ABSTRACT

Title of Dissertation:	SIMPLY (NEURO-)STIMULATING: THE EFFECTS OF TRANSCUTANEOUS VAGUS NERVE STIMULATION ON PHONOLOGICAL AND LEXICAL TONE LEARNING INDEXED BY BEHAVIOR AND PUPILLOMETRY
	Nicholas Balint Pandza, Doctor of Philosophy, 2022
Dissertation directed by:	Professor Kira Gor, Second Language Acquisition

Mandarin lexical tone learning has repeatedly been identified as a difficult linguistic feature for non-native speakers of tonal languages like English, even for native English learners of Mandarin at high proficiencies (e.g., Pelzl et al., 2019b). Sound perception training has been shown to help native English speakers perceive lexical tone differences, but acquiring lexical tone as a feature still remains difficult, even after as many as 18 training sessions (Bowles et al., 2016; Chandrasekaran et al., 2010; Li & DeKeyser, 2017; 2019; Liu & Chandrasekaran, 2013; Wang et al., 1999; 2003; Wong et al., 2011; Wong & Perrachione, 2007). While much of the tone learning literature has focused on different training interventions to overcome learning plateaus, another type of intervention that could augment learning is non-invasive neurostimulation. Transcutaneous auricular vagus nerve stimulation (taVNS) is a type of safe, non-invasive neurostimulation that delivers electrical current to the ear canal that has been

shown to enhance cognition and learning (e.g., Jacobs et al., 2015). This dissertation investigated taVNS and its potential impact as tool to enhance Mandarin tone learning.

Participants in three groups, peristim taVNS, priming taVNS, and a sham taVNS control participated in a double-blind two-day Mandarin phonological and lexical tone training study. Behavioral data including accuracy and reaction time were collected, as was physiological data in the form of pupillometry due to its ties both to cognitive effort and the most well-studied taVNS mechanism of action, the production of norepinephrine. Active taVNS groups received stimulation before or during multiple training and testing tasks across the two days.

This body of work revealed: (1) priming and peristim administrations of taVNS differentially facilitated vocabulary learning of words with Mandarin tone, (2) priming and peristim administrations of taVNS differentially facilitated learning of new phonological tone categories, and (3) the effects of individual differences were substantially and differentially impacted by priming and peristim administrations taVNS, all results compared to a sham control. The evidence herein supports the potential of taVNS as a practical treatment intervention for enhancing language learning and reveals a number of considerations for its use and implementation to be explored in future research.

SIMPLY (NEURO-)STIMULATING: THE EFFECTS OF TRANSCUTANEOUS VAGUS NERVE STIMULATION ON PHONOLOGICAL AND LEXICAL TONE LEARNING INDEXED BY BEHAVIOR AND PUPILLOMETRY

by

Nicholas Balint Pandza

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Kira Gor, Chair
Associate Professor Jared Novick
Dr. Stefanie Kuchinsky
Dr. Jared Linck
Professor Min Wang, Dean's Representative

Dedication

To my parents, for encouraging my curiosity and making this possible.

Acknowledgements

This dissertation was made possible by the support, input, and encouragement of a great many people. I can't overstate the impact of the village of folks throughout this process cheering me on, and surely I have inadvertently left out a few important names and contributions below.

First and foremost, thanks are due to my committee for letting me pursue something so ambitious and for being constructively critical; every critique along the way made me a better scientist. My advisor, Kira Gor, has been patient, encouraging, and a positive force throughout my graduate program. She challenged me to think more deeply, be more (constructively) critical, and make more theoretical connections. She also taught me that babka had been missing from my life for too long. Jared Linck, a colleague and one-time supervisor, has made me a far better, more thoughtful statistician who can keep his eye on the bigger picture at least every now and again. Jared Novick, also one-time supervisor (albeit briefly), helped fuel my interest in executive function and individual differences near the start of my graduate career. I thank Min Wang for her thoughtful reading of this beast of a document and her useful and practical insights. Finally, Stefanie Kuchinsky, whose influence I can't do justice, let me barge into her office to pick her brain and bounce ideas, and she patiently taught me so many things—ranging from what pupillometry was to how to operate an MRI and more than a few things about R and statistics. She has made me a better scientist and a better person, and I simply can't thank her enough.

