31.73, *SE* = 1.08

Discussion

Learnability of our stimuli

This pilot study indicated that the sound-sequence categories we generated are learnable and that learning them produces CP. Only four subjects (11%) could reach our success criterion of 80% correct in the last learning block (and only 3/4 showed CP), but the correlation between the percent correct and (global) CP across all 35 subjects was positive and significant (and there still remained a near-significant positive correlation when we excluded the 9 lab members (including the four 80%-learners) from the sample). This suggests that with a larger sample size a significant effect would be likely. A potential explanation for CP is that the subjects (or their

brains) produce “dimensionality reduction” as they learn the categories: they learn to pay selective attention to only the category distinguishing feature and ignore the rest (Dupont et al., 2013; Pérez-Gay et al., 2017).

There are many potential explanations as to why only a few select subjects are attaining our 80% success criterion. Our categories may be a little too difficult to learn in 400 trials. After subjects have completed the study, many of them reported back to us that the stimuli were too difficult and confusing and that it was almost impossible to differentiate a K from an L. While self-report can be unreliable at times (Althubaiti, 2016), it is nonetheless important to take the feedback we get from our volunteers seriously; the statistics do indeed reflect their difficulty in learning. Previous studies evaluating category-learning and difficulty have emphasized that if there are more trials added to difficult categories, learning is more likely to occur (Pérez-Gay et al., 2017; Véronneau et al., 2020). While it is possible that there might be a “lab-member bias” in our sample, it is doubtful that this bias is a major reason why non-lab subjects didn’t reach the criterion. It is important to keep in mind that all subjects completed our study in their homes, using their own equipment. This means that we have no control over what goes on in their personal living spaces (distractions could arise unexpectedly, and this can add noise to our data and hinder our subjects’ ability to learn). We also do not have any control as to what equipment subjects use to participate (some subjects could have partially defective earphones, an unstable internet connection, etc.,). All these factors would normally be controlled and minimize variance in a laboratory setting.

Potential stimulus and response biases

Two potential stimulus biases are position bias in how often the category-distinguishing feature appears in chunk 1, 2 or 3 across all 400 trials (should be 1/3 each) and category bias in

how often chunk 1, 2, or 3 was part of a K or an L overall and per block (should be 1/2 each). We also tested for a performance bias in the number of correct trials depending on whether the category-distinguishing chunk appeared in chunk 1, 2 or 3. We found no evidence for either position or category bias, therefore any performance bias when the distinguishing feature was in chunk 1, 2 or 3 would be perceptual rather than a consequence of unbalanced stimuli.

We did find two statistically significant instances of performance bias in our sample's performance. The first was that our subjects (N = 35) had a significantly higher percentage correct when the distinguishing feature (whether for Ks or Ls) was in chunk 1 compared to chunk 2. Subjects were more likely categorize a K or L correctly if the distinguishing feature was at the beginning, rather than the middle of our sound-sequence. Many of our subjects did report that the stimuli sounded holistic and that the middle chunk was the hardest to decipher because subjects were focused on how the first chunk sounded and could not adequately focus on the second chunk. With that taken into consideration, this could be a classic case of the anchoring bias (the first piece of information presented to you is more salient than the information that comes afterwards; Tversky & Kahneman, 1974). The second performance bias was only in the non-lab member's sample (N = 26). In this case, subjects scored more correct trials when the distinguishing feature for K was in Chunk 2 (C2K) compared to when the distinguishing feature for L was in Chunk 2 (C2L). The cause is less evident, and perhaps despite the significant Chi-Square this was not a robust effect, particularly as the overall performance was closer to chance in the non-lab sample.

Analyzing for any bias (whether stimulus or response) in this pilot study may help in designing more balanced stimuli for the follow-up full-scale experiment. The bias analyses could also be done separately for each of the four 100-trial learning blocks.

Preliminary conclusion

Bringing all this together, our findings demonstrate a classic CP effect for serial auditory pattern categories that have now been replicated many times for visual categories (Pérez-Gay et al., 2017; Andrews et al., 2015). Our pilot results demonstrate that there is a positive correlation between how well a category is learned and the size of the perceptual change called CP. The present study contributes to the existing literature on auditory categorical perception, while simultaneously expanding on which type of sounds can be used to induce category learning and CP.

Limitations

Our pilot study had several limitations. First, if we remove the 9 lab members that participated in our study (who were not naïve to what a categorization/categorical perception study entails), we lose statistical significance (and all four of our 80% learners). A plausible reason for this is that the lab members, with their previous experience/knowledge of what we are studying (categorization and CP), were better "prepared" for what the study entails, hence less likely to be deterred by the difficulty of the learning task. (Their knowledge would not, however, clue them in any way to what the category-distinguishing feature was). Second, the inclusion/exclusion criteria of our study were respected for the most part, but we still had subjects that would normally have been excluded if this had been a paid study (i.e., any with musical experience, or above the age of 55). Third, our experiment was not done in a laboratory setting (due to the COVID-19 pandemic). This means that the equipment people used was highly variable, introducing an unknown amount of noise and variance into our results. Last, it is possible that our stimuli were simply too hard to learn for the majority of our sample. Many of

them reported that they had no idea what the category differences were, and that they were guessing for most of it.

The follow-up to this pilot study will have to take these variables into account to reduce these problems, including conducting it in the laboratory rather than just online.

Conclusion

This pilot study investigated the effect of category learning on the perception of auditory stimuli. The categories did turn out to be learnable and the learning did generate a global CP effect, but there was not enough successful learning among the non-lab-member subjects. Further analyses on our sample's level of CP were conducted to test whether there was any bias in our stimuli or our subjects' performance. Because all four of our learners were members of Dr. Harnad's lab, we re-analyzed our sample after removing all nine lab members from our sample of 35 to see whether there was still an effect in the remaining 26. The positive correlation between learning and perception was there, but it no longer reached statistical significance. More measures to evaluate bias had to be considered to ensure that our Kalaphones (K) and Lakaphones (L) were equally and randomly distributed for each subject. Two instances were found where performance was skewed towards one chunk over the other. The first was where our entire sample scored more correct trials in chunk 1 over chunk 2, and the other one was when the non-lab member sample scored more correct trials in chunk 2K over chunk 2L.

These results indicate that our stimuli are learnable, but that subsequent studies may have to adjust the difficulty to make the learning less difficult. Either by calibrating presentation speed between or within-chunk speeds, or by adding another 200 trials to allow our subjects to further

familiarize themselves with our stimuli. There will also be vocal imitation of the stimuli in later experiments, and this too will slow down presentation rate.

Appendix A

Statement of contribution

The current project was built entirely from scratch. As previously stated, this is a pilot project to properly calibrate morse-code-like sequences for a subsequent study on verbal imitation and category learning. All the data collected in this study are new, as well as the stimuli that were generated.

During the whole process, I have received extensive help and guidance from my supervisor, Dr. Stevan Harnad, his graduate student, Michel Mercier, and fellow undergraduate, Ada Yetiş – for which I am very grateful. Throughout the project, I was helping with the data analysis, subject recruitment and keeping up with the literature, as our project is investigating a relatively new area of auditory categorical perception. The paper was written up entirely by me, and was subsequently reviewed and edited by Dr. Harnad before it was submitted. Outside of my research paper, I was involved in several meetings with Dr. Harnad's lab (Laboratoire de Cognition et Communication), to discuss the findings of our project and how we can use it for future studies (on top of the weekly meetings between Dr. Harnad, Michel, Ada, and myself). Furthermore, I was very fortunate to receive a lot of coding assistance from Michel and Ada during this whole process, as I am a naïve coder, and our project relied heavily on the knowledge of basic coding skills in order to properly analyze the data.

References

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., & Lang, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *The Journal of the Acoustical Society of America*, *101*(2), 1090-1105.

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, *9*, 211.

Andrews, J. K., de Leeuw, J., Larson, C., & Xu, X. (2017). A Preliminary P-Curve Meta-Analysis of Learned Categorical Perception Research. In *CogSci*.

Blondin-Massé, A., Harnad, S., Picard, O., & St-Louis, B. (2013). Symbol grounding and the origin of language: From show to tell. In *C. Lefebvre, B. Combir, & H. Cohen (Eds.), Current perspective on the origins of language*. John Benjamins Publishing Company.

Dupont, S., Ravet, T., Picard-Limpens, C., & Frisson, C. (2013, July). Nonlinear dimensionality reduction approaches applied to music and textural sounds. In *2013 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.

Ell, S. W., Ashby, F. G., & Hutchinson, S. (2012). Unsupervised category learning with integral-dimension stimuli. *Quarterly Journal of Experimental Psychology*, *65*(8), 1537-1562.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227-240.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814-823.

Gibson, J. J. (1979). The theory of affordances. The ecological approach to visual perception.

Godøy, R. I., Jensenius, A. R., & Nymoen, K. (2010). Chunking in music by coarticulation. *Acta Acustica united with Acustica*, *96*(4), 690-700.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69-78.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *The Journal of the Acoustical Society of America*, *106*(5), 2900-2912.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. *Categorical perception: The groundwork of cognition*, 1-52.

Harnad, S. (2003). Categorical perception.

Harnad, S. (2017). To cognize is to categorize: cognition is categorization. In *Handbook of categorization in cognitive science* (pp. 21-54). Elsevier.

Jacobs, G. (2013). *Comparative color vision*. Elsevier.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, *25*(1), 322-349.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets?. *Memory & Cognition*, *6*(4), 462-472.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual review of psychology*, *32*(1), 89-115.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition*, *95*(2), B1-B14.

Pérez-Gay, F., Christian, T., Gregory, M., Sabri, H., Harnad, S., & Rivas, D. (2017). How and why does category learning cause categorical perception?. *International journal of comparative psychology*, *30*.

Sauter, D. A., LeGuen, O., & Haun, D. (2011). Categorical perception of emotional facial expressions does not require lexical categories. *Emotion*, *11*(6), 1479.

Teng, X., Tian, X., Doelling, K., & Poeppel, D. (2018). Theta band oscillations reflect more than entrainment: behavioral and neural evidence demonstrates an active chunking process. *European Journal of Neuroscience*, *48*(8), 2770-2782.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Véronneau, M., Sicotte, T., & Harnad, S. (2020). The effect of learning and overlearning on acquired categorical perception. [Unpublished doctoral dissertation]. UQAM