

Individual differences in working memory impact the trajectory of non-native speech category learning

Casey L. Roark^{*1,2}, Giorgio Paulon³, Giovanni Rebaudo³, Jacie R. McHaney¹, Abhra Sarkar³, & Bharath Chandrasekaran^{*1,2}

¹University of Pittsburgh, Communication Science & Disorders

²Center for the Neural Basis of Cognition

³University of Texas at Austin, Statistics and Data Sciences

* Corresponding authors

Email: casey.l.roark@gmail.com; b.chandra@pitt.edu

Abstract

What is the role of working memory over the course of non-native speech category learning?

Prior work is mixed with respect to the role of working memory in speech learning and has focused on how this important cognitive skill might influence learning assessed at a single timepoint. Here, we substantially extend this prior work by examining the role of working memory on speech learning performance over time (i.e., over several months) and leverage a multifaceted approach that provides key insights into how working memory influences learning accuracy, maintenance of knowledge over time, generalization ability, and decision processes.

We found that the role of working memory in non-native speech learning depends on the timepoint of learning and whether individuals learned the categories at all. Among learners, across all stages of learning, working memory was associated with higher accuracy as well as faster and slightly more cautious decision making. Further, while learners and non-learners did not have substantially different working memory performance, learners had faster evidence accumulation and more cautious decision thresholds throughout all sessions. Working memory may enhance learning by facilitating rapid category acquisition in initial stages and enabling faster and slightly more careful decision-making strategies that may reduce the effort needed to learn. These results have important implications for developing interventions to improve learning in naturalistic language contexts.

Keywords: category learning; working memory; speech perception; language learning; drift diffusion models

Categorization involves mapping variable inputs to discrete labels and is an important process that supports complex cognitive processes, such as object recognition [1] and speech perception [2]. Humans can learn novel categories throughout the lifespan across different perceptual modalities. However, there are also large individual differences in the underlying learning processes and outcomes. As such, there is a need to better understand what contributes to successful or less successful learning. In this study, we systematically examine the contributions of an ability that has been linked to category learning in prior work – working memory capacity.

Working memory (WM) reflects the resources available for the temporary storage and manipulation of information relevant for a given task [3,4]. Category learning involves many processes that are dependent on WM. Learners need to attend to task-relevant features and ignore task-irrelevant features, maintain features of a stimulus in mind as relevant or irrelevant for a decision, hold hypotheses in mind about stimulus-category-response mapping, compare representations of the stimulus to previous stimuli or rules, and incorporate feedback to update existing category representations and hypotheses about category identity. The ability to learn categories across sensory modalities has generally been found to be positively associated with WM [5–10]. WM is thought to support faster initial category learning [5] by allowing learners to hold multiple hypotheses about category identity at mind and test these hypotheses and specifically to rapidly and efficiently find a useful hypothesis [11].

Importantly, prior studies have primarily focused on the role of WM in initial learning, and, as a result, it is unclear how WM may play a role in maintenance of performance or learning patterns over time. In the earliest stages of learning, learners must be highly flexible with their behavior and search a large pool of potential hypotheses about category identity. As performance

improves and becomes more stable over time, WM processes may be less relevant because learners may be making small refinements to existing rules rather than keeping many competing hypotheses in mind. As a result, it is necessary to examine learning beyond very initial learning especially for categories that are difficult or challenging to acquire within a single session.

In the current study, we examine a specific case of category learning that is an important skill in second language acquisition – non-native speech category learning. The ability to learn a new language has been positively associated with individual abilities like WM capacity [12–15]. Assessed in a single session, the ability to learn sounds of a non-native language in adulthood has been positively linked to WM capacity [9,10]. However, other studies examining learning across longer training periods (e.g., multiple sessions across many days) have found that WM ability does not predict the ability to learn non-native speech categories [16,17]. The role of WM across the trajectory of non-native speech category learning is not yet clear. It is possible that WM supports initial, but not later speech learning.

In the current study, we train participants on non-native Mandarin tone categories. In Mandarin, distinct pitch patterns are lexically contrastive – the same syllable produced with four different pitch patterns (e.g., high-flat, low-rising, low-dipping, and high-falling) has four unique meanings. Learning to distinguish sounds based on these pitch patterns can be difficult for non-native listeners and there are large individual differences in learning [18–23].

For both speech and artificial perceptual categories, training beyond one session can be very successful, leading to significant learning and retention over time. In studies not focused on WM, participants learn through extensive training over several weeks [23–28] and then sometimes are brought back for a test of retention months later (e.g., [26] – 3 months; [23] – 8

91 weeks). Neural representations of categories rapidly emerge within a single session of initial
92 learning [29,30], but continue developing over time with more experience [23].

93 The role of WM beyond initial category acquisition is not well understood. Whereas
94 initial learning involves testing a large range of possible hypotheses about stimulus-response
95 mapping and using feedback to update these hypotheses, learning beyond the novice stage
96 involves refining existing hypotheses, learning about idiosyncratic stimuli, and continuing to
97 develop and refine representations. Additionally, after a delay in experience, learners must
98 reactivate existing representations and hypotheses to continue refining their category knowledge.
99 It is possible that these processes rely less on WM than initial testing among multiple hypotheses
100 as is necessary during initial learning. In the current study, we examine the role of WM in both
101 an initial learning session and learning sessions after one and three months from the initial
102 session.

103 Our approach involves inviting participants who previously completed a single session of
104 training on Mandarin tone categories [10] back for additional training sessions. McHaney et al.
105 [10] demonstrated that WM abilities were related to success in initial non-native speech category
106 learning across two experiments – one behavioral (revisited here) and one with pupillometry.
107 Specifically, individuals with higher WM capacity were better at learning, better at finding task-
108 appropriate strategies, and had pupil responses that reflected better stimulus-related attention.
109 Based on this, McHaney et al. [10] concluded that WM may support learning by enhancing
110 attention to task-relevant information. Critically, because this prior study tested only a single
111 session of learning, it is possible this conclusion may only apply to initial learning. In the current
112 study, we invite participants from Experiment 1 of McHaney et al. [10] back for two additional
113 sessions – one session one month after their initial training and another session two months after

the second session. We follow up with the same sample from McHaney et al. [10] to understand how individual differences in WM relate to individual differences in learning beyond initial acquisition.

An important aspect of understanding individual differences in learning is the acknowledgement that many individuals perform at chance levels even after extensive training. We see two main possibilities that could explain this pattern – (1) these participants are actively engaged and trying but persistently fail to learn and/or (2) these participants are actively disengaged and are not trying to learn, so they fail to learn. Disentangling these two possibilities is challenging. Prior work on category learning takes one of two approaches regarding participants performing at chance levels. Some studies remove these participants entirely, typically by removing participants who perform at or below chance levels by the end of learning [31–34]. Other studies retain these participants in the sample as it is impossible to know if their performance reflects a true inability to learn or whether they are disengaged [35–38]. The lack of consistency in these approaches across studies makes it difficult to understand this poor performing subset of the population. In the current study, we take a hybrid version of these approaches to better understand the underlying challenges facing less successful performers. We examine both the entire set of participants and participants who perform at above-chance levels (i.e., learners vs. non-learners who do not perform at above-chance levels). By examining the patterns while considering if participants eventually learned or not, we can better understand behaviors and abilities that lead to success.

We employ a multifaceted approach to understand what WM does or does not do for initial and later learning of speech categories. Specifically, we assess if WM is related to (1) performance in initial and later learning sessions, (2) maintenance of category knowledge over

137 time, (3) generalization of category knowledge to different talkers, (4) rate of evidence
 138 accumulation and (5) response caution during decision making (Table 1).

139

140 **Table 1. Hypothesized role of working memory across measures.**

	Hypothesized role of working memory	Relevant measure(s) in current study
Initial learning	Hold multiple hypotheses in mind, better and faster learning	Accuracy in session 1
Later learning	Enhanced attention and motivation	Accuracy in sessions 2 and 3
Maintenance of category knowledge	Quickly reactivate and flexibly use existing representations	Accuracy in first block of sessions 2 and 3 compared to final block of sessions 1 and 2
Generalization	Flexibly apply rules to new contexts	Accuracy in generalization test with different talkers and no feedback
Evidence accumulation	More efficient processing, mobilization of attentional resources	Evidence accumulation (drift) rate parameter from drift diffusion modeling
Response caution	More cautious, gather more information and test against multiple hypotheses before a decision is made	Decision threshold (boundary) parameter from drift diffusion modeling

141

142 **Initial and later learning**

143 Based on prior work, we expect that higher WM will be beneficial to initial acquisition
 144 (i.e., session 1) of non-native speech categories [10]. This prediction stems from prior work that
 145 has demonstrated that higher WM is associated with faster and better initial artificial category
 146 learning [5–7,9,39–41]. We also expect to observe this pattern given that the first session of
 147 training was published in McHaney et al. [10] where among all 195 participants, WM was

positively related to learning. A subset of these participants (107/195) returned for the current study.

Building on the prior study, we will probe the extent to which WM is related to performance in subsequent learning sessions. It is possible that WM provides benefits only in initial learning by quickly allowing learners to test many different hypotheses and find the ones that maximize their performance (e.g., [11]) and that WM is unrelated to learning beyond this novice stage. This prediction would be consistent with the observation that WM is not related to speech category learning when assessed after eight days of training [16,17]. However, it is also possible that WM provides benefits to learning beyond initial acquisition, allowing for enhanced further refinement of category representations.

Maintenance of category knowledge

By probing performance after one and two months of no additional exposure or training, we will examine the maintenance of performance over time. One possibility is that higher WM may allow learners to quickly reactivate and flexibly use their category representations developed in prior session(s). However, it is possible that maintenance of performance over time may be independent of WM and could reflect long-term memory abilities instead.

Generalization

The ability to accurately identify novel category exemplars is a hallmark of categorization. We will assess generalization in each session by presenting learners with novel stimuli spoken by novel talkers that they do not encounter during training and never receive feedback about the correct category. To successfully generalize to these novel talkers, they will need to apply their existing knowledge flexibly to the new context. It is possible that generalization relies on WM, as the ability to flexibly apply rules (e.g., cognitive flexibility) is

correlated with WM capacity [42,43] and generalization to novel contexts is related to individual differences in WM capacity [44,45].

Decision processes during learning

Using a drift diffusion modeling (DDM) approach [46,47], we will examine whether different components of the decision process (e.g., rate of evidence accumulation and response caution) are related to WM. DDMs are popular tools to understand decision making processes from accuracy and response time measures [48–52]. DDMs assume that during decision making, sensory evidence for multiple decision alternatives is accumulated in the human brain at varying rates, and a decision is made when such evidence reaches a particular boundary [46,47]. In the case of learning non-native speech sound categories like Mandarin tone categories, as a participant hears a stimulus, they begin accumulating evidence towards all four response options (e.g., high-flat, low-rising, low-dipping, high-falling). Each of the four response options has its own decision threshold, with higher thresholds requiring more evidence to be accumulated before the decision will be made, reflecting more cautious responding. Evidence is also accumulated toward each threshold at its own rate, with faster rates reflecting higher quality of evidence extracted from the stimulus. Below we consider the possibility that WM relates to these two components of the decision process.

The classical literature on DDMs has focused almost exclusively on binary decision-making in static settings and typically focuses on group-level analyses rather than heterogeneity across individuals. Recently, Paulon et al. [52] extended these models significantly, accounting for situations with more than two decision alternatives, heterogeneity across individuals, and longitudinal evolution of the decision-making processes by considering individual-specific and

time-varying accumulators of evidence. As such, we will examine decision processes over time with estimates at both the group and *individual subject level*.

Rate of evidence accumulation

We predict that more WM resources may enable learners to acquire information from the stimulus more quickly, thereby reducing the perceived difficulty of the task and effort needed to learn. The rate of evidence accumulation reflects the quality of information extracted from the stimulus, with faster rates reflecting a faster evidence accumulation process. The evidence accumulation process may also reflect efficiency of retrieval or access to categorization exemplars or other representations in memory. Faster evidence accumulation rates are associated with motivation and better task performance [53]. Prior work has demonstrated that evidence accumulation rates are related to WM abilities, with faster evidence accumulation associated with higher WM capacity [54,55].

Response caution

We predict that more WM resources may allow learners to be more cautious and less impulsive in their responses and to collect more evidence for a particular category response before making a decision. Response caution is reflected in the decision threshold. Higher thresholds reflect more cautious responses that need more evidence before a decision is made, whereas lower thresholds reflect more impulsive responses based on less evidence [56]. More difficult tasks result in more cautious response patterns, requiring that participants gather more information to make decisions [47,57]. Individuals with higher WM capacity may have sufficient resources to gather and consult more information during decision making. As such, they may be more cautious in their responses, gathering more information to hold in WM as they learn to make more accurate decisions. This may ensure that the learner builds up enough of a

representation of the stimulus before they make a response and, thus, enhance learning.

Alternatively, individuals with higher WM capacity may have sufficient resources to maintain similar decision thresholds as individuals with lower WM capacity, enabling them to respond faster without making sacrifices in performance.

Summary

To summarize, we examine the relationship between WM capacity and non-native Mandarin tone speech category learning in an extended training task with three sessions separated by one and two months, respectively. To gain mechanistic insights on the putative relationship between WM and individual differences in category learning over time, we assess behavior from multiple angles. Specifically, we examine how initial and later learning performance, maintenance of performance across delays, generalization to novel talkers, rate of evidence accumulation, and response caution are related to WM capacity (Table 1).

Methods

Participants completed three sessions of Mandarin tone category learning separated by at least one and two months (Session 1 to 2: $M = 32.1$ days, $SD = 0.68$, range 31.7-35.6 days; Session 2 to 3: $M = 61.4$ days, $SD = 2.56$, range 56.6-70.9 days). Data from the first session appeared in a previously published study [10], and the second and third sessions have not appeared elsewhere.

Participants

Participants were adults ages 18-35 recruited from Prolific (prolific.co) and participated via Gorilla Experiment Builder [58]. A total of 198 participants completed session 1 (99 Female (F), 99 Male (M), $M = 25.0$ years, $SD = 4.97$). Three participants were excluded because they

did not follow instructions on the WM task, leaving a total of 195 participants in session 1 (98 F, 97 M, $M = 24.9$ years, $SD = 4.89$). There was substantial attrition from session-to-session, and we excluded participants who did not complete all sessions – 153 completed session 2 (70 F, 83 M, $M = 24.9$ years, $SD = 5.05$), and 107 completed session 3 (47 F, 60 M, $M = 24.8$ years, $SD = 5.07$). Participants who completed only one or two sessions did not differ in WM or categorization accuracy compared to those who completed all sessions (Supporting Information, Figure S1).

Participants completed a language history questionnaire prior to participating. All participants were native speakers of non-tonal languages and reported no prior experience with any tonal languages, including Mandarin. Participants were given a sound check before the start of each session to ensure they could hear the sounds and were wearing headphones. Participants received \$10/session for their participation (total up to \$30 across three sessions). Informed consent was obtained from all participants. The study protocol was approved by the Institutional Review Board at the University of Pittsburgh.

Stimuli

The stimuli were natural speech productions recorded from four native speakers (2 M, 2 F) of Mandarin Chinese (Figure 1A). Each tone category (e.g., high-flat, low-rising, low-dipping, and high-falling) was produced by each speaker in five syllable contexts (/bu/, /di/, /lu/, /ma/, and /mi/) for a total of 80 stimuli (20/category). The stimuli from two speakers (1 F, 1 M) were used during the training blocks and the stimuli from the other speakers (1 F, 1 M) were withheld for the generalization block. The same 40 generalization stimuli were presented in the generalization block of each session and participants never received feedback about these stimuli. To reduce incidental differences in duration across categories, the stimuli were duration-normalized to 440

ms and RMS-amplitude normalized to 70 dB. The stimuli are shown in Figure 1A in a two-dimensional space (relative pitch, pitch change) that can be used to separate the stimuli into categories and is linked to neural representations of these categories [59,60].

Fig. 1. Stimuli and procedure. A. Two-dimensional representation of stimuli used during category learning and generalization with colors reflecting different tone categories. B. Session procedure. C. Task procedure.

Procedure

Category learning

Participants completed three separate sessions of category learning (Figure 1B). Sessions 1 and 2 were separated by one month. Sessions 2 and 3 were separated by two months. In each session, participants completed six blocks of an identical category learning task and an additional generalization block with different stimuli and no feedback. The stimuli were the same across sessions. Participants never received feedback about the generalization stimuli. At the beginning of the experiment, participants were told that they would be grouping sounds into different categories based on corrective feedback. They were not given any specific instructions about the stimuli or what might differentiate the categories from one another.

In the category learning task, there were six blocks of 40 trials each. In the generalization task, there was one block of 40 trials. Participants heard the 440 ms duration sound, followed by a prompt about the category identity (“Which category?”) (Figure 1C). They pressed the 1, 2, 3, and 4 buttons on the keyboard to respond. Participants received trial-by-trial feedback in the category learning task where they were informed about whether their decision was ‘Correct’ or

‘Incorrect.’ The feedback was presented immediately for 750 ms. Participants did not receive feedback in the generalization task. In both tasks, there was an intertrial interval of 1 sec.

Working memory capacity

In the first session, participants first completed the category learning and generalization blocks and then completed an operation span task [61] as a measure of WM capacity. Participants were shown simple arithmetic problems and reported whether the presented solutions were correct or incorrect (e.g., $(1 + 7) \times 2 = 16$) and were then shown a letter on the screen (e.g., A). A sequence of these arithmetic problems and letters from three to seven items in length made up a trial. After a full sequence was presented, participants were instructed to recall the letters presented in order. There were 15 trials. Participants’ WM capacity was calculated based on the OSPAN score – the sum of the length of all correctly recalled spans. For example, if a participant correctly recalled a sequence of four letters (e.g., A, I, D, F), four points were added to their score. The minimum possible OSPAN score is 0 and the maximum possible OSPAN score is 75. We did not filter scores based on accuracy on the arithmetic problems [62] and participants were generally very accurate ($M = 85\%$, $SD = 14\%$; Supporting Information Figure S2).

Drift diffusion modeling

We applied a variant of the DDMs developed in Paulon et al. [52]. The model estimates the evidence accumulation rate (i.e., drift) $\mu_{d,s}$ for each combination of decision response d and stimulus category s and decision thresholds (i.e., boundaries) b_d for each decision response d . Additionally, the model also fits offset parameters δ_s for each stimulus category, which characterize the times taken by the actions that are not directly relevant to the actual decision-making processes (e.g., the time required to encode the s -th stimulus before evidence

accumulation begins, to press a computer key, to record a response after a decision is reached, etc.). The model lets the parameters $\mu_{d,s}$, b_d and δ_s to vary between participants, which accommodates the substantial variability across participants. Importantly, the model also allows $\mu_{d,s}$ and b_d to evolve smoothly over time (across training blocks), explaining the changes in the decision-making processes as the participants learn over time. We allowed the drift rates to vary across both stimulus category and response and assume that participants gather evidence towards each of the four possible response options at different rates depending on the true identity of the stimulus category. The decisions participants make in this task are tied directly to the sound category. Exemplars from within a sound category share characteristics and differ from exemplars from other sound categories. Due to the stimulus characteristics, participants may accumulate evidence at different rates for the different stimulus-response combinations. Boundaries only varied across response and different levels of response caution were not dependent on the true stimulus category.

The data were filtered to exclude very fast and very slow responses by removing the top and bottom 1% of trials based on reaction time. The remaining data, comprising both correct and incorrect trials, were used to estimate the parameters. Since gradual improvements in making correct decisions characterize learning, in our discussions below, we emphasize heavily on inferring the drift rates associated with successful identification of the stimulus ($\mu_{d,s}$ for correct responses with $s = d$). Consideration of all responses does not change the overall results (see Supporting Information).

We adopted a Bayesian framework for these analyses, assigning priors to the parameters and relying on samples drawn from the posterior using a Markov chain Monte Carlo (MCMC) algorithm for estimation and inference. The algorithm was run for 6,000 iterations with the initial

2,000 iterations discarded as burn-in. The remaining samples were further thinned by an interval of 5 to reduce autocorrelation. MCMC diagnostics such as trace-plots of the parameters, Geweke test for stationarity of the chains, etc. indicated no convergence or mixing-related issues. Posterior predictive checks indicated good model fit. Finally, posterior means are reported as point estimates and pointwise credible intervals are used to assess uncertainty. For more details on the implementation of these models, see Supporting Information.

Data availability

The stimulus materials, data, and analysis code are publicly available through the Open Science Framework repository and can be accessed online [64]. Data were visualized and analyzed using R, version 4.3.1 [65] and the following R packages: *tidyverse*, version 1.3.2, [66], *ggplot2*, version 3.4.3 [67], *ggthemes*, version 4.2.4 [68], *lddmm*, version 0.4.2 [69], *lme4*, version 1.1.34 [70], *lmerTest*, version 3.1.3 [71], *rstatix*, version 0.7.2 [72].

Results

Learning performance

On average, participants learned the Mandarin tone categories with substantial individual variability in performance (Figure 2A). For context, we also plot the reaction times (Figure 2B). We note that for visualization of performance across blocks, we grouped participants by their WM scores based on a median split ($Mdn = 46$), with values equal to or higher than the median defined as high WM and values lower than the median being defined as low WM. The analyses were conducted using raw OSPAN scores as a continuous variable with linear mixed effects models using the *lme4* package in R [70] and are also shown (Figure 2C).

Fig 2. Working memory and learning performance across all participants. A. Accuracy and B. Reaction times after removing the shortest and longest 1% of responses. Error bars reflect *SEM*. For purposes of illustration, high and low working memory groups are defined based on a median split of working memory (OSPAN) scores. C. Relation between OSPAN score and proportion correct across blocks and sessions for all participants.

We examined the extent to which WM capacity, indexed by the OSPAN score, was associated with performance in the category learning task. We used linear mixed effects models with session (as categorical variable), block, WM capacity, all possible interactions as fixed effects, participant (intercept) as a random effect, and average accuracy across a block as the continuous outcome variable. Session 1 was treated as the baseline session. Full results are presented in Table 2.

Table 2. Summary of results on WM capacity and category learning performance.

	β	<i>SE</i>	<i>p</i>
Intercept	26.0	5.09	< .0001
OSPAN	-0.011	0.11	.92
Block	1.84	0.62	.0032
Session 2	6.51	3.43	.058
Session 3	14.5	3.43	< .0001
OSPAN * Block	0.055	0.013	< .0001
OSPAN * Session 2	0.31	0.074	< .0001
OSPAN * Session 3	0.22	0.074	.0026
Block * Session 2	-0.28	0.88	.75
Block * Session 3	-1.12	0.88	.20
OSPAN * Block * Session 2	-0.053	0.019	.0057
OSPAN * Block * Session 3	-0.032	0.019	.095

Overall, accuracy improved linearly across blocks in all sessions ($\beta_{\text{Block}} = 1.84$, *SE* = 0.62, *p* = .0032; $\beta_{\text{Block*Session2}} = -0.28$, *SE* = 0.88, *p* = .75; $\beta_{\text{Block*Session3}} = -1.12$, *SE* = 0.88, *p* = .20)

and improved marginally in session 2 from session 1 ($\beta_{\text{Session2}} = 6.51$, $SE = 3.43$, $p = .058$) and significantly in session 3 from session 1 ($\beta_{\text{Session3}} = 14.5$, $SE = 3.43$, $p < .0001$).

Collapsing across blocks, the relationship between WM score and accuracy was not significant in session 1 ($\beta_{\text{OSPAN}} = -0.011$, $SE = 0.11$, $p = .92$), but was significantly stronger in sessions 2 and 3 ($\beta_{\text{OSPAN}*\text{Session2}} = 0.31$, $SE = 0.074$, $p < .0001$; $\beta_{\text{OSPAN}*\text{Session3}} = 0.22$, $SE = 0.074$, $p = .0026$). Importantly, the relationship between WM score and accuracy interacted with both block and session. In session 1, there was a positive relationship between WM and accuracy that became stronger across blocks ($\beta_{\text{OSPAN}*Block} = 0.055$, $SE = 0.013$, $p < .0001$). A one unit increase in WM score was associated with an additional 0.055% increase in accuracy in each block. While in the first block, the relationship between WM score and accuracy was very weak (0.044%), by the final block, the relationship was clearly positive (0.32%). As a reminder, WM scores could range from 0 to 75, so even a relatively modest increase in WM score of 10 points would be associated with an additional increase in accuracy of 3.2% in the final block of session 1. A larger difference in WM score of 30 points would be associated with an additional increase in accuracy of 9.6% in this block.

One month later, in session 2, there was a positive relationship between WM and accuracy. While the relationship between WM and accuracy became stronger across blocks, the relative change was significantly smaller than in session 1 ($\beta_{\text{OSPAN}*Block*\text{Session2}} = -0.053$, $SE = 0.019$, $p = .0057$). In session 2, a one unit increase in WM score was associated with an additional 0.002% increase in accuracy in each block. Across blocks, the relationship between WM score and accuracy was similar to session 1 (range 0.30% - 0.31%).

Two months after session 2, in session 3, there was a positive relationship between WM and accuracy that became stronger across blocks in a way that was not significantly different

from session 1 ($\beta_{\text{OSPAN*Block*Session3}} = -0.032$, $SE = 0.019$, $p = .095$). In session 3, a one unit increase in WM score was associated with an additional 0.023% increase in accuracy in each block. In the first block, the relationship between WM score and accuracy was 0.24% and by the final block, the relationship was similar to the final blocks of the other sessions (0.35%).

Taken together, we found that working memory ability was positively associated with speech category learning accuracy across training sessions, becoming relatively stronger across blocks in sessions 1 and 3 and was stable in session 2. While in the very initial stages of learning, WM score was not significantly related to accuracy (0.044% in first block of session 1), by the end of session 1 and persisting through the other sessions, WM score was positively related to accuracy (range 0.24% to 0.35%). The positive relationship between WM ability and performance emerged within the first session and remained relatively stable throughout follow up sessions 2 and 3.

Learners and non-learners

Importantly, we also aimed to understand if the relationship between WM capacity and accuracy was present when considering only participants who learned the categories. We identified participants who performed at or below chance levels in the final block of session 3 (defined by 95% cumulative binomial probability, 40 trials, 0.25 probability of correct response = 25% +/- 10%) as ‘non-learners’ and those who performed better than chance as ‘learners’ (Figure 3A). Even though the non-learners were defined based on their accuracy in the final block of session 3, non-learners had significantly lower accuracy throughout all blocks (Bonferroni-corrected pairwise comparisons, $p < .001$), except for the first block of session 1 ($p = .078$). This underlines the necessity of considering learners separately from non-learners.

Fig 3. Working memory and learning performance across learners and non-learners. A.

Accuracy and B. Reaction times after removing the shortest and longest 1% of responses. Error bars reflect *SEM*. For purposes of illustration, high and low working memory groups are defined based on a median split of working memory (OSPAN) scores. Groups are additionally separated into learners and non-learners based on session 3 block 6 accuracy and whether it was greater (learners) or less than (non-learners) chance performance. C. Relation between OSPAN score and proportion correct across blocks for learners only.

A total of 32% (34/107) of participants were classified as non-learners. WM scores for learners ($M = 44.1$) were marginally higher than non-learners ($M = 36.7$; $t(65.6) = 1.79$, $p = .078$, 95% CI [-0.84, 15.6]). This may indicate that individuals with lower WM may be more likely to be non-learners. It is important to note that we cannot completely rule out that non-learners with seemingly lower WM may have been generally disengaged in the experiment, leading to poorer performance in both the WM task and the category learning task. If this is the case, WM scores for these individuals may not reflect their true WM abilities. As post-hoc evidence that some participants may have been disengaged across tasks, we found that learners ($M = 90\%$) performed better than non-learners ($M = 79\%$) at identifying the arithmetic equations as correct or incorrect in the WM task ($t(44) = 3.49$, $p = .0011$, 95% CI [4.58, 17.1]; Supporting Information, Figure S2). In the following analyses, we focus on the remaining 68% (73/107) of participants who are operationally defined as ‘learners’ in the category learning task. Because the accuracies of non-learners were within a low and highly restricted range by definition, we examined the relationship between WM score and accuracy for learners only.

To understand if the relationship between WM and category learning performance was present when examining learners only, we ran the same linear model analysis with learners only (Figure 3B; Table 3). Session 1 was treated as a baseline.

Table 3. Summary of results on WM capacity and category learning performance across groups.

	β	SE	p
Intercept	27.5	4.94	< .0001
OSPAN	-0.029	0.10	.78
Block	3.46	0.71	< .0001
Session 2	13.0	3.90	.00090
Session 3	25.0	3.90	< .0001
OSPAN * Block	0.050	0.015	.00059
OSPAN * Session 2	0.30	0.081	.00018
OSPAN * Session 3	0.18	0.081	.023
Block * Session 2	-0.56	1.00	.58
Block * Session 3	-1.12	1.00	.26
OSPAN * Block * Session 2	-0.058	0.021	.0048
OSPAN * Block * Session 3	-0.043	0.021	.039

Of critical interest is whether WM score and accuracy were still positively related when examining only those who learned the categories. In session 1 ignoring block, the relationship between WM and accuracy was not significant ($\beta_{\text{OSPAN}} = -0.029$, $SE = 0.10$, $p = .76$). However, this relationship became stronger across blocks ($\beta_{\text{OSPAN} \times \text{Block}} = 0.050$, $SE = 0.015$, $p = .00059$; $\beta_{\text{OSPAN} \times \text{Block} \times \text{NonLearners}} = -0.043$, $SE = 0.026$, $p = .11$). A one unit increase in WM score was associated with an additional 0.050% increase in accuracy in each block for learners. By the final block of session 1, a one unit increase in WM score was associated with a 0.27% increase in accuracy for learners.

In session 2, the relationship between WM and accuracy was positive and significantly stronger than session 1 ($\beta_{\text{OSPAN} \times \text{Session2}} = 0.30$, $SE = 0.081$, $p = .00018$). Ignoring block, a one unit

increase in WM was associated with an increase in accuracy of 0.27%. This relationship was relatively stable, becoming mildly weaker across blocks. The relationship between WM score and accuracy across blocks was significantly different from session 1 ($\beta_{\text{OSPAN*Block*Session2}} = -0.058$, $\text{SE} = 0.021$, $p = .0048$). A one unit increase in WM score was associated with an additional 0.008% decrease in accuracy in each block. By the final block of session 2, a one unit increase in WM score was associated with a 0.23% increase in accuracy for learners.

In session 3, the relationship between WM and accuracy was positive and significantly stronger than session 1 ($\beta_{\text{OSPAN*Session3}} = 0.18$, $\text{SE} = 0.081$, $p = .023$). Ignoring block, a one unit increase in WM was associated with an increase in accuracy of 0.15%. The relationship was relatively stable, becoming mildly stronger across blocks. The relationship between WM score and accuracy across blocks was significantly different from session 1 ($\beta_{\text{OSPAN*Block*Session3}} = -0.043$, $\text{SE} = 0.021$, $p = .039$). A one unit increase in WM score was associated with an additional 0.007% increase in accuracy in each block for learners. By the final block of session 3, a one unit increase in WM score was associated with a 0.20% increase in accuracy for learners.

Among learners only, higher WM ability was associated with better non-native speech category learning performance. This relationship emerged within the first session and was persistent across sessions 2 and 3 and, unsurprisingly, was slightly weaker than the relationship including all participants. The slope of the relationship between WM score and accuracy was 0.27% in the final block of session 1, 0.23% in the final block of session 2, and 0.20% in the final block of session 3.

Maintenance of Category Knowledge Over Time

By examining learning across several sessions separated by one and two months, respectively, we can assess the maintenance of categorization performance and category

knowledge over time. We assessed category knowledge maintenance by comparing adjacent training blocks that were either separated by no delay (i.e., blocks 5 and 6 of the same session) or a delay of one or two months (i.e., block 6 of one session and block 1 of the next session). Performance across these blocks and sessions for learners and non-learners separately is shown in Figure 4A. Because we are interested in how knowledge is retained over time, we focus our analyses only on learners.

Fig 4. Working memory and performance maintenance. A. Error bars reflect *SEM*. For purposes of illustration, high and low working memory groups are defined based on a median split of working memory (OSPAN) scores. Groups are additionally separated into learners and non-learners based on session 3 block 6 accuracy and whether it was greater (learners) or less than (non-learners) chance performance. B. Relation between OSPAN score and percent difference from block 5 to 6 within a session (No Delay) and block 6 to block 1 (Delay) for learners only.

Learners were somewhat able to maintain their category knowledge after a month or more of no additional training. Between sessions 1 and 2, accuracy fell an average of 7.2% (58.0% in block 6 to 50.9% in block 1) and between sessions 2 and 3, accuracy fell an average of 7.0% (65.6% in block 6 to 58.6% in block 1). In contrast, accuracy was relatively stable in the end of the sessions with accuracy increasing by 1.8% in session 1 (56.0% in block 5 to 57.8% in block 6) and by 0.3% in session 2 (65.2% in block 5 to 65.5% to block 6).

The ability to maintain category performance in adjacent blocks both with no delay (i.e., block 5 vs block 6) and after a one- or two-month delay (i.e., block 6 and block 1 of the next

session) was unrelated to learners' WM capacity (Figure 4B, Table S2). We examined the percent difference between adjacent blocks across sessions using a linear mixed effects model with time (session 1 to 2 as baseline), delay (delay as baseline), WM score (OSPAN), and all interactions as fixed effects and participant as a random effect. WM was unrelated to the retention of performance across sessions 1 to 2 ($\beta_{\text{OSPAN}} = 0.032$, $SE = 0.068$, $p = .64$) and 2 and 3 ($\beta_{\text{OSPAN} \times \text{Sessions 2 to 3}} = 0.020$, $SE = 0.094$, $p = .83$). The relationship between WM and retention did not depend on whether there was a delay of a month ($\beta_{\text{OSPAN} \times \text{Delay}} = 0.012$, $SE = 0.094$, $p = .90$) or two months ($\beta_{\text{OSPAN} \times \text{Delay} \times \text{Sessions 2 to 3}} = -0.14$, $SE = 0.13$, $p = .29$).

Generalization to Novel Speakers

By examining how participants respond to new speakers about which they never receive feedback, we can assess the generalizability of their category knowledge. We first calculated a generalization score by subtracting the final training block accuracy from the test accuracy. Overall, learners were successful at generalizing their knowledge to the new speakers (Figure 5A). Once again, we focus our analyses on learners as there is no clear category knowledge for non-learners to generalize. We examined whether generalization performance across sessions was related to WM capacity by examining session (session 1 as baseline), WM score (OSPAN), and the interaction between session and WM score as fixed effects and participant as a random effect (Figure 5B, Table S3).

Fig 5. Working memory and category generalization. A. Error bars reflect *SEM*. For purposes of illustration, high and low working memory groups are defined based on a median split of working memory (OSPAN) scores. Groups are additionally separated into learners and non-learners based on session 3 block 6 accuracy and whether it was greater (learners) or less than

(non-learners) chance performance. B. Relation between OSPAN score and generalization test score (mean generalization accuracy – mean block 6 accuracy) across sessions for learners only.

WM ability was not significantly related to learners' generalization ability in session 1 ($\beta_{\text{OSPAN}} = 0.080$, $\text{SE} = 0.055$, $p = .14$). There were no significant differences in the relationship between WM and generalization accuracy in sessions 1 and 2 ($\beta_{\text{OSPAN} \times \text{Session2}} = -0.049$, $\text{SE} = 0.076$, $p = .52$) or sessions 1 and 3 ($\beta_{\text{OSPAN} \times \text{Session3}} = -0.085$, $\text{SE} = 0.076$, $p = .26$). Overall, these results demonstrate that, among learners, WM ability is not significantly related to the ability to generalize Mandarin tone category knowledge to novel speakers.

Decision processes

We examined participants' decision processes based on the parameters from the drift diffusion models. We focus on the evidence accumulation rate (i.e., drift rate; Figure 6A) and decision threshold (i.e., boundary; Figure 6C) parameters. As these are Bayesian analyses, we interpret differences between groups where there is no overlap in the 95% credible intervals. We estimated the parameters for each individual and block, separately across sessions, with all subjects together (i.e., both learners and non-learners). As in prior work, we focus on the results for accumulators where the stimulus category is the same as the response category (i.e., correct responses) [50]. This allows for examination of decision processes at play on trials where participants made correct responses. The overall pattern of results does not change when examining responses from all accumulators (Supporting Information: Figure S3, Table S4).

Fig 6. Working memory and decision processes. A and C: error bars reflect 95% credible intervals. For purposes of illustration, high and low working memory groups are defined based

on a median split of working memory (OSPAN) scores. Groups are additionally separated into learners and non-learners based on session 3 block 6 accuracy and whether it was greater (learners) or less than (non-learners) chance performance. B and D: relation between OSPAN score and evidence accumulation rate and decision threshold for learners only.

First, we note the difference between learners and non-learners. Learners had higher evidence accumulation rates and higher decision thresholds than non-learners. In learners, the evidence accumulation rates increased over time, indicating that they became faster at accumulating evidence towards the correct decision. In contrast, the evidence accumulation rates in non-learners were low and flat throughout training, providing evidence of their general disengagement from the task. The decision thresholds were lower in non-learners than learners throughout the sessions deviating from one another after the very first block of training, indicating that non-learners needed less evidence to make their decision. This pattern may indicate that non-learners' decisions were based on optimizing speed rather than categorization accuracy.

Critically, our modeling approach enables estimation of the decision parameters at the *individual participant level*, allowing for examination of how these parameters relate to WM capacity. To understand how decision parameters differed based on WM in learners, we ran separate linear mixed effects models on the two parameters with block, session, WM score (OSPAN), and all interactions as fixed effects and participant as a random effect. Session 1 was treated as a baseline. Full results are shown in Tables 5 and 6. We focus on the results on the relationship between WM capacity and evidence accumulation rates and decision thresholds.

Table 5. Summary of results on WM capacity and evidence accumulation rate.

	β	SE	p
Intercept	0.15	0.17	.38
OSPAN	0.00075	0.0035	.83
Block	0.11	0.021	< .0001
Session 2	0.47	0.12	< .0001
Session 3	0.64	0.12	< .0001
OSPAN * Block	0.0013	0.00044	.0030
OSPAN * Session 2	0.0064	0.0024	.0084
OSPAN * Session 3	0.0072	0.0024	.0030
Block * Session 2	-0.066	0.030	.029
Block * Session 3	-0.051	0.030	.087
OSPAN * Block * Session 2	-0.00059	0.00062	.34
OSPAN * Block * Session 3	-0.00065	0.00062	.29

Overall, learners with higher WM capacity accumulated evidence more quickly towards the correct decision in each session (Figure 6B). In session 1, there was not a significant relationship between WM and evidence accumulation rate ($\beta_{\text{OSPAN}} = 0.00075$, $SE = 0.0035$, $p = .83$). However, the relationship became significantly stronger across blocks ($\beta_{\text{OSPAN} \times \text{Block}} = 0.0013$, $SE = 0.00044$, $p = .0030$). A one unit increase in WM score was associated with an increase in evidence accumulation rate of 0.0021 units for learners in the first block of session 1 and 0.0086 units for learners in the final block of session 1.

The strength of the relationship between WM score and evidence accumulation rate also increased across sessions ($\beta_{\text{OSPAN} \times \text{Session2}} = 0.0064$, $SE = 0.0024$, $p = .0084$; $\beta_{\text{OSPAN} \times \text{Session3}} = 0.0072$, $SE = 0.0024$, $p = .0030$). In session 2, a one unit increase in WM score was associated with an increase in evidence accumulation rate of 0.0071 units for learners and this relationship was not significantly different across blocks ($\beta_{\text{OSPAN} \times \text{Block} \times \text{Session2}} = -0.00059$, $SE = 0.00062$, $p = .34$). In session 3, a one unit increase in WM score was associated with an increase in evidence accumulation rate of 0.0079 units for learners and this relationship was not significantly different across blocks ($\beta_{\text{OSPAN} \times \text{Block} \times \text{Session3}} = -0.00065$, $SE = 0.00062$, $p = .29$).

Table 6. Summary of results on WM capacity and decision threshold.

	β	SE	p
Intercept	1.35	0.085	< .0001
OSPAN	-0.0024	0.0018	.17
Block	-0.014	0.013	.31
Session 2	0.0023	0.074	.98
Session 3	-0.049	0.074	.51
OSPAN * Block	0.00063	0.00028	.022
OSPAN * Session 2	0.0047	0.0015	.0022
OSPAN * Session 3	0.0035	0.0015	.021
Block * Session 2	-0.0091	0.019	.63
Block * Session 3	0.035	0.019	.068
OSPAN * Block * Session 2	-0.00074	0.00039	.058
OSPAN * Block * Session 3	-0.00066	0.00039	.090

In contrast, there was no clear relationship between WM capacity and decision thresholds in any session (Figure 6B). In session 1, a one unit increase in WM score was associated with a non-significant decrease in threshold of 0.0024 units for learners ($\beta_{\text{OSPAN}} = -0.0024$, $SE = 0.0018$, $p = .17$). The relationship between WM and threshold became slightly less negative across blocks in session 1 ($\beta_{\text{OSPAN} \times \text{Block}} = 0.00063$, $SE = 0.00028$, $p = .022$). A one unit increase in WM score was associated with a *decrease* in threshold of 0.0018 units for learners in the first block but an *increase* of 0.0014 units in the final block of session 1. Overall, in session 1, there was no clear relationship between WM score and decision threshold.

The relationship between WM and threshold differed in sessions 2 and 3 compared to session 1 ($\beta_{\text{OSPAN} \times \text{Session2}} = 0.0047$, $SE = 0.0015$, $p = .0022$; $\beta_{\text{OSPAN} \times \text{Session3}} = 0.0035$, $SE = 0.0015$, $p = .021$). However, this difference appears to stem from changing from a negligible negative relationship in session 1 to a negligible positive relationship in sessions 2 and 3. In session 2, one unit increase in WM score was associated with an increase in threshold of 0.0023 for learners, which did not significantly differ across blocks ($\beta_{\text{OSPAN} \times \text{Block} \times \text{Session2}} = -0.00074$, $SE = 0.00039$, $p =$

.058). In session 3, a one unit increase in WM score was associated with an increase in threshold of 0.0011 units for learners, which did not significantly differ across blocks ($\beta_{\text{OSPAN} \times \text{Block} \times \text{Session3}} = -0.00066$, $SE = 0.00039$, $p = .090$). In sum, decision thresholds did not strongly relate to WM capacity in any session.

Overall, learners with higher WM capacity had *faster* evidence accumulation rates. The relationship began to emerge in the first session and was clearly present in the second and third sessions. In contrast, learners' decision thresholds did not depend on WM capacity. Together, these results indicate that WM capacity impacts specific elements of decision-making differently across the trajectory of learning.

Discussion

We investigated non-native speech category learning in initial learning sessions and in two follow up sessions with one and two months between each session, respectively. We examined the extent to which WM capacity was related to initial and later learning sessions and in which ways (Figure 7). Considering all participants, higher WM was associated with better speech category learning across learning stages. Participants with higher WM may also have been more likely to learn the categories than participants with lower WM. When considering only individuals who performed at above-chance levels (i.e., learners), WM was associated with better performance by later blocks of initial acquisition (session 1) and in intermediate and later sessions (session 2-3) becoming somewhat weaker over time. WM ability was generally unrelated to maintenance of category knowledge over delays or generalization of category knowledge to new talkers. Finally, among learners, higher WM capacity was associated with faster evidence accumulation rates across learning sessions and was not associated with decision thresholds in any session.

Fig 7. Role of working memory in different stages of category learning. Visualization of relationship between behavioral measures and working memory for learners based on the regression model coefficients. Error bars reflect *SEM*.

Learners and non-learners

Our results demonstrate that simply grouping all participants together does not tell a complete story because some participants clearly do not demonstrate learning, performing at chance levels even after extensive training. However, swiftly removing these non-learners as is common practice in the field [31–34] may obscure parts of the story as well. Participants who performed at or below chance levels at the end of three sessions of training were consistently poor performers across all blocks and sessions had marginally lower WM scores than learners. Importantly, it is possible that non-learners with lower WM scores may have been generally disengaged in the experiment, performing poorly across all measures (Supporting Information, Figure S2). In support of the interpretation that non-learners were generally disengaged in the task, they had very low and flat evidence accumulation rates across learning, which may be indicative of general task disengagement [53,73].

Regardless of WM ability, we found that a substantial number of participants (32%) were classified as non-learners. These individuals returned for three separate sessions of the same task that they were unable to consistently perform above chance levels. It is important to consider participants' goals and motivation for completing the task and compare this with experimenter-defined goals. Whereas we instructed them to respond as accurately as possible, their goal seemed to be to respond as quickly as possible regardless of accuracy evidenced by non-learners'

much lower decision thresholds than learners. Decision thresholds (i.e., response caution) are related to the speed-accuracy tradeoff [56], with lower decision thresholds reflecting favoring speed over accuracy. As such, we interpret these low decision thresholds as a mark of these participants' disengagement in the category learning task. Importantly, favoring speed over accuracy is an adaptive strategy if your goal is not to *learn* the categories, but instead to complete the experiment as quickly as possible [74].

It is necessary to understand and adapt to the goals of our participants. This study was conducted using an online population, rather than a more typical convenience sample of college students leveraged in prior studies. This approach presents challenges, but also highlights that the goals and motivations to perform a simple experimental task may be different among a broader population than in student populations often examined in experimental psychology research.

It is important to understand how task disengagement is related to WM ability to understand potential interventions to improve learning. It is unclear if some non-learners want to learn, but they are unable to or if they are actively deciding to disengage from the task. Future work should include dynamic measures of task engagement, such as pupil dilation, to better understand how task engagement is related to WM and contributes to differences learning outcomes. If task disengagement is truly related to WM and we want to improve learning for individuals with lower WM, a first step should be ensuring that they are engaged in the task in the first place.

Together, these results highlight the importance of consideration of individual differences in learning. In particular, these results call for the need of special consideration of individuals who may be disengaged from the task. It is possible that a role that WM plays in learning is ensuring that resources are available for engagement in complex tasks.

671 **Initial learning and learning over time**

672 The main goal of the current study was to understand the role of working memory in
673 learning beyond initial acquisition. In line with prior work, we found that WM was positively
674 related to learning by the end of the first session [5–10]. The benefit of higher WM in initial
675 learning may stem from the ability to hold in mind many possible hypotheses which helps
676 learners home in on the best one and use it faster and more efficiently [5,10,11]. Our results are
677 in line with this prior work and suggest a role for WM in initial non-native speech category
678 acquisition.

679 As a novel contribution, our results extend these findings and demonstrate that among
680 participants who eventually learn the speech categories, WM was related to learning
681 performance starting at the end of session 1 and persisting in sessions 2 and 3. This pattern of
682 results conflicts with other work on speech category learning that demonstrates that given
683 multiple days of training, there is no clear link between WM and performance [16,17]. However,
684 these prior studies trained participants on across days separated by very short delays, rather than
685 delays of over a month or more without additional training. Our results indicate that WM helps
686 in initial acquisition of category knowledge, but individuals with lower WM may be able to
687 ‘catch up’ given more time. Specifically, our results provide some preliminary evidence that the
688 relationship between WM and non-native speech category learning may become weaker over
689 time. Lower WM is not a sentence to poor learning forever. As long as participants remain
690 engaged, they are able to learn.

691 This work also connects with prior investigations of learning from initial acquisition in
692 novices to overtrained performance in experts in both language (e.g., [23]) and other perceptual
693 contexts (e.g., [75]). While category representations start to emerge within a single session of

training [29,76], it is clear that further learning continues to shape representations and the networks supporting learning. For example, as individuals move from initial acquisition to highly experienced experts, there is a decrease in activation in sensory and frontal brain regions [77,78], potentially reflecting increased neural efficiency with learning. Research from visual category learning demonstrates that similar neural networks support initial and well-learned categorization behavior, but that these networks become more coordinated with extensive practice [79]. Together, these results highlight the need to understand how learning and the cognitive abilities and processes that support categorization change from the very initial novice stages of learning to behavior in overtrained experts. This is particularly relevant for speech and language learning contexts, where expert or even genuinely stable levels of performance are unlikely to emerge in a single training session.

Task difficulty and effort

We found that WM was consistently related to faster evidence accumulation among learners. These results are in line with prior work that demonstrates that evidence accumulation rates are linked to individual differences in WM [54,55]. Faster evidence accumulation rates reflect higher motivation [53], faster mobilization of attentional resources [80], and lower task difficulty [81–85].

We then might interpret the persistently higher evidence accumulation rate in learners with higher WM as reflective of heightened motivation, rapid mobilization of available attentional resources, and perhaps perceived difficulty of the task. That is, even when accuracies were similar, learners with higher WM may have achieved that level of performance with lower perceived difficulty and perceived or exerted effort. Conversely, lower evidence accumulation rates observed in learners with lower WM may be associated with slower mobilization of

motivational or attentional resources and more perceived difficulty in the task. Future research should clarify how WM relates to perceived difficulty and perceived and exerted effort during learning.

In summary, these results indicate that higher WM capacity is not a guarantee of better learning. Rather, it reflects better initial acquisition and general performance due to the ability to hold multiple hypotheses in mind and more rapid decision-making processes throughout learning. Lower WM also does not doom one to poor performance and, instead, lower WM may be linked to more time and resource-dependent decision processes which may be more effortful for the learner. Future work should address the perceived and exerted effort in learning and how this is related to WM.

Limitations

We note that there was significant attrition across sessions. Whereas 195 individuals completed the first session, only 107 returned for both follow up sessions. This is a challenge for longitudinal designs using online samples but is a necessary challenge to overcome to understand learning beyond initial acquisition. While we considered non-learners who completed all three sessions, it is also important to consider participants who failed to complete all parts of the experiment. In future work, it will be important to understand participants' reasons for returning or not returning to better understand what is motivating their performance in the task. Importantly, we found that WM did not differ based on how many sessions participants completed (Supporting Information, Figure S1A). This indicates that it was not just lower or higher WM individuals who failed to return for follow up sessions. There was also no difference in categorization accuracy based on the number of sessions participants completed. That is,

739 within the same session, those who completed one, two, or all three sessions did not differ in
740 their accuracy (Supporting Information, Figure S1B).

741 Another limitation of the current work is that we used a single measure of WM, measured
742 at a single timepoint [86]. Specifically, we used an operation span measure based on ability to
743 manipulate and remember a sequence of letters given a mathematical task interference.
744 Operation span is extensively used and is a highly reliable measure of WM [86,87]. Even still,
745 one measure likely does not reflect the true complexity of WM. Further, because of the nature of
746 the complex span task we used to assess WM capacity, it is possible that performance was
747 influenced by some combination of WM and long-term memory [88]. As a result, the observed
748 relationship between WM score and speech category learning performance may reflect the ability
749 to hold onto and manipulate information in WM *as well as* retrieve exemplars or rules from long-
750 term memory. However, it is important to note that measures that should theoretically be related
751 to long-term memory or activation of exemplars stored in memory (e.g., maintenance,
752 generalization) were not significantly related to WM score. Future studies should collect multiple
753 measures of WM including visuospatial and auditory WM as well as measures of long-term
754 memory to better understand how speech category learning relies on WM and long-term memory
755 abilities.

756 Finally, participants learned four difficult categories with minimal feedback (e.g.,
757 “correct” or “incorrect”). Because this kind of feedback is ambiguous when the response is
758 incorrect, it is possible that performance may have improved if we had provided full feedback
759 (e.g., “correct, that was category 1”). However, prior work has demonstrated that Mandarin tone
760 learning, as we examined here, is better with minimal feedback relative to full feedback [89].
761 Future studies will need to address the role of WM in learning with full and minimal feedback.

Conclusion

We examined non-native speech category learning, maintenance of category knowledge across sessions, generalization to novel talkers, and decision processes involved in learning. The results demonstrate that higher WM is not a guarantee of learning, nor is lower WM a sentence to long-term learning difficulties. WM is one important ability in supervised category learning. Here, we highlight the need for an approach that considers the stage of learning and whether participants eventually learn. By leveraging a drift diffusion modeling approach and examining behavior from several angles over time, we conclude that WM may help learners by facilitating rapid category acquisition in initial stages and enhanced performance during subsequent stages of learning due to rapid evidence accumulation that may reduce the effort needed to learn. These results have important implications for developing interventions to improve learning in naturalistic language contexts and understanding what it means to be engaged in a task.

774 **Acknowledgements**

775 All data, stimulus materials, and analysis code are publicly available at the Open Science
776 Framework and can be accessed at <https://doi.org/10.17605/OSF.IO/WDPYU>. Data from the
777 first session appeared in McHaney et al. (2021). Data from the second and third sessions have
778 not appeared previously.

779

References

1. Richler JJ, Palmeri TJ. Visual category learning. *Wiley Interdiscip Rev Cognitive Sci.* 2014;5: 75–94. doi:10.1002/wcs.1268
2. Holt LL, Lotto AJ. Speech perception as categorization. *Atten Percept Psychophys.* 2010;72: 1218–1227. doi:10.3758/app.72.5.1218
3. Baddeley A. Working memory. *Science.* 1992;255: 556–559. doi:10.1126/science.1736359
4. Baddeley AD, Hitch G. Working memory. *Exploring working memory: selected works of Alan Baddeley.* Taylor & Francis Group; 2017. pp. 43–79.
5. Sewell DK, Lewandowsky S. Attention and Working Memory Capacity: Insights From Blocking, Highlighting, and Knowledge Restructuring. *J Exp Psychology Gen.* 2012;141: 444–469. doi:10.1037/a0026560
6. Craig S, Lewandowsky S. Whichever way you Choose to Categorize, Working Memory Helps you Learn. *Q J Exp Psychol.* 2011;65: 439–464. doi:10.1080/17470218.2011.608854
7. Lewandowsky S. Working Memory Capacity and Categorization: Individual Differences and Modeling. *J Exp Psychology Learn Mem Cognition.* 2011;37: 720–738. doi:10.1037/a0022639
8. Lewandowsky S, Yang L-X, Newell BR, Kalish ML. Working memory does not dissociate between different perceptual categorization tasks. *J Exp Psychology Learn Mem Cognition.* 2012;38: 881–904. doi:10.1037/a0027298
9. Maddox WT, Chandrasekaran B, Smayda K, Yi H-G. Dual systems of speech category learning across the lifespan. *Psychol Aging.* 2013;28: 1042–56. doi:10.1037/a0034969
10. McHaney JR, Tessmer R, Roark CL, Chandrasekaran B. Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry. *Brain Lang.* 2021;222: 105010. doi:10.1016/j.bandl.2021.105010
11. Lloyd K, Sanborn A, Leslie D, Lewandowsky S. Why Higher Working Memory Capacity May Help You Learn: Sampling, Search, and Degrees of Approximation. *Cognitive Sci.* 2019;43: e12805. doi:10.1111/cogs.12805
12. Miyake A, Friedman NP. Individual Differences in Second Language Proficiency: Working Memory as Language Aptitude. In: Healy AF, Bourne LE, editors. *Foreign Language Learning: Psycholinguistic Studies on Training and Retention.* Mahwah, NJ: Lawrence Erlbaum Associates; 1998. pp. 339–364.
13. Baddeley A, Gathercole S, Papagno C. The phonological loop as a language learning device. *Psychol Rev.* 1998;105: 158–173. doi:10.1037//0033-295x.105.1.158

- 812 14. Cheung H. Nonword Span as a Unique Predictor of Second-Language Vocabulary Learning.
813 Dev Psychol. 1996;32: 867–873. doi:10.1037/0012-1649.32.5.867
- 814 15. Speciale G, Ellis NC, Bywater T. Phonological sequence learning and short-term store
815 capacity determine second language vocabulary acquisition. Appl Psycholinguist. 2004;25: 293–
816 321. doi:10.1017/s0142716404001146
- 817 16. Perrachione TK, Lee J, Ha LYY, Wong PCM. Learning a novel phonological contrast
818 depends on interactions between individual differences and training paradigm design. J Acoust
819 Soc Am. 2011;130: 461–472. doi:10.1121/1.3593366
- 820 17. Ingvalson EM, Nowicki C, Zong A, Wong PCM. Non-native Speech Learning in Older
821 Adults. Front Psychol. 2017;8: 148. doi:10.3389/fpsyg.2017.00148
- 822 18. Wang Y, Jongman A, Sereno JA. Acoustic and perceptual evaluation of Mandarin tone
823 productions before and after perceptual training. J Acoust Soc Am. 2003;113: 1033–1043.
824 doi:10.1121/1.1531176
- 825 19. Wang Y, Spence MM, Jongman A, Sereno JA. Training American listeners to perceive
826 Mandarin tones. J Acoust Soc Am. 1999;106: 3649–3658. doi:10.1121/1.428217
- 827 20. Wong PCM, Perrachione TK. Learning pitch patterns in lexical identification by native
828 English-speaking adults. Appl Psycholinguist. 2007;28: 565–585.
829 doi:10.1017/s0142716407070312
- 830 21. Xu Y, Gandour JT, Francis AL. Effects of language experience and stimulus complexity on
831 the categorical perception of pitch direction. J Acoust Soc Am. 2006;120: 1063–1074.
832 doi:10.1121/1.2213572
- 833 22. Llanos F, McHaney JR, Schuerman WL, Yi HG, Leonard MK, Chandrasekaran B. Non-
834 invasive peripheral nerve stimulation selectively enhances speech category learning in adults.
835 Npj Sci Learn. 2020;5: 12. doi:10.1038/s41539-020-0070-0
- 836 23. Reetzke R, Xie Z, Llanos F, Chandrasekaran B. Tracing the Trajectory of Sensory Plasticity
837 across Different Stages of Speech Learning in Adulthood. Curr Biol. 2018;28: 1419-1427.e4.
838 doi:10.1016/j.cub.2018.03.026
- 839 24. Roeder JL, Ashby FG. What is automatized during perceptual categorization? Cognition.
840 2016;154: 22–33. doi:10.1016/j.cognition.2016.04.005
- 841 25. Waldschmidt JG, Ashby FG. Cortical and striatal contributions to automaticity in
842 information-integration categorization. Neuroimage. 2011;56: 1791–1802.
843 doi:10.1016/j.neuroimage.2011.02.011

- 844 26. Bradlow AR, Akahane-Yamada R, Pisoni DB, Tohkura Y. Training Japanese listeners to
845 identify English /r/ and /l/: Long-term retention of learning in perception and production. *Percept*
846 *Psychophys.* 1999;61: 977–985. doi:10.3758/bf03206911
- 847 27. H  lie S, Waldschmidt JG, Ashby FG. Automaticity in rule-based and information-integration
848 categorization. *Atten Percept Psychophys.* 2010;72: 1013–1031. doi:10.3758/app.72.4.1013
- 849 28. H  lie S, Roeder JL, Ashby FG. Evidence for Cortical Automaticity in Rule-Based
850 Categorization. *J Neurosci.* 2010;30: 14225–14234. doi:10.1523/jneurosci.2393-10.2010
- 851 29. Ester EF, Sprague TC, Serences JT. Categorical Biases in Human Occipitoparietal Cortex. *J*
852 *Neurosci.* 2020;40: 917–931. doi:10.1523/jneurosci.2700-19.2019
- 853 30. Feng G, Li Y, Hsu S-M, Wong PCM, Chou T-L, Chandrasekaran B. Emerging Native-
854 Similar Neural Representations Underlie Non-Native Speech Category Learning Success.
855 *Neurobiology Lang.* 2021;2: 280–307. doi:10.1162/nol_a_00035
- 856 31. Folstein JR, Palmeri TJ, Gauthier I. Category Learning Increases Discriminability of
857 Relevant Object Dimensions in Visual Cortex. *Cereb Cortex.* 2013;23: 814–823.
858 doi:10.1093/cercor/bhs067
- 859 32. Hammer R, Sloutsky V, Grill-Spector K. Feature saliency and feedback information
860 interactively impact visual category learning. *Front Psychol.* 2015;6: 74.
861 doi:10.3389/fpsyg.2015.00074
- 862 33. Maddox WT, Chandrasekaran B. Tests of a dual-system model of speech category learning.
863 *Biling Lang Cognition.* 2014;17: 709–728. doi:10.1017/s1366728913000783
- 864 34. Rosedahl LA, Eckstein MP, Ashby FG. Retinal-specific category learning. *Nat Hum Behav.*
865 2018;2: 500–506. doi:10.1038/s41562-018-0370-z
- 866 35. Ell SW, Ashby FG. The effects of category overlap on information-integration and rule-
867 based category learning. *Percept Psychophys.* 2006;68: 1013–1026. doi:10.3758/bf03193362
- 868 36. Kaplan AS, Murphy GL. Category learning with minimal prior knowledge. *J Exp*
869 *Psychology Learn Mem Cognition.* 2000;26: 829–846. doi:10.1037/0278-7393.26.4.829
- 870 37. Roark CL, Chandrasekaran B. Individual variability in strategies and learning outcomes in
871 auditory category learning. In: Fitch T, Lamm C, Leder H, Tessmar K, editors. *Proceedings of*
872 *the 43rd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science
873 Society; 2021. pp. 154–160.
- 874 38. Roark CL, Holt LL. Perceptual dimensions influence auditory category learning. *Atten*
875 *Percept Psychophys.* 2019;81: 912–926. doi:10.3758/s13414-019-01688-6

- 876 39. DeCaro MS, Thomas RD, Beilock SL. Individual differences in category learning:
877 Sometimes less working memory capacity is better than more. *Cognition*. 2008;107: 284–294.
878 doi:10.1016/j.cognition.2007.07.001
- 879 40. DeCaro MS, Carlson KD, Thomas RD, Beilock SL. When and how less is more: reply to
880 Tharp and Pickering. *Cognition*. 2009;111: 397–403. doi:10.1016/j.cognition.2009.03.002
- 881 41. Tharp IJ, Pickering AD. A note on DeCaro, Thomas, and Beilock (2008): Further data
882 demonstrate complexities in the assessment of information-integration category learning.
883 *Cognition*. 2009;111: 411–415. doi:10.1016/j.cognition.2008.10.003
- 884 42. Diamond A. Executive Functions. *Annu Rev Psychol*. 2013;64: 135–168.
885 doi:10.1146/annurev-psych-113011-143750
- 886 43. McCabe DP, Roediger HL, McDaniel MA, Balota DA, Hambrick DZ. The Relationship
887 Between Working Memory Capacity and Executive Functioning: Evidence for a Common
888 Executive Attention Construct. *Neuropsychology*. 2010;24: 222–243. doi:10.1037/a0017619
- 889 44. Lenaert B, Ven V van de, Kaas AL, Vlaeyen JWS. Generalization on the Basis of Prior
890 Experience Is Predicted by Individual Differences in Working Memory. *Behav Ther*. 2016;47:
891 130–140. doi:10.1016/j.beth.2015.10.001
- 892 45. Wills AJ, Barrasin TJ, McLaren IPL. Working Memory Capacity and Generalization in
893 Predictive Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*. 2011.
894 pp. 3205–3210. doi:10.1016/j.cogsys.2005.02.002
- 895 46. Ratcliff R. A Theory of Memory Retrieval. *Psychological Review*. 1978;85: 59–108.
896 doi:10.1037/0033-295x.85.2.59
- 897 47. Ratcliff R, Smith PL, Brown SD, McKoon G. Diffusion Decision Model: Current Issues and
898 History. *Trends Cogn Sci*. 2016;20: 260–281. doi:10.1016/j.tics.2016.01.007
- 899 48. Nosofsky RM. Attention, Similarity, and the Identification-Categorization Relationship.
900 *Journal of Experimental Psychology: General*. 1986;115: 39–57.
- 901 49. Nosofsky RM, Palmeri TJ. An Exemplar-Based Random Walk Model of Speeded
902 Classification. *Psychol Rev*. 1997;104: 266–300. doi:10.1037/0033-295x.104.2.266
- 903 50. Roark CL, Paulon G, Sarkar A, Chandrasekaran B. Comparing perceptual category learning
904 across modalities in the same individuals. *Psychon B Rev*. 2021;28: 898–909.
905 doi:10.3758/s13423-021-01878-0
- 906 51. Heffernan EM, Adema JD, Mack ML. Identifying the neural dynamics of category decisions
907 with computational model-based functional magnetic resonance imaging. *Psychon B Rev*. 2021;
908 10-Jan. doi:10.3758/s13423-021-01939-4

- 909 52. Paulon G, Llanos F, Chandrasekaran B, Sarkar A. Bayesian Semiparametric Longitudinal
910 Drift-Diffusion Mixed Models for Tone Learning in Adults. *J Am Stat Assoc.* 2021; 14-Jan.
911 doi:10.1080/01621459.2020.1801448
- 912 53. Hübner R, Schlösser J. Monetary reward increases attentional effort in the flanker task.
913 *Psychon B Rev.* 2010;17: 821–826. doi:10.3758/pbr.17.6.821
- 914 54. Schmiedek F, Oberauer K, Wilhelm O, Süß H-M, Wittmann WW. Individual Differences in
915 Components of Reaction Time Distributions and Their Relations to Working Memory and
916 Intelligence. *J Exp Psychology Gen.* 2007;136: 414–429. doi:10.1037/0096-3445.136.3.414
- 917 55. Ester EF, Ho TC, Brown SD, Serences JT. Variability in visual working memory ability
918 limits the efficiency of perceptual decision making. *J Vision.* 2014;14: 2-Feb. doi:10.1167/14.4.2
- 919 56. Bogacz R, Wagenmakers E-J, Forstmann BU, Nieuwenhuis S. The neural basis of the speed–
920 accuracy tradeoff. *Trends Neurosci.* 2010;33: 16-Oct. doi:10.1016/j.tins.2009.09.002
- 921 57. Philiastides MG, Ratcliff R, Sajda P. Neural Representation of Task Difficulty and Decision
922 Making during Perceptual Categorization: A Timing Diagram. *J Neurosci.* 2006;26: 8965–8975.
923 doi:10.1523/jneurosci.1655-06.2006
- 924 58. Anwyl-Irvine A, Massonnié J, Flitton A, Kirkham N, Evershed J. Gorilla in our Midst: An
925 online behavioral experiment builder. *bioRxiv.* 2020; 438242. doi:10.3758/s13428-019-01237-x
- 926 59. Feng G, Gan Z, Wang S, Wong PCM, Chandrasekaran B. Task-General and Acoustic-
927 Invariant Neural Representation of Speech Categories in the Human Brain. *Cereb Cortex.*
928 2017;28: 3241–3254. doi:10.1093/cercor/bhx195
- 929 60. Li Y, Tang C, Lu J, Wu J, Chang EF. Human cortical encoding of pitch in tonal and non-
930 tonal languages. *Nat Commun.* 2021;12: 1161. doi:10.1038/s41467-021-21430-x
- 931 61. Turner ML, Engle RW. Is working memory capacity task dependent? *J Mem Lang.* 1989;28:
932 127–154. doi:10.1016/0749-596x(89)90040-5
- 933 62. Đokić R, Koso-Drljević M, Đapo N. Working memory span tasks: Group administration and
934 omitting accuracy criterion do not change metric characteristics. *Plos One.* 2018;13: e0205169.
935 doi:10.1371/journal.pone.0205169
- 936 63. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian
937 models. *Stat Comput.* 2014;24: 997–1016. doi:10.1007/s11222-013-9416-2
- 938 64. Roark CL, Chandrasekaran B. Stable, flexible, common, and distinct behaviors supporting
939 rule-based and information-integration category learning. *npj Science of Learning.* Available:
940 <https://doi.org/10.17605/OSF.IO/TFB6A>

- 941 65. Team RC. R: A language and environment for statistical computing. R Foundation for
942 Statistical Computing. Vienna, Austria; 2022. Available: <https://www.R-project.org/>
- 943 66. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the
944 Tidyverse. *J Open Source Softw.* 2019;4: 1686. doi:10.21105/joss.01686
- 945 67. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* 2016.
- 946 68. Arnold JB. *ggthemes: Extra Themes, Scales and Geoms for “ggplot2.”* 2018. Available:
947 <https://CRAN.R-project.org/package=ggthemes>
- 948 69. Paulon G, Sarkar A. *lddmm: Longitudinal Drift-Diffusion Mixed Models (LDDMM).* 2021.
949 Available: <https://CRAN.R-project.org/package=lddmm>
- 950 70. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4.
951 *J Stat Softw.* 2015;67. doi:10.18637/jss.v067.i01
- 952 71. Kuznetsova A, Brockhoff PB, Christensen RHB. {lmerTest} Package: Tests in Linear Mixed
953 Effects Models. *Journal of Statistical Software.* 2017;82: 26-Jan. doi:10.18637/jss.v082.i13
- 954 72. Kassambara A. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests.* 2021. Available:
955 <https://CRAN.R-project.org/package=rstatix>
- 956 73. Balsdon T, Wyart V, Mamassian P. Confidence controls perceptual evidence accumulation.
957 *Nat Commun.* 2020;11: 1753. doi:10.1038/s41467-020-15561-w
- 958 74. Hawkins GE, Brown SD, Steyvers M, Wagenmakers E-J. An optimal adjustment procedure
959 to minimize experiment time in decisions with multiple alternatives. *Psychon B Rev.* 2012;19:
960 339–348. doi:10.3758/s13423-012-0216-z
- 961 75. Diamond R, Carey S. Why Faces Are and Are Not Special: An Effect of Expertise. *J Exp*
962 *Psychology Gen.* 1986;115: 107–117. doi:10.1037/0096-3445.115.2.107
- 963 76. Feng G, Gan Z, Yi HG, Ell SW, Roark CL, Wang S, et al. Neural dynamics underlying the
964 acquisition of distinct auditory category structures. *Neuroimage.* 2021; 118565.
965 doi:10.1016/j.neuroimage.2021.118565
- 966 77. de Souza ACS, Yehia HC, Sato M, Callan D. Brain activity underlying auditory perceptual
967 learning during short period training: simultaneous fMRI and EEG recording. *Bmc Neurosci.*
968 2013;14: 8. doi:10.1186/1471-2202-14-8
- 969 78. Little DM, Thulborn KR. Correlations of cortical activation and behavior during the
970 application of newly learned categories. *Cognitive Brain Res.* 2005;25: 33–47.
971 doi:10.1016/j.cogbrainres.2005.04.015

- 972 79. DeGutis J, D’Esposito M. Network Changes in the Transition from Initial Learning to Well-
973 Practiced Visual Categorization. *Front Hum Neurosci.* 2009;3: 44.
974 doi:10.3389/neuro.09.044.2009
- 975 80. Sarter M, Gehring WJ, Kozak R. More attention must be paid: The neurobiology of
976 attentional effort. *Brain Res Rev.* 2006;51: 145–160. doi:10.1016/j.brainresrev.2005.11.002
- 977 81. Gold JJ, Shadlen MN. The Neural Basis of Decision Making. *Annu Rev Neurosci.* 2007;30:
978 535–574. doi:10.1146/annurev.neuro.29.051605.113038
- 979 82. Mulder MJ, van Maanen L, Forstmann BU. Perceptual decision neurosciences – A model-
980 based review. *Neuroscience.* 2014;277: 872–884. doi:10.1016/j.neuroscience.2014.07.031
- 981 83. O’Connell RG, Dockree PM, Kelly SP. A supramodal accumulation-to-bound signal that
982 determines perceptual decisions in humans. *Nat Neurosci.* 2012;15: 1729–1735.
983 doi:10.1038/nn.3248
- 984 84. van-Maanen L, Portoles O, Borst JP. The Discovery and Interpretation of Evidence
985 Accumulation Stages. *Comput Brain Behav.* 2021;4: 395–415. doi:10.1007/s42113-021-00105-2
- 986 85. van Maanen L, Forstmann BU, Keuken MC, Wagenmakers E-J, Heathcote A. The impact of
987 MRI scanner environment on perceptual decision-making. *Behav Res Methods.* 2016;48: 184–
988 200. doi:10.3758/s13428-015-0563-6
- 989 86. Conway ARA, Kane MJ, Al CET. Working memory span tasks: A methodological review
990 and user’s guide. *Psychonomic Bulletin & Review.* 2005;12: 769–786.
- 991 87. Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation span
992 task. *Behavior Research Methods.* 2005;37: 498–505.
- 993 88. Unsworth N, Engle RW. The Nature of Individual Differences in Working Memory
994 Capacity: Active Maintenance in Primary Memory and Controlled Search From Secondary
995 Memory. *Psychol Rev.* 2007;114: 104–132. doi:10.1037/0033-295x.114.1.104
- 996 89. Chandrasekaran B, Yi H-G, Maddox WT. Dual-learning systems during speech category
997 learning. *Psychon B Rev.* 2014;21: 488–95. doi:10.3758/s13423-013-0501-5
- 998
- 999