Second, thanks are due to the army of CASL/ARLIS folks that contributed to Phase 1 of TNT-Speech (and the larger TNT research effort) for which my work here directly and indirectly benefits. It has been a privilege be a part of this large team: Stefanie Kuchinsky and Polly O'Rourke as PIs and main drivers of the work, Val Karuzis, Ian Phillips, Matthew Turner, Sara

McConnell, Jarrett Lee, Henk Haarmann, Gina Calloway, Eric Pelzl, Alison Tseng, Meredith Hughes, Jason Struck, and Mike Johns. Some extra thanks are due to Mike for sending me down the GAMMs rabbit hole to begin with, and for the members of the GAMMs Club for helping me think through this brain-breaking method. I also have to thank Lara and a different Alison, who inadvertently pushed me to get the first part of this dissertation on paper and out the door.

Thanks broadly to the DARPA Targeted Neuroplasticity Training program: this material is based upon work supported by the Naval Information Warfare Center and Defense Advanced Research Projects Agency under Cooperative Agreement No. N66001-17-2-4009. The identification of specific products or scientific instrumentation is considered an integral part of the scientific endeavor and does not constitute endorsement or implied endorsement on the part of the author, DoD, or any component agency. The views expressed in this dissertation are those of the author and do not reflect the official policy of the Department of Army/Navy/Air Force, Department of Defense, or U.S. Government.

Next, I need to thank my other SLA professors: Robert DeKeyser for his support on earlier pieces of my tVNS research and his engaging course on individual differences; Nan Jiang for his stimulating courses on L2 and bilingual processing; Steve Ross for classes on statistics and validity that resonated with me early on; and Cathy Doughty and Mike Long for encouragement in smaller ways that helped me feel like I made the right choice in pursuing my PhD. I also have to thank my statistics professors in EDMS, in particular Greg Hancock, Laura Stapleton, and Jeff Harring, some of the best teachers I have ever met who filled my brain to the brim and gave me so many tools that will continue to pay dividends for the rest of my career. I also have to thank some CASL and ARLIS folks I haven't already mentioned: Ewa Golonka for her endless support and encouragement, Scott Jackson who made R make sense for the first time

(and put a number of tools in my stats toolbox after), Mike Bunting for his support and confidence in my ability, and a host of others that have also made me a better scientist along the way: Adam, Alexa, Carrie, Erica, Lana, Martyn, Petra, Susan, and Susannah to name a few.

Finally, thanks to the friends and colleagues who kept me sane throughout so many phases of my decade-long graduate career and one way or another helped me finally get to this point. Phillip Hamrick made conferencing actually…easy, and his friendship came with both intellectual and brotherly support for which I will forever count myself lucky. I have to thank Rebecca Sachs for a number of things, but in particular her brilliance and infectious passion for linguistics. I also have to thank some of my fellow Terps, in particular Sunhee Kim, Alia Lancaster, Eric Pelzl, and Başak Karataş (but also Fatima, Katie, Megan, Pete, Stephen, Rachel, and Yuichi, to name a few) for not only dinners, game nights, and/or advice, but also for being exceptional researchers themselves and keeping me motivated in and out of class. I have to thank some Hoyas for the same: Karen Feagin, Kristine Nugent, Lauren Park, and Janire Zalbidea helped keep me sane over the years, and I'm grateful and privileged by their friendships. So many others to thank, including, but not limited to: Aaron, Ariel, Charlie, Christie, Dan, Daniel, Eric, the GU PsyLab, Jennifer, Jesse, John, Justin, Luis, Kate, Mackenzie, Mari, Marisa, Marta, Mina, Mona, Nina, Noam, Ross, Sandra, Sarah, Tim, Toby, and Tommy. I also owe thanks to, in chronological order, Lara, Gretchen, Carolina, my BVC fam (especially Gwen, Angela, and Dorothea), Rebecca, Cristina, Ande, and Lourdes for their kindness and intellect and for helping me get from one stage to the next on this journey. Lastly, I thank Mikey, my FRA (feline research assistant), for keeping me company on many late nights, and I thank Janickren for giving me dedicated, collegial, and friendly work time to help me through the home stretch.

All remaining errors are my own. Thank you all for 'participanