Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry

Jacie R. McHaney\textsuperscript{a,1}, Rachel Tessmer\textsuperscript{b,1}, Casey L. Roark\textsuperscript{a,c}, Bharath Chandrasekaran\textsuperscript{a,7}

\textsuperscript{a} Department of Communication Science and Disorders, University of Pittsburgh, United States
\textsuperscript{b} Department of Speech, Language, and Hearing Sciences, University of Texas at Austin, United States
\textsuperscript{c} Center for the Neural Basis of Cognition, Pittsburgh, PA, United States

ARTICLE INFO

Keywords:
- Speech category learning
- Working memory
- Pupillometry

ABSTRACT

Across two experiments, we examine the relationship between individual differences in working memory (WM) and the acquisition of non-native speech categories in adulthood. While WM is associated with individual differences in a variety of learning tasks, successful acquisition of speech categories is argued to be contingent on WM-independent procedural-learning mechanisms. Thus, the role of WM in speech category learning is unclear. In Experiment 1, we show that individuals with higher WM acquire non-native speech categories faster and to a greater extent than those with lower WM. In Experiment 2, we replicate these results and show that individuals with higher WM use more optimal, procedural-based learning strategies and demonstrate more distinct speech-evoked pupillary responses for correct relative to incorrect trials. We propose that higher WM may allow for greater stimulus-related attention, resulting in more robust representations and optimal learning strategies. We discuss implications for neurobiological models of speech category learning.

1. Introduction

Adults struggle to learn speech sound contrasts that do not exist in their native language (Flege, 1995; Holt & Lotto, 2008; Iverson et al., 2003; Vallabha et al., 2007). While adults can acquire non-native speech sound categories with training (Flege, 1995; Holt & Lotto, 2006, 2010; Iverson et al., 2003; Vallabha et al., 2007; Wang et al., 1999; Wong & Perrachione, 2007), there is a great deal of individual variability in successful acquisition of speech categories in adulthood (Chandrasekaran et al., 2015). The individual variability may be driven in part by individual differences in working memory (WM) and executive attention (EA; Kidd et al., 2018). Operationally, WM refers to the ability to represent and manipulate information in a buffer that is consciously accessible (Engle & Kane, 2004; Tsukahara et al., 2016). EA provides the scaffolding for this operation by enabling the maintenance of information and preventing interference from external sources (Engle & Kane, 2004). Tasks like the operation span (OSPAN; Unsworth et al., 2005) can be used to measure both WM and EA. WM and EA are two cognitive processes that work together synergistically and can be difficult to tease apart in an experimental setting. Thus, we collectively refer to both processes as WM, keeping in mind that EA interacts with WM.

Within a population, there is consistent variation in WM that systematically relates to the successful acquisition of various components of language (Kidd et al., 2018). However, the precise role of WM in mediating individual differences in speech category acquisition has not been established. Across two experiments, we examine the extent to which individual differences in WM relate to successful acquisition of a non-native speech contrast (Mandarin tones). The first experiment is a large-scale study conducted online (n = 196 participants) where we examined the role of WM in speech category learning. In a second experiment conducted in-person (n = 28 participants), we leveraged pupillometry and computational modeling to specify mechanisms and computational strategies driving a potential WM advantage in speech category learning success.

Despite the individual differences in learning success, most adults can acquire even perceptually subtle non-native contrasts with training. Explicit training regimens that have engendered significant, generalizable learning typically incorporate three characteristics: natural speech productions, talker and contextual-variability, and feedback (Bradlow, 2008). The use of natural speech productions, multiple talkers, and multiple contexts are beneficial because they expose learners to the natural variability in the learning environment, allowing learners to

\textsuperscript{1} Contributed equally.

https://doi.org/10.1016/j.bandl.2021.105010

Received 8 January 2021; Received in revised form 26 July 2021; Accepted 10 August 2021
Available online 25 August 2021

0093-934X/© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license

http://creativecommons.org/licenses/by-nc-nd/4.0/
focus on and weight salient dimensions that vary less across talkers (Bradlow, 2008; Bradlow & Bent, 2008). Listeners may use feedback to monitor errors, fine-tune their learning strategies to focus on relevant information, and/or ignore irrelevant information. However, the extent and quality of feedback required is unclear. While the presence of some amount of feedback confers benefits to adult learners (Chandrasekaran et al., 2015; Chandrasekaran, Yi, et al., 2014; Lim & Holt, 2011; McClelland et al., 2002; Tricomi et al., 2006; Yi et al., 2016), adults can acquire speech categories incidentally (Gabay et al., 2015; Lim et al., 2019; Luthra et al., 2019; Roark & Holt, 2018) or with task-irrelevant feedback (Goudbeek et al., 2008; McClelland et al., 2002). However, explicit feedback has been shown to support greater perceptual learning, guiding mappings between acoustic patterns and phonetic categories (Lehet et al., 2020).

In explicit training regimens, individual differences in speech category learning success could arise from individual variation in the ability to select relevant dimensions and to ignore irrelevant or distracting dimensions. Individuals with lower WM demonstrate a relatively reduced ability to attend to and select task-appropriate dimensions (D’Esposito & Postle, 2015; Unsworth & Robison, 2017b; Wöstmann & Oehlsen, 2016). A recent framework posits that individual differences in WM are driven by moment-to-moment fluctuations in locus coeruleus-norepinephrine (LC-NE) system activity (Unsworth & Robison, 2017a). Individuals with lower WM may have greater attentional fluctuations that are likely caused by a disruption in LC-NE functioning when compared to individuals with higher WM (Unsworth & Robison, 2017a). The LC-NE system receives reciprocal input from the prefrontal cortex, such that these reciprocal projections can modulate attention for salient and irrelevant events (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003). Thus, individuals with lower WM may show reduced attention to salient dimensions relative to those with higher WM, resulting in less effective weighting of task-relevant dimensions during category learning.

In addition to attention to relevant dimensions, individual differences in speech category learning success during explicit training regimens could also arise from individual variation in the use of corrective feedback to guide appropriate learning strategies. Individual differences in WM may impact the selection and maintenance of learning strategies during training. Adapted from the Competition between Verbal and Implicit Systems (COVIS) model of multiple learning systems in the visual domain (Ashby et al., 1998), the dual-learning systems (DLS) framework posits that at least two competitive learning systems mediate feedback-based learning of novel speech categories in adulthood (Chandrasekaran, Koslov, et al., 2014; Chandrasekaran, Yi, et al., 2014; Feng, Yi, et al., 2018; Maddox & Chandrasekaran, 2014). The reflective learning system, involving the prefrontal cortex and the hippocampus, uses WM (Decaro et al., 2008; Filoteo et al., 2010; Miles et al., 2014; Reetzke et al., 2016; Zeithamova & Maddox, 2006) to develop and test verbalizable rules based on feedback (Maddox & Ashby, 2004). In contrast, the reflexive learning system associates perception with rewarding actions derived from feedback (Maddox & Chandrasekaran, 2014; Seger, 2008; Seger & Miller, 2010). The reflexive learning system requires pre-decisional integration across perceptual dimensions for optimal categorization and is thought to operate independently of WM (Ashby & Gott, 1988; Ashby & Maddox, 2011).

The dissociation of WM involvement posited by the DLS model may be too simplistic, as memory systems including WM do not have neatly dissociable roles in human category learning (Ashby & Valentin, 2017). The complex interplay of WM, attention, and perceptual encoding involved in learning new categories is not perfectly accounted for by any singular model (Roark & Holt, 2019a). There also remains debate about the nature of WM and whether it is a discrete capacity or a resource that can be distributed across multiple memory representations (Ma et al., 2014; Suchow et al., 2014). Individual differences in category learning success could be due to differences in WM that impact constraints on cognitive resources, interference, and/or decaying sound representations (Lloyd et al., 2019). While some studies have shown that procedural-based learning is impaired by increasing WM demands (Miles et al., 2014; Zeithamova & Maddox, 2006; but see Newell et al., 2010), others have demonstrated that the procedural-based learning system is facilitated by WM demands (Filoteo et al., 2010; but see Newell et al., 2013). Several studies have also found that WM does not disassociate between rule-based and procedural learning, and can, in fact, aid performance across both category structures (Kalish et al., 2017; Newell et al., 2010; Xing & Sun, 2017). The precise role and extent to which WM influences feedback-based speech category learning strategies is unclear and warrants continued investigation.

Mandarin tone categories are optimally learned through procedural strategies, as evidenced by the relationship between activation in the putamen during correct trials and learning success (Yi et al., 2016). Yi et al. (2016) demonstrated greater putamen activity in native English speakers using procedural learning strategies during Mandarin tone category learning and observed greater auditory- striatal functional connectivity patterns as a function of learning. WM is hypothesized to be associated with both the ability to acquire novel categories and the ability to switch between categorization strategies (Lewandowsky, 2011; Lloyd et al., 2019; Sewell & Lewandowsky, 2012). Prior studies have leveraged computational models to show that during the acquisition of non-native speech categories, successful learners tend to switch to procedural-based learning strategies, while less successful learners perseverate with suboptimal rule-based learning strategies (Chandrasekaran, Koslov, et al., 2014; Maddox & Chandrasekaran, 2014). A potential downside of higher WM is that individuals may perseverate with rule-based strategies when searching for rules in procedural learning conditions, which could slow learning speed (Ashby & Valentin, 2017; Kalish et al., 2017). In this scenario, individuals with lower WM may switch to procedural-based learning strategies, which are more independent of WM. This scenario would be consistent with findings that show that WM demands impair category learning in procedural learning conditions (Miles et al., 2014; Zeithamova & Maddox, 2006). In contrast, several studies have demonstrated that individuals with higher WM learn categories faster, and that individuals with lower WM, not higher WM, perseverate with suboptimal rule-based strategies in categorization tasks (Decaro et al., 2009; Tharp & Pickering, 2009). WM may influence or predict strategy choice in speech category learning, as found across other domains (Bailey et al., 2008; Craig & Lewandowsky, 2012; Dunlosky & Kane, 2007). As another alternative for the role of WM during category learning, Craig and Lewandowsky (2012) found that WM did not influence the strategy an individual chose during visual category learning. Instead, WM influenced how well an individual used the selected strategy.

The current study utilizes a combination of behavioral, physiological, and computational approaches to examine the extent to which individual differences in WM are associated with the acquisition of Mandarin tone categories by native English speakers. In Mandarin, pitch changes within a syllable can differentiate word meaning. Mandarin has four linguistically-relevant pitch contrasts, or tones, that function similarly to consonants and vowels in modifying word meaning. At least two pitch dimensions, relative pitch and pitch change, phonetically contrast the tone categories (Chandrasekaran et al., 2007; Gandour & Harshman, 1978). Adult English listeners often have difficulty learning Mandarin tone contrasts because these contrasts do not differentiate word meaning in English (Wang et al., 1999, 2003). Prior studies have demonstrated that neural representations of tone categories can emerge within a single session of sound-to-category training (Feng, Yi, et al., 2018; Yi et al., 2016). Specifically, Feng, Gan, et al. (2018) and Feng, Yi, et al. (2018) showed rapid emergent representations of tone categories in the left superior temporal gyrus within a few hundred training trials. Interestingly, during the second half of training, there was greater functional coupling between the left superior temporal gyrus and the striatum during incorrect responses, suggesting that representations are ‘fine-tuned’ as a function of corrective feedback. Successful learners
tended to show greater striatal activation as well as more robust, multidimensional category representations in the left superior temporal gyrus (Feng, Yi, et al., 2018).

In the first experiment, we predicted that WM differences in a large, highly diverse population of non-tonal language speakers would relate to individual differences in the successful acquisition of tone categories. In the second experiment, we replicate Experiment 1 in a smaller cohort with the addition of computational modeling and pupillometry to provide mechanistic insight into the role of WM in non-native speech category learning. Computational modeling approaches allowed us to go beyond accuracies and discern underlying neurobiological learning strategies. Pupillary responses serve as an indirect index of LC-NE activity (Alnæs et al., 2014; Aston-Jones & Cohen, 2005; Gabay et al., 2011; Murphy et al., 2014). Prior studies have shown distinct pupillary responses as a function of individual differences in WM (Heitz & Engle, 2007; Tsukahara et al., 2016). Pupillary measures during the process of speech category learning allowed us to discern the extent to which LC-NE activity differences between individuals with higher and lower WM underlies variability in speech category learning success.

In both experiments, we predicted that individuals with higher WM would show greater tone category learning accuracy and greater generalizability to untrained talkers relative to those with lower WM. This prediction aligns with findings from numerous studies that have shown that greater WM relates to better learning (Decaro et al., 2009; Filoteo et al., 2010; Kalish et al., 2017; Newell et al., 2010; Tharp & Pickering, 2009; Xing & Sun, 2017). Regarding strategy usage in Experiment 2, we predicted that individuals with higher WM would use more optimal procedural-based learning strategies during training than individuals with lower WM. This prediction is in line with prior computational modeling studies that found successful learners switched from rule-based to procedural-based learning strategies over the course of training (Chandrasekaran, Koslov, et al., 2014; Maddox & Chandrasekaran, 2014). Regarding pupillometric measures, we hypothesized that the switch from rule-based to procedural-based learning strategies would be discerned online via reduced pupillometric responses, reflecting the reduced cognitive effort associated with procedural-based learning strategies. Prior studies demonstrate that pupillary responses tend to be larger for incorrect trials than for correct trials during behavioral tasks (Braem et al., 2015; Gritchley et al., 2005). Aligning with the LC-NE hypotheses, we predicted that correct and incorrect trials would have similar stimulus-evoked pupillary responses in individuals with lower WM, while individuals with higher WM would have larger pupillary responses for incorrect trials compared to correct trials. Overall, we sought to provide novel mechanistic insights into an important variable (i.e., WM) that may underlie individual differences in speech category learning success.

2. Experiment 1

2.1. Methods

2.1.1. Participants

A total of 198 participants between the ages of 18–35 (99 females; mean age = 24.965, SD = 4.973) were recruited online using Prolific (www.prolific.co). Prolific provides access to participants from all over the world for a representative study sample. Participants first completed a language-history questionnaire to ensure they reported that they were native speakers of English with no prior experience with a tonal language (e.g., language courses, immersion experiences). Participants received monetary compensation for their participation. Two participants did not follow instructions on the OSPAN task assessing working memory (see Section 2.1.3) and were removed from data analysis. Removal of these two participants resulted in a final sample of 196 participants (98 females; mean age = 24.939, SD = 4.934). Sixty-three of the final 196 participants reported prior musical experience ranging 1–24 years (M = 9.825, SD = 6.155). While formal music training has previously been shown to enhance speech processing (Bidelman et al., 2011; Schön et al., 2004; Wong et al., 2007), the average years of music experience for the entire sample (M = 3.158, SD = 5.763) was less than the amount of experience shown to enhance performance on the speech category learning task (>10 years; Smayda et al., 2015). This research protocol was approved by the Institutional Review Board at the University of Pittsburgh.

2.1.2. Stimuli

Participants learned to categorize Mandarin tones that vary along relative pitch and pitch change dimensions. Four lexical tones were produced in five syllable contexts (/bu/, /di/, /ri/, /ma/, /mi/) by four native Mandarin Chinese speakers (two females; Fig. 1A; Feng, Gan, et al., 2018). The Mandarin tone stimuli were duration (440 ms) normalized using Praat (Boersma & Weenink, 2005). A scatterplot of the speech stimuli is displayed in Fig. 1B.

2.1.3. Procedure

The experiment was created and hosted using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were instructed to use headphones for the training and generalization tasks and were instructed to set the stimulus volume to a comfortable listening level in a quiet area without distractions. Each trial in the training task began with the presentation of the auditory stimulus for 440 ms, followed by a response prompt of “Which Category” where participants provided their category response (untimed). Corrective feedback in the form of “Correct” or “Wrong” was displayed for 750 ms immediately after the category response was provided, followed by a 1000 ms intertrial interval presented with a fixation cross on the screen (Fig. 1C). The training task consisted of six blocks of 10 trials of each of the four tone categories produced by two native Mandarin speakers (one female). Each block consisted of 40 trials for a total of 240 trials across the training task. Participants received a self-timed break at the end of each block. A generalization task immediately followed the training task. The generalization task consisted of a single block of 10 trials of each of the four tone categories (40 trials total) with no corrective feedback and used speech stimuli produced by two novel native Mandarin speakers (one female) that were not used in the training blocks.

All participants also completed the automated version of the OSPAN task (Unsworth et al., 2005). The OSPAN task has been widely used to measure WM and EA (e.g., Turner & Engle, 1989; Xie et al., 2015). Participants were shown simple arithmetic problems and asked to decide whether presented solutions to the problems were correct or incorrect. A letter was displayed on the screen after each problem. Following a series of arithmetic problems, participants were required to recall the letters that were displayed in the order that they appeared. The task consisted of 15 letter sequences that spanned three to seven letters (three repetitions of each span). If a participant correctly recalled all letters from a sequence, the span length was added to their score. For instance, if a participant recalled all letters in the correct order from a seven-letter span sequence, seven points were added to their score. The two participants who were removed from data analysis were the answer to the arithmetic problems, rather than the letters that were presented. Because the task was performed incorrectly, a score could not be calculated for these participants. The maximum possible score on the OSPAN task was 75 (Mdn = 45). Each participant’s OSPAN score was used as a measure of WM in the data analyses.

2.1.4. Behavioral data analysis

A binomial generalized linear mixed-effects model was fit to examine learning performance in the tone categorization training task as a function of WM using the lme4 package (Bates et al., 2015) in R (R Core Team, 2019) and the lmerTest package (Kuznetsova et al., 2017) to estimate p-values. The outcome variable was trial-by-trial accuracy (correct, incorrect) for each participant. The optimal model fit included fixed effects of trial, WM (i.e., individual participant OSPAN scores), and...
the interaction of trial and WM, with a random slope of trial for each subject and a random intercept of tone category: Accuracy ~ trial*WM + (trial|subject) + (1|tone category). This model provided a significantly lower Akaike information criteria (AIC) value and better fit than the model including only random intercepts of subject and tone category ($\chi^2(2) = 161.33, p < .001; \text{AIC}_{M1} = 7779.918; \text{AIC}_{M2} = 7622.592$).

We fit a similar binomial generalized linear mixed-effects model to examine categorization accuracy in the generalization block as a function of WM. However, this model differed in that the outcome variable was trial-by-block response outcomes in the generalization block and block 6 of the training task. Block 6 was used as the reference block to account for individual differences in performance at the end of training. A model including a random slope of trial for each subject and a random intercept of tone category, as in the above training task model, did not converge. The best fit model included fixed effects of block, WM, and the interaction of block and WM, with random intercepts of subject and tone category: Accuracy ~ block*WM + (1|subject) + (1|tone category).

Fig. 1. Experimental stimuli and training procedure. A) Spectrograms for four Mandarin tones (T1, T2, T3, T4) produced by two female and two male native speakers in five syllable contexts: /bu/, /di/, /lu/, /ma/, and /mi/. B) Scatterplot of all stimuli from the training and generalization blocks along the two acoustic dimensions of relative pitch and pitch change in arbitrary units (a.u.). C) Representative trial of the Mandarin tone training task procedure.

Fig. 2. Experiment 1 behavioral results. A) Distribution of scores on the OSPAN task for all 196 participants. B) Categorization accuracy across the six category training blocks and the generalization block (abbreviated as “Gen”) for higher ($n = 101$) and lower ($n = 95$) working memory (WM) participants, as defined by a median split ($\text{Mdn} = 45$) of OSPAN scores for visualization purposes. The dark, solid lines denote the average accuracy with error bars reflecting standard error of the mean for each WM group. The lighter lines and points represent each individual participant’s accuracies. Categorization accuracy is defined as the proportion of correct trials per block. The dashed line reflects chance-level performance (0.25).
3. Results

We first examined accuracy during the training task as a function of WM. There was a significant effect of WM on accuracy ($\beta = 0.161, z = 2.588, p = 0.010$), suggesting that participants with higher WM had greater overall accuracy in the category learning training task. Additionally, there was a positive and significant interaction of trial and WM ($\beta = 0.105, z = 3.437, p < 0.001$). These results suggest that not only did the participants with higher WM demonstrate greater categorization accuracy throughout the training task, but they also had a larger trial-by-trial improvement in categorization accuracy relative to participants with lower WM (Fig. 2B). These findings demonstrate that participants with higher WM scores learned non-native speech categories to a greater degree than participants with lower WM scores.

We also examined differences in categorization accuracy as a function of WM in the generalization block, where participants categorized stimuli produced by novel speakers without corrective feedback. There was a positive and significant effect of WM ($\beta = 0.282, z = 3.417, p < 0.001$), suggesting that accuracy was greater in both block 6 and the generalization block for participants with higher WM. However, we did not observe a significant interaction of block and WM ($\beta = 0.036, z = 0.998, p = 0.318$). This non-significant interaction suggests that, although categorization accuracy was greater for participants with higher WM scores, the change in accuracy from the final block of training to the generalization block did not differ based on WM status. These results suggest that WM status does not affect one’s ability to generalize learned categories to a context involving untrained talkers.

4. Experiment 2

Experiment 1 demonstrated significant differences in non-native speech category learning based on WM status. In Experiment 2, we probe the mechanisms underlying the WM advantage in learning tone categories. Experiment 2 takes a computational and mechanistic approach to examine the role of the LC-NE system, as indexed by stimulus-evoked pupillary responses, on WM involvement during non-native speech category learning.

3.1 Methods

3.1.1 Participants

Twenty-eight participants (21 females; mean age = 20.54, SD = 3.00) were recruited from The University of Texas at Austin and the greater Austin community for Experiment 2. All participants were native English speakers, had hearing thresholds < 25 dB for 250, 500, 1000, 2000, 4000, and 8000 Hz, and had less than six years of formal music training. Individuals with greater than six years of musical training were not included in this study as music experience has been shown to influence speech processing (Bidelman et al., 2011; Schon et al., 2004; Wong et al., 2007). Participants also completed a language-history questionnaire to ensure no prior tonal language experience (e.g., language courses, immersion experiences). Participants received either monetary compensation or research credit for their participation. This research protocol was approved by the Institutional Review Board at The University of Texas at Austin.

3.1.2 Procedure

The speech stimuli for the training task were identical to those used in Experiment 1 (see Section 2.1.2. Stimuli). In addition to the four Mandarin tone categories, we also included a fifth category of silence trials to examine the extent to which the observed pupillary responses were sound-evoked and not caused by other physiological responses (e.g., motor preparation). Silence trials were identical to sound trials, except participants were not presented with an auditory stimulus during the sound period of the trial. The addition of silence trials increased the number of trials per block to 50 (40 speech trials, 10 silence trials), for a total of 300 trials across six blocks of training. To rule out pupillary changes associated with a physical button-press during the category response period (McGinley et al., 2015), mapping of the category and key response was counterbalanced across participants, with a different category designated as a no-press response. For instance, some participants were instructed to not press a key if they believed the stimulus belonged in category 1, but to press keys ‘2’, ‘3’, and ‘4’ for categories 2, 3, and 4, respectively. All participants were instructed to press key ‘5’ to categorize silence trials. The training task was created and presented using Experiment Builder (SR Research), and the speech stimuli were presented via insert earphones (ER-3C, Etymotic Research, Elk Grove Village, IL).

Monocular left eye pupil size was monitored using an Eyelink 1000 Plus Desktop Mount (SR Research) with a chin and forehead rest for stabilization. Data were recorded at a sampling rate of 1000 Hz. Luminance of the visual field was controlled by setting consistent room lighting across all participants. Nine-point eye tracker calibration was performed prior to the start of the experiment. Consistent with procedures from Experiment 1, each trial began with a fixation cross, followed by the speech stimulus presentation, category response, and corrective feedback. However, we included delays after each trial event to allow for changes in pupil size (Kooi et al., 2018; Winn et al., 2018; Zekveld et al., 2013). Participants were required to fixate on the cross in the center of the screen for a minimum of 500 ms to begin each trial. An invisible boundary surrounded the fixation cross to restrict the area of the 500 ms criterion and consisted of the following visual angles: [0 – 45, 45 – 90, 90 – 135, 135 – 180, 180 – 225, 225 – 270, 270 – 315, 315 – 360, 360 – 45, -45 – 0]. This fixation criteria to begin each trial was implemented in an effort to control for the effects of saccades, which can alter pupil diameter, and to minimize pupil foreshortening errors (Hayes & Petrov, 2016; Knapen et al., 2016). The trial stimulus (i.e., sound or silence) was presented following a two-second delay after meeting the fixation criteria. There was a four-second delay from the onset of the trial stimulus to the category response prompt on the screen in the form of: “Which Category?”: Participants had two seconds to respond. Following the response, there was a two-second delay before corrective feedback was displayed on the screen in the form of “Correct” or “Wrong” for two seconds. At the end of each block, tone categorization accuracy from the previous block was displayed onscreen in the form of a percentage. Manual drift correction was performed between each block by the experimenter to ensure high quality tracking of the pupil. Pupil data for the four-second period starting from the onset of the stimulus for each trial were used for further processing (see Section 3.1.4 for preprocessing details).

Following the training task, participants completed a tone categorization generalization block and the OSPAN task as described in Section 2.1.3. Analysis of the behavioral data from the category learning task was identical to Experiment 1 (see Section 2.1.4. Behavioral data analysis), with the exception that silence trials were removed for behavioral analyses. Each participant’s OSPAN score was used as a measure of WM in the behavioral and pupillary analyses. Participants were also categorized into higher ($n = 14$) and lower ($n = 14$) WM groups based on a median split (median OSPAN score = 41) for use in the computational modeling.

Upon completion of the training task and the generalization block, participants filled out the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) questionnaire. Participants completed the NASA-TLX as a self-report measure to assess effort, frustration level, mental demand, physical demand, temporal demand, and overall performance on the speech category learning task (Table 1).

3.1.3 Decision bound computational modeling

While behavioral accuracies provide important insights, they provide no information about how participants learned the categories and what strategies they used. To understand the strategies participants used, we applied several classes of decision-bound computational...
Table 1

| Demographic                  | Higher Working Memory | Lower Working Memory | t    | p    |
|------------------------------|-----------------------|----------------------|------|------|
| n (Female)                   | 14 (8)                | 14 (13)              | –    | –    |
| Age (Mean, SD)               | 21.500 (3.937)        | 19.571 (1.089)       | 1.766| 0.098|
| OSPAN (Mean, SD)             | 52.071 (7.022)        | 27.286 (8.453)       | 8.440| <0.001***|
| Years Music Experience (Mean, SD) | 4.357 (4.733) | 3.929 (4.747) | 0.239 | 0.813 |
| NASA-TLX (Mean, SD)          | Effort                | 73.571 (11.507)      | 0.837| 0.412 |
| Frustration Level            | 58.571 (22.311)       | 47.857 (28.333)      | 1.112| 0.277 |
| Mental Demand                | 67.143 (24.629)       | 74.643 (12.929)      | –1.009| 0.325 |
| Overall                      | 46.786 (26.502)       | 47.143 (22.678)      | –0.038| 0.970 |
| Performance                  | 41.071 (21.766)       | 33.214 (23.420)      | 0.919| 0.366 |
| Physical Demand              | 30.000 (19.315)       | 41.071 (22.203)      | –1.408| 0.171 |

*** p < .001.

t models (Ashby, 1992; Ashby & Maddox, 1993; Chandrasekaran et al., 2016; Smydia et al., 2015). Decision bound models assume that participants divide the categories in the two-dimensional stimulus space (i.e., relative pitch and pitch change) with decision boundaries that can rely on rule-based or procedural-based learning processes.

Originally developed in the visual modality, decision bound models were recently extended to examine auditory category learning (Chandrasekaran et al., 2016; Goudbeek et al., 2008, 2009; Maddox et al., 2002, 2013, 2014; Maddox & Chandrasekaran, 2014; Roark & Holt, 2019b; Scharinger et al., 2013; Smydia et al., 2015). These models are fit individually for each participant and each block to mitigate challenges interpreting fits to aggregate data (Ashby et al., 1994; Estes, 1956; Estes & Maddox, 2005). Consistent with prior research, we specified three classes of models, with multiple instantiations possible within a class (Chandrasekaran et al., 2015, 2016; Maddox et al., 2014, 2016; Smydia et al., 2015; Yi et al., 2016). The classes included Rule-Based models, the Striatal Pattern Classifier model (SPC), and the Inconsistent/Random Responder models.

3.1.3.1. Rule-Based models. The Rule-Based class of models represents hypothesis-testing mechanisms that are dependent on WM resources and mediated by the reflective learning system. Rule-Based models assume that participants selectively attend to individual dimensions during learning and can be either unidimensional (attending to a single dimension) or multidimensional (attending to both dimensions).

Unidimensional models assume that participants draw boundaries between categories based on one of the dimensions (i.e., relative pitch or pitch change). Unidimensional models have four free parameters: three reflect the placement of the decision boundaries along the relevant dimension and one reflects perceptual and criterial noise. Separate unidimensional models were fit assuming participants made categorization decisions using only the relative pitch dimension or only the pitch change dimension.

Multidimensional models assume that the participant places two decision boundaries (one along each dimension) that are combined to determine category membership. When a participant uses a multidimensional strategy, they selectively attend to both dimensions. One subclass of multidimensional models has three free parameters: two reflect the placement of a single boundary along each dimension and one reflects perceptual and criterial noise. Models from this subclass are also referred to as ‘conjunctive’ models as the boundaries require the conjunctive combination of the boundaries along the two dimensions to separate all four categories and forms the shape of a plus sign (Fig. 3). A separate subclass of multidimensional models has four free parameters: two reflect the placement of two boundaries along one dimension, one reflects the placement of a single boundary along the other dimension, and one reflects perceptual and criterial noise. Models from this subclass are sometimes called ‘conjunctive-H’ models as the boundaries form the shape of an H (Fig. 3).

3.1.3.2. Striatal Pattern Classifier model. The implicit SPC model is a neurobiologically grounded model thought to represent procedural-based learning mechanisms (Ashby et al., 1998). The SPC model assumes that participants use feedback to learn stimulus–response associations instantiated within the striatum (Ashby & Waldron, 1999). The SPC model assumes that participants build category representations based on theoretical ‘striatal’ units (based on the neurobiology of medium-spiny neurons), where similar category responses to stimuli are clustered together in perceptual space, supported by the procedural-
3.1.3.3. Inconsistent/Random Responder model. The Inconsistent or Random Responder model assumes that participants respond to the stimuli with all responses being equally probable. Thus, this model captures participants who respond randomly to the stimuli and participants who may be changing strategies too often to be accounted for by the other models. We refer to these participants as Inconsistent strategy users.

3.1.3.4. Summary. The different classes of decision bound models make different assumptions about how participants divide the stimuli into categories based on the underlying dimensions that define the stimuli. As a demonstration of these models, Fig. 3 shows a hypothetical response pattern for each model class. Mathematical formulation of these models is outside of the scope of the current article and can be found elsewhere (Ashby, 1992; Ashby & Maddox, 1993).

As input, each model uses the dimensional coordinates (i.e., relative pitch and pitch change) of each stimulus and the participant’s actual response to that stimulus for a given block of each participant’s data. Importantly, these models are agnostic about the actual category identity of the stimuli and are based on the participant’s response. To visualize this, for each block of trials, we can plot all of the stimuli that the participant encountered in the two-dimensional space and divide that space using decision boundaries (Fig. 3) with each of the classes of models that have different assumptions with the goal of finding the best model to account for this participant’s response pattern in this block of data. The key differences between the models are in the assumptions they make about how participants use the stimulus dimensions to separate the categories. In some instances, they use only one dimension (unidimensional Rule-Based models). In other cases, they use both dimensions in a manner that is not orthogonal to the dimensions but rather consistent with selectively attending to that dimension (multidimensional Rule-Based models). In other cases, they use only one dimension (unidimensional Rule-Based models). In total, 13 unidimensional models, 11 multidimensional models, one SPC model, and one Random Responder model were fit to each block of each participant’s data. Based on the parameters of the SPC model, the model is very flexible about assignment of responses to regions of space. Similarly, the Random Responder model makes a basic assumption that subjects are randomly responding to the stimuli, and multiple versions with different response regions of space assumptions were not needed. Models were fit to the early (blocks 1–3) and late (blocks 4–6) blocks for each participant. As such, multiple models were fit across participants (26 models × 28 participants × 2 blocks = 1456 models fit).

3.1.3.6. Assessment of goodness of fit. To understand how well the best-fit model for each participant and block accounted for that participant’s actual response pattern, we computed the proportion of participant’s responses accurately predicted by the best-fit model (i.e., model accuracy) as a measure of goodness-of-fit. Once the best-fitting model was selected using BIC (comparing against all other models), we computed the predicted response based on the fitted parameters for the best-fit model. That is, as these models are categorical models, this is the category response (i.e., T1, T2, T3, or T4) that the best-fitting model predicts that the participant would have made if they applied this strategy consistently throughout that block. Next, we computed the predicted response of the best-fitting model to the observed response made by the participant for each stimulus presented in a given block. For example, for Stimulus-1, if the model predicted that the participant would have responded 1 and the participant actually responded 1, this would be coded as a ‘Correct’ prediction. However, if the model predicted 1 and the participant actually responded 2, this would be coded as an ‘Incorrect’ prediction. For each participant and each block, we computed the model accuracy by comparing the best-fit model’s predicted response and the participant’s actual (observed) response.

Using this goodness-of-fit measure of model accuracy, we observed that the best-fit models accurately captured participants’ response patterns with an average predicted vs. observed accuracy of 50% in the early block and 61% in the late block. With probability of success at 25% for a single trial and 120 trials, an accuracy of 32% has a 95% cumulative probability based on a binomial distribution, suggesting that accuracy higher than 32% can be considered ‘better than chance.’ The model accuracies of 50% and 61% in the early and late blocks both give > 99% cumulative probability. As such, our model fits can be considered better than chance, even when taking into account a confidence interval around 25% chance. Because the accuracy was better than chance, we know that the best-fit models provided an accurate account of participants’ response patterns. It was expected that the model accuracy would be less than 100% (i.e., not account for every single one of the participant’s responses in a given block of data). The reason model accuracy may be less than 100% is if a participant applied the strategy even somewhat inconsistently across all trials or they had a lapse in attention during even a single trial. Instead, the approach of these models was to capture the most likely strategy across a given block of trials and provide an estimate (with relatively high accuracy) of the strategy participants were using over the block of trials.

3.1.4. Pupilometry preprocessing

Consistent with prior research, pupilometry data were preprocessed to remove noise from the analysis (Winn et al., 2015; Zekveld & Kramer, 2014), using data from approximately 500 to 4000 ms time-locked to the onset of the stimulus presentation. Pupilometry data were down-sampled to 50 Hz (Wierda et al., 2012) and trials with more than 15% of samples detected as blinks were removed (Higher WM group: 1013 out of 4200 rejected; Lower WM group: 1031 out of 4200 rejected; McMahon et al., 2016). Missing samples due to blinks were linearly interpolated to 120 ms before and after the blink. Additional blinks were identified and linearly interpolated.
interpolated based on the first derivative of the blink threshold. Pupil responses were baseline normalized using the average pupil size in the 500 ms prior to the onset of the speech stimulus on a trial-by-trial basis (Peyshakhovich et al., 2015), wherein the outcome variable reported is the proportion change in pupil size relative to baseline.

### 3.1.5. Growth curve analysis

Pupil responses within the time window from 0 to 2500 ms time-locked to the speech stimulus onset were analyzed using growth curve analysis (GCA; Mirman, 2014). GCA is appropriate for modeling time-series data, such as pupillary responses, because it provides a statistical approach for modeling changes in the timing and shape of the pupillary response over time. GCA has an advantage over traditional approaches, like time-binned analysis of variance, for several reasons: 1) GCA does not require time-binned samples, which eliminates the trade-off between temporal resolution and statistical power, 2) there is no experimenter bias in selecting time windows in an arbitrary manner, and 3) the model can account for individual differences (Mirman, 2014).

The pupillary response unfolds over time and often does not follow a linear trajectory (Winn et al., 2015, 2016). Therefore, GCA uses orthogonal polynomial time terms to capture distinct functional forms of the pupillary response. A GCA was fit to model the interactions between first-, second-, third-, and fourth-order orthogonal polynomials and the independent variables of interest. The fourth-order polynomial was chosen to model the pupillary response because it was consistent with the trajectory of the data and provided a better fit than a model that included only the first-, second-, and third-order polynomials ($\chi^2(6) = 848.78, p < .001$; AIC$_{G1} = -51,460$; AIC$_{G2} = -52,297$). This fourth-order model uses four parameters to capture the complexity of the pupillary response. The intercept refers to the overall change in the pupillary response over the entire time window and can be interpreted as the average change in the pupillary response from start to finish (Mirman, 2014). The linear (ot1) term reflects the slope of the pupillary response and can be interpreted as the rate of dilution over time (Kuchinsky et al., 2014; Morett et al., 2020). The quadratic (ot2) term represents the curvature of the pupillary response. Given the trajectory of the pupillary response, the quadratic term should be negative. Here, a larger, negative quadratic term can be interpreted as a steeper curvature, while a quadratic term closer to zero indicates a more linear shape (Kuchinsky et al., 2014).

Lastly, the cubic (ot3) and quartic (ot4) terms represent the extent to which two or three inflection points occur in the pupillary response, respectively. GCA were conducted using the lme4 package (Bates et al., 2015) with log-likelihood maximization using the BOBYQA optimizer to promote convergence (Mirman, 2014), and p-values were estimated using the lmerTest package (Kuznetsova et al., 2017).

We first estimated a GCA to examine differences in pupillary responses between sound trials and silence trials. The pupillary data were averaged across sound and silence trials separately for each participant and entered into the model (i.e., two entries per participant). The optimal, final model included fixed effects of trial type (sound or silence; reference = silence) on all time terms with random slopes of subject on each time term and trial type. A model including a random effect for the interaction between trial type and subject on all time terms did not converge. The final random effect structure provided a better model fit than when including only random slopes of subject on each time term ($\chi^2(6) = 5,763.000, p < .001$; AIC$_{M1} = -34,269$; AIC$_{M2} = -40,020$).

We re-estimated a second GCA to examine pupillary changes as a function of WM on correct and incorrect trials between the early and late halves of training. Here, each participant’s pupillary responses were averaged across correct and incorrect trials separately for early (blocks 1–3) and late (blocks 4–6) training (i.e., four entries per subject) and entered into the model. This model included fixed effects of WM (i.e., each participant’s OSPAN score), trial accuracy (reference = incorrect trials), and training half (reference = early half) on all time terms, and the interaction of WM, accuracy, and training half on all time terms. This model also included random slopes of subject on each time term and a random slope of the interaction between subject, accuracy, and WM on the linear, quadratic, and cubic time terms. Models with random effect structures including interactions of subject, accuracy, WM, and training half on all time terms failed to converge. Our final model and random effect structure provided a better model fit than including only random slopes of subject on each time term ($\chi^2(6) = 6843.54, p < .001$; AIC$_{M1} = -53,860$; AIC$_{M2} = -54,532$).

### 3.2. Results

#### 3.2.1. Behavioral

Consistent with the findings from Experiment 1, participants with higher WM had significantly greater overall average accuracy ($p = 0.516, z = 2.587, p = 0.010$). However, we did not observe a significant interaction of trial and WM ($p = 0.186, z = 1.847, p = 0.065; \text{Fig. 4B}$). This suggests that while WM status had an effect on categorization accuracy, WM status did not have a significant effect on the rate of increase in accuracy across training. We observed similar patterns of accuracy in the generalization block. There was a significant and positive effect of WM ($p = 0.573, z = 2.230, p = 0.026$), such that participants with higher WM had greater accuracy across the final block of training and the generalization block. However, the interaction of block and WM was non-significant ($p = -0.093, z = -0.907, p = 0.365$), which demonstrates that the change in categorization accuracy from the final block of training to the generalization block did not differ based on WM status. Collectively, the results from the categorization task in Experiment 2 are similar to the findings from Experiment 1 and demonstrate that participants with higher WM scores have better Mandarin tone category learning accuracy during training than those with lower WM scores. Additionally, higher and lower WM participants did not significantly differ on any of the NASA-TLX subscales (see Table 1), which indicates that participants with higher and lower WM found the speech category learning task equally demanding on workload. Thus, the differences observed between higher and lower WM participants derive from specific aspects of learning.

Participants used different strategies during category learning. We examined the extent to which categorization accuracy differed in early and late training between Rule-Based strategy users (early: $n = 5$, $M = 0.475$, $SD = 0.232$; late: $n = 8$, $M = 0.734$, $SD = 0.174$), SPC strategy users (early: $n = 6$, $M = 0.594$, $SD = 0.174$; late: $n = 15$, $M = 0.695$, $SD = 0.180$), and Inconsistent strategy users (early: $n = 17$, $M = 0.382$, $SD = 0.121$; late: $n = 5$, $M = 0.383$, $SD = 0.093$), regardless of WM group membership. There was a significant effect of strategy type in early (F(2, 25) = 4.275, p = 0.025, $\eta^2 = 0.255$) and late training (F(2, 25) = 7.905, p = 0.002, $\eta^2 = 0.387$). Tukey-HSD post-hoc testing revealed that the mean categorization accuracy in early training of SPC strategy users was significantly greater than the categorization accuracy of Inconsistent strategy users (p = .021). However, categorization of Rule-Based strategy users in early training did not significantly differ from Inconsistent (p = .481) or SPC strategy users (p = .420). In late training, Inconsistent strategy users had significantly poorer accuracy than both Rule-Based (p = .003) and SPC (p = .004) strategy users, but accuracy did not differ between Rule-Based and SPC strategy users (p = .856).

When examining strategy usage between higher and lower WM participants, as determined by a median split of OSPAN scores, we observed no differences in early training (Fisher’s Exact Test, $p = 0.184$). However, in late training, strategy usage differed between higher and lower WM participants (p = 0.034; Fig. 4C). Specifically, participants with lower WM used a mix of learning strategies, while participants with higher WM were more likely to use optimal procedural-based (i.e., SPC) learning strategies. Although strategy usage differed between higher and lower WM participants in late training, there were no differences in categorization accuracy within each learning strategy. For those using SPC learning strategies in late training, there were no significant differences in categorization accuracy between participants with higher
Fig. 4. Experiment 2 behavioral results. A) Distribution of scores on the OSPAN task for all 28 participants. B) Categorization accuracy across training blocks and generalization block (abbreviated as “Gen”) for higher \((n = 14)\) and lower \((n = 14)\) working memory (WM) participants, as defined by a median split \((\text{Mdn} = 41)\) of OSPAN scores for visualization purposes. Categorization accuracy is defined as the proportion of correct trials within each block. Average accuracy for each WM group is denoted by the darker lines and points. The lighter lines and points denote each individual participant’s accuracies. The dashed line reflects chance-level performance \((0.25)\). C) Distribution of strategy usage among higher and lower WM participants in early \((blocks 1–3)\) and late \((blocks 4–6)\) training halves. There were no differences in strategy usage in early training between WM groups, but strategy usage significantly differed in late training between higher and lower WM participants \((p = .034)\).

Fig. 5. Pupillary responses (baseline normalized) to stimulus onset. A) Proportion change in pupil size for silence and sound trials. Shaded regions represent the standard error of the mean. B) Proportion change in pupil size as a function of working memory (WM) group for correct and incorrect trials in early \((blocks 1–3)\) and late \((blocks 4–6)\) training halves. Shaded regions represent the standard error of the mean. Participants were split into higher \((n = 14)\) and lower \((n = 14)\) WM groups based on a median split \((\text{Mdn} = 41)\) of OSPAN scores for visualization purposes only. C) Decomposition by growth curve time terms of the proportion change in pupil size as a function of WM group and training half for participants with incorrect and correct trials. The combination growth curve (abbreviated as “Combo”) represents the sum of each individual time term component to model the pupillary response.
WM (Mdn = 0.756) and participants with lower WM (Mdn = 0.554; W = 14,000, p = 0.327, r = 0.270). Additionally, there were no significant differences in late training for higher and lower WM participants using Rule-Based learning strategies (higher: Mdn = 0.825; lower: Mdn = 0.633; W = 2,500, p = 0.314, r = 0.415) nor Inconsistent learning strategies (higher: Mdn = 0.250; lower: Mdn = 0.308; W = 3,500, p = 0.468, r = 0.487).

3.2.2. Pupillometry

We first compared pupillary responses between sound trials and silence trials to confirm that pupillary responses on sound trials differed from silence trials as a proof of concept (Fig. 5A). Pupillary responses on sound trials were significantly different from silence trials on the intercept (β = 0.085, t = 14.511, p < .001), linear (β = 0.371, t = 102.825, p < 0.001), quadratic (β = −0.309, t = −85.513, p < 0.001), cubic (β = −0.038, t = −10.617, p < 0.001), and quartic (β = 0.093, t = 25.869, p < 0.001) terms. These results indicate that pupillary changes on sound trials were larger, increased more gradually over time, were stronger in curvature, and had significantly different secondary and tertiary inflection points. Silence trials were not included in the subsequent GCA model. Full model details are provided in Table 2.

3.2.2.1. Effects of WM and category learning accuracy in early training.

Next, we examined pupillary responses as a function of WM for correct and incorrect trials in early training (see Table 3; Fig. 5B). For correct trials in early training, we observed no significant main effects on any of the time terms, indicating that correct trials were similar in size, rate of dilation, curvature, and secondary and tertiary inflection points to incorrect trials in early training (Fig. 5C). Additionally, we did not observe any significant main effects of WM on incorrect trials, indicating that the pupillary response on incorrect trials in early training was similar in size, slope, curvature, and secondary and tertiary inflection points regardless of WM status. Finally, we did not observe any significant two-way interactions between accuracy and WM. Taken together, these findings suggest that pupillary responses in early training do not differ based on trial accuracy or WM status.

3.2.2.2. Effects of WM between training halves and category learning accuracy.

Next, we examined the extent to which pupillary responses during accurate categorization were modulated across training as a function of WM scores (see Table 3; Fig. 5B). We observed significant main effects of training half on the intercept, linear, quadratic, cubic, and quartic terms. These significant results indicate that pupillary responses on incorrect trials during late training were overall smaller in size, increased at a slower rate over time, were shallower in curvature, and had less prominent secondary and tertiary inflection points in comparison to incorrect trials in the early training half (Fig. 5C).

Additionally, working memory significantly interacted with training half on the intercept, linear, and quadratic terms, suggesting that the change in the size, rate of dilation, and the curvature of the pupillary response on incorrect trials from early to late training differed as a function of increasing WM scores. These results indicate that individuals with higher WM scores had a smaller decrease in the overall size of the pupillary response from early to late training compared to those with lower WM scores. Additionally, the rate of dilation increased from early to late training as a function of increasing WM scores, while the curvature of the pupillary response was shallower for higher WM scores.

We also observed significant interactions between accuracy and
training half on the intercept, linear, and quadratic terms. Simple effect analyses revealed that on correct trials in late training, the pupillary response was significantly smaller ($t = 16.470, p < .001$), increased at a slower rate over time ($t = 3.333, p = .005$), and had less prominent secondary ($t = -5.846, p < .001$) and tertiary ($t = 2.990, p = .015$) inflection points relative to correct trials in early training. This same analysis also revealed the pupillary response for correct trials in late training increased at a slower rate over time ($t = 2.740, p = .031$) than incorrect trials in late training. However, there were no significant differences in the overall size, curvature, nor secondary or tertiary inflection points ($p > 0.05$) between correct and incorrect trials in late training. Collectively, these results indicate that the overall size, rate of dilation, and curvature of the pupillary response reduced from early to late training to a greater extent for correct trials, while correct and incorrect trials primarily differed in the rate of dilation in late training.

Finally, we observed significant negative three-way interactions between accuracy and training half as a function of WM scores on the intercept, linear, and quadratic terms. Simple effect analyses revealed that on correct trials in late training, the overall pupillary response was significantly smaller ($t = 16.465, p < .001$) and had a less prominent secondary inflection point ($t = -5.846, p < .001$) for participants with lower WM scores, but participants with higher WM scores had a significantly slower rate of dilation over time ($t = 3.334, p = .041$) in the late training half compared to their correct trials in the early training half. The curvature and secondary inflection point of the pupil response did not significantly differ from early to late training ($p > 0.05$). However, the significant three-way interaction on the quadratic term suggests that the decrease in the curvature of the pupillary response from early to late training on correct trials was significantly larger as a function of increasing WM scores. This same analysis also revealed that pupillary responses on correct trials in late training increased at a significantly slower rate over time ($t = 4.374, p < .001$) as a function of increasing WM scores, compared to incorrect trials in late training. However, the overall size, curvature, and secondary and tertiary inflection points did not differ between correct and incorrect trials in late training as a function of increasing WM scores ($p > 0.05$). Taken together, these significant three-way interactions indicate that pupillary responses on correct trials significantly decreased in the overall size, rate of dilation, and curvature of the pupillary response to a greater extent for participants with higher WM compared to those with lower WM. Further, these significant three-way interactions demonstrate that in late training, participants with higher WM had a larger difference in the rate of dilation between correct and incorrect trials than participants with lower WM scores.

Collectively, these results provide evidence that the effects of trial accuracy and WM on pupillary responses during speech category learning were modulated by training half. Specifically, pupillary response on trials that were categorized correctly reduced to a greater extent from early to late training in participants with higher WM compared to participants with lower WM. Participants with higher WM also had a larger difference in the pupillary response between trials categorized correctly and those categorized incorrectly in late training than participants with lower WM. These results demonstrate that pupillary responses may reflect greater use of optimal, procedural-based learning strategies for categorization by late training during speech category learning.

4. General discussion

We examined the extent to which individual differences in WM influence non-native speech category learning. In Experiment 1, we showed that adults with higher WM learned Mandarin tone categories better and faster than individuals with lower WM. In Experiment 2, we also observed a learning enhancement for individuals with higher WM in a smaller sample of participants who learned to categorize Mandarin tones while pupillometry was simultaneously recorded. Our neurobiologically-inspired computational models revealed that individuals with higher WM used more optimal, procedural-based strategies in late training, while those with lower WM perseverated with rule-based learning strategies. Pupillary responses revealed that participants with higher WM scores experienced a greater reduction in stimulus-evoked pupil responses for correct trials by late training and also showed larger differences between correct and incorrect trials in late training compared to participants with lower WM scores. Taken together, our findings suggest that higher WM promotes earlier use of optimal procedural-based learning strategies during non-native speech category learning. A substantial reduction in stimulus-evoked pupillary responses on correct trials during late training may further reflect a reduction in cognitive effort induced by the WM-independent, procedural-based strategy use.

Mandarin tone categories are optimally learned with procedural-based strategies, as the multiple acoustic dimensions that differentiate these categories are difficult to verbalize (Chandrasekaran, Koslov, et al., 2014; Maddox & Chandrasekaran, 2014; Yi et al., 2016). Our results demonstrate that strategy use was similar between higher and lower WM individuals during early training, where a single strategy did not dominate learning. However, individuals with higher WM used more procedural-based learning strategies in late training compared to individuals with lower WM. Prior work has shown a bias towards using rule-based learning strategies in early training, wherein successful learners switch to procedural-based learning strategies by late training, while less successful learners continue using suboptimal strategies (Maddox & Chandrasekaran, 2014; Seger, 2008; Seger & Miller, 2010).

The established involvement of WM during rule-based learning (Decaro et al., 2008; Filoteo et al., 2010; Miles et al., 2014; Reetzke et al., 2016; Zeithamova & Maddox, 2006) may enable individuals with higher WM to better maintain speech stimulus-related information, which may, in turn, allow learners to extract and integrate key perceptual dimensions for accurate categorization. While this would be important in early training, it could be less critical in later training once initial category templates have been established. Our finding that individuals with lower WM were more likely to use non-procedural strategies in late training replicates prior findings that individuals with lower WM perseverate with suboptimal strategies in categorization (Decaro et al., 2009; Tharp & Pickering, 2009).

Our results are interesting to interpret in the context of Craig and Lewandowsky (2012), who examined the relationship between WM and visual category learning. In their study, WM predicted the speed of category learning. Strategy choice was partially predicted by category learning performance, and participants with higher training accuracy showed fewer strategy shifts. However, WM did not predict which strategies participants used; rather, WM predicted how well participants used their selected strategy. The authors argued that WM modulates associative learning (i.e., stimulus-to-category mapping), regardless of strategy usage. In contrast, our study supports the perspective that WM may influence the selection of optimal strategies for learning.

Importantly, strategy usage was indexed by pupillary responses. Reductions in the stimulus-evoked pupillary responses across training reflected a reduction in cognitive effort, which corresponded with the shift to procedural-based learning strategies in individuals with higher WM. During early training, there were few stimulus-evoked pupillary differences between higher and lower WM individuals. By late training, differences were robust, wherein individuals with higher WM scores experienced a greater reduction in the overall size, rate of dilation, and secondary inflection point on correct trials and had a greater difference between correct and incorrect trials in late training than individuals with lower WM. The effect of WM on the reduction in the stimulus-evoked pupillary response across training on trials that were categorized correctly likely reflects reduced cognitive effort (Kuchinsky et al., 2014; Morett et al., 2020; Winn et al., 2015). Moreover, less cognitive effort is required when categories are optimally learned using procedural-based learning strategies (Maddox & Chandrasekaran,
2014). It is also possible that the rate of dilation over time (i.e., linear term) on correct trials may reflect the implementation of procedural-based learning strategies. An alternative interpretation of these findings could be that individuals with higher WM exerted more effort early on or that they experienced greater salience for the stimuli relative to individuals with lower WM (Liao et al., 2016). Taken together, greater changes in the pupillary response on correct trials were observed across training as a function of WM, which aligned with the changes in learning strategies across training in individuals with higher WM.

A recent study by Lewis and Bidelman (2020) utilized pupillometry to examine speech categorization in the presence of noise. They found that pupillary response latencies were strongly related to the speed of listener decisions, where slower reaction times were associated with slower pupillary responses to ambiguous speech tokens. The authors posited that delays in pupillary responses during speech sound categorization may reflect ambiguity resolution. In contrast, our findings showed individuals with higher WM had a significantly slower rate of dilation on correct trials relative to incorrect trials in late training. It is possible that task differences may explain the discrepancies between our results and the findings from Lewis and Bidelman (2020). For instance, their experiment was a categorical perception task with two category representations comprised of single vowel sounds masked in noise, which requires acoustical resolution for categorization, as opposed to our category learning experiments of four speech categories comprised of syllables with superimposed pitch contrasts in quiet. In non-native speech category learning experiments, as in the current study, distinct category representations are emerging, and fine-grained acoustic disambiguation may not be necessary, particularly in late training. Further, Kong and Edwards (2016) also examined individual differences in categorical speech perception. They found that listeners with a more gradient response to syllable speech stimuli, rather than a categorical response, had greater flexibility in cue utilization. While their experimental tasks differed from those in the current study, their findings have interesting implications for more complex non-native speech category learning. Unlike our findings on the relationship between WM and Mandarin tone categorization, Kong and Edwards (2016) did not observe a relationship between executive functioning and categorical perception performance. Their study used measures of inhibition and task shifting and did not specifically examine a relationship between WM and categorization. Thus, it is unclear how WM may influence categorical perception. Future studies investigating the role of WM and other cognitive processes on non-native speech category learning may consider including categorical perception tasks to tease apart these discrepancies.

Similar to prior neuroimaging findings that observed greater functional coupling between the left superior temporal gyrus and the striatum on incorrect responses during late training (Feng, Yi, et al., 2018), we observed greater pupillary differences between correct and incorrect trials as a function of WM. Our findings demonstrate that incorrect trials had a faster rate of dilation than correct trials in late training for individuals with higher WM scores. Interestingly, the increased coupling between the left superior temporal gyrus and the striatum on incorrect responses during late training in Feng et al. (2018) was observed during corrective feedback presentation. In the current study, we observed pupillary differences between correct and incorrect trials as early as the stimulus encoding period. While this suggests that individuals with higher WM may have focused greater attention on less established stimuli in late training, it also suggests that individuals with higher WM were better able to attend to the relevant dimensions of these less established stimuli to fine-tune their category representations in late training. These findings demonstrate that a combination of corrective feedback and attention to relevant dimensions during stimulus encoding, modulated by individual differences in WM, drive speech category learning success.

Individual differences between individuals with higher and lower WM is purported to be caused by a deficit in LC-NE system functioning in individuals with lower WM (Unsworth & Robson, 2017a). One interpretation of these results is that a disruption in LC-NE functioning leads to attentional fluctuations in individuals with lower WM, such that individuals with higher WM were better able to attend to the speech stimuli and category-relevant dimensions. It could be that individuals with higher WM have either 1) more efficient release of norepinephrine by the LC-NE system or 2) greater activation of norepinephrine receptors in the prefrontal cortex compared to those with lower WM. Intermediate levels of norepinephrine (Arnsten, 1999; Berridge & Spencer, 2016) and greater norepinephrine receptor activation in the prefrontal cortex (Arnsten, 1999; Spencer & Berridge, 2019) have been shown to influence attention and WM processes. Thus, enhanced attention in individuals with higher WM related to LC-NE system activity may facilitate the formation of novel category representations and optimal strategy usage to a greater degree than in individuals with lower WM. Collectively, the differences observed in speech category learning performance, learning strategies, and pupillary responses between higher and lower WM individuals in the current set of experiments suggest that the LC-NE system may be an important modulator of individual differences in WM during non-native speech category learning success.

4.1. Limitations and directions for future study

The two experiments presented in this paper only included younger adult participants who were relatively within the normal ranges of WM. Further research should investigate non-native speech category learning across the lifespan (i.e., in young children and older adult participants) and in populations with clinical deficits in WM. Future research should also consider employing attention tasks to better parse the relationship between attentional fluctuations and WM during category learning. Prior work on speech category learning has shown that changes in selective attention, motivational factors, and performance pressure enhance learning (Chandrasekaran et al., 2016; Lim & Holt, 2011; Maddox et al., 2016). Pupillometry research paired with manipulation of these factors could shed more light on their contribution to speech category learning.

Overall learning performance varied between experiments. The large online sample in Experiment 1 obtained approximately 50% accuracy by the end of training, while the smaller lab-based sample in Experiment 2 reached closer to 60% accuracy by the end of training. While these differences between experiments could be due to the integrity of online data collection, the accuracy in Experiment 1 is significantly better than chance (25%) and aligns with prior lab-based studies with approximately 200 participants (Chandrasekaran et al., 2016). Learning effects observed in Experiment 1 are similar to other large sample sized experiments investigating Mandarin tone category learning in native English speakers, which lends to the validity of the data collected in an online setting. Regarding the transfer of knowledge to speech stimuli from novel talkers, WM did not modulate one’s ability to generalize to untrained talkers in either experiment immediately after training. While unexpected, this finding is in line with work from Ingvaison et al. (2017) who also found that WM was not associated with generalization performance in older adults learning Mandarin tone pseudowords. Instead, they found learning and generalization performance related to declarative memory capacity and that older adults employed declarative-supported rule-based strategies.

Lastly, we observed a significant interaction of trial and WM on learning accuracy in Experiment 1, but this interaction was not significant in Experiment 2. This interaction provides information on the trial-by-trial increase in accuracy as a function of WM status, or in other words, the rate of learning across the training task. The lack of a significant interaction in Experiment 2 could be due to the nature of the pupillometry task. Trial events were extended to allow time for the slowing reacting pupil to respond, but this extra timing may have introduced an additional factor that affected the rate of learning that was not present in Experiment 1. Therefore, while it is evident from the main effects of WM
status on accuracy in both experiments that higher WM had a positive effect on overall learning accuracy, it remains unclear whether WM status enhances the rate of learning non-native speech categories.

4.2. Conclusion

In conclusion, we demonstrate that individual differences in WM facilitate non-native speech category learning in adults. Individuals with higher WM acquired non-native speech categories with greater accuracies, relative to those with lower WM. Computational modeling and pupillometry revealed greater usage of procedural-based learning strategies and distinct stimulus-evoked pupillary signatures in individuals with higher WM as a function of training. We posit that individuals with lower WM may be more prone to attentional fluctuations, which may result in poorer encoding of key stimulus dimensions and suboptimal use of corrective feedback for error monitoring. The findings from the current set of experiments encourage further investigation of the role of individual differences in WM and the LC-NE system on non-native speech category learning success.

CRediT authorship contribution statement

Jacie R. McHaney: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Project administration, Writing – original draft, Writing – review & editing.

Rachel Tessmer: Conceptualization, Data curation, Formal analysis, Investigation, Software, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Casey L. Roark: Formal analysis, Methodology, Writing – review & editing.

Bharath Chandrasekaran: Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Dr. Zilong Xie for his contribution to experimental design and early versions of preprocessing pupillometry data. This work used the Extreme Science and Engineering Discovery Environment (XSEDE, Towns et al., 2014), which is supported by NSF (ACI-1548562). Specifically, it used the Bridges system (Nystrom et al., 2015), which is supported by NSF (ACI-1445606), at the Pittsburgh Supercomputing Center.

Funding

This research was supported by the Defense Advanced Research Projects Agency as part of the Targeted Neuroplasticity Program (contract number: N66001-17-2-4008), the National Institutes of Health grant (R01DC015504) to B. Chandrasekaran, and the National Institutes of Health grant (T32-DC011499) to K. Kandler and B. Yates (trainee: J. R. McHaney).

References

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking. Journal of Vision, 14(4), 1-20. https://doi.org/10.1167/14.4.1doi.

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

Arntzen, A. F. T. (1999). Through the looking glass: Differential noradrenergic modulation of prefrontal cortical function. Neural Plasticity, 7(1–2), 133-146.

Ashby, F. G. (1992). Multidimensional models of categorization. In Multidimensional models of perception and cognition (pp. 449–483). Lawrence Erlbaum Associates Inc.

Ashby, F. G., Alfonso-Reese, L. A., Turk, E., & Waldron, E. M. (1996). A neuropsychological theory of multiple systems in category learning. Psychological Review, 103(3), 442-481.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. Trends in Cognitive Sciences, 5(5), 204-210. https://doi.org/10.1016/S1364-6613(00)00162-4.

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. (1995). Relations between prototype, exemplar, and decision bound models of categorization. Journal of Mathematical Psychology, 37(3), 372-400.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. Annals of the New York Academy of Sciences, 1224(1), 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. Psychological Science, 5(2), 144-151.

Ashby, F. G., & Valentim, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In Handbook of Categorization in Cognitive Science (pp. 157–188). Elsevier. https://doi.org/10.1016/b978-0-08-101107-2.00007-5.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. Psychonomic Bulletin & Review, 6(3), 363–378.

Ashby, J. R., & Maddox, W. T. (2020). On the role of attention in perceptual learning. Journal of Vision, 20. https://doi.org/10.1167/20.4.1.doi.

Bradlow, A. R. (2008). Training non-native language sound patterns Lessons from training Japanese adults on the English /r/-/l/ contrast.

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014a). Toward a dual-learning systems model of speech category learning.

Chandrasekaran, B., Yi, H. G., Smayda, K. E., & Maddox, W. T. (2016). Effect of explicit corrective feedback for error monitoring. The findings from the cur- rent set of experiments encourage further investigation of the role of individual differences in WM and the LC-NE system on non-native speech category learning success.

CRediT authorship contribution statement

Jacie R. McHaney: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Project administration, Writing – original draft, Writing – review & editing.

Rachel Tessmer: Conceptualization, Data curation, Formal analysis, Investigation, Software, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Casey L. Roark: Formal analysis, Methodology, Writing – review & editing.

Bharath Chandrasekaran: Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Dr. Zilong Xie for his contribution to experimental design and early versions of preprocessing pupillometry data. This work used the Extreme Science and Engineering Discovery Environment (XSEDE, Towns et al., 2014), which is supported by NSF (ACI-1548562). Specifically, it used the Bridges system (Nystrom et al., 2015), which is supported by NSF (ACI-1445606), at the Pittsburgh Supercomputing Center.

Funding

This research was supported by the Defense Advanced Research Projects Agency as part of the Targeted Neuroplasticity Program (contract number: N66001-17-2-4008), the National Institutes of Health grant (R01DC015504) to B. Chandrasekaran, and the National Institutes of Health grant (T32-DC011499) to K. Kandler and B. Yates (trainee: J. R. McHaney).

References

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking. Journal of Vision, 14(4), 1-20. https://doi.org/10.1167/14.4.1doi.

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

Ashby, J. R., McHaney, et al. Brain and Language 222 (2021) 105010

Dunlavy, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. The Quarterly Journal of Experimental Psychology, 60(1), 1227–1245. https://doi.org/10.1080/17470210600926075.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. Psychology of Learning and Motivation, 44, 145–199.

Estes, W. (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53(2), 134–146.

Estes, W., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. Psychonomic Bulletin & Review, 12(3), 403–408. https://doi.org/10.3758/BF03197894.

Feng, G., Gan, Z., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2018). Task-General and Acoustic-Speech Representation in the Human Brain. Cerebral Cortex, 28, 3241–3254. https://doi.org/10.1093/cercor/bhx195.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2018). The role of the human auditory corticostriatal network in speech learning. Cerebral Cortex, 29(10), 4077–4089. https://doi.org/10.1093/cercor/bhy294.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. Psychological Science, 21(3), 415–423. https://doi.org/10.1177/0956797610362646.

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In English, 46(11), 1177–1227. https://doi.org/10.1177/0013915795461112.3.

Estes, W. (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53(2), 134–146.

Estes, W., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. Psychonomic Bulletin & Review, 12(3), 403–408. https://doi.org/10.3758/BF03197894.

Feng, G., Gan, Z., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2018). Task-General and Acoustic-Speech Representation in the Human Brain. Cerebral Cortex, 28, 3241–3254. https://doi.org/10.1093/cercor/bhx195.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2018). The role of the human auditory corticostriatal network in speech learning. Cerebral Cortex, 29(10), 4077–4089. https://doi.org/10.1093/cercor/bhy294.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. Psychological Science, 21(3), 415–423. https://doi.org/10.1177/0956797610362646.

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In English, 46(11), 1177–1227. https://doi.org/10.1177/0013915795461112.3.

Estes, W. (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53(2), 134–146.

Estes, W., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. Psychonomic Bulletin & Review, 12(3), 403–408. https://doi.org/10.3758/BF03197894.

Feng, G., Gan, Z., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2018). Task-General and Acoustic-Speech Representation in the Human Brain. Cerebral Cortex, 28, 3241–3254. https://doi.org/10.1093/cercor/bhx195.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2018). The role of the human auditory corticostriatal network in speech learning. Cerebral Cortex, 29(10), 4077–4089. https://doi.org/10.1093/cercor/bhy294.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. Psychological Science, 21(3), 415–423. https://doi.org/10.1177/0956797610362646.

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In English, 46(11), 1177–1227. https://doi.org/10.1177/0013915795461112.3.

Estes, W. (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53(2), 134–146.

Estes, W., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. Psychonomic Bulletin & Review, 12(3), 403–408. https://doi.org/10.3758/BF03197894.

Feng, G., Gan, Z., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2018). Task-General and Acoustic-Speech Representation in the Human Brain. Cerebral Cortex, 28, 3241–3254. https://doi.org/10.1093/cercor/bhx195.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2018). The role of the human auditory corticostriatal network in speech learning. Cerebral Cortex, 29(10), 4077–4089. https://doi.org/10.1093/cercor/bhy294.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. Psychological Science, 21(3), 415–423. https://doi.org/10.1177/0956797610362646.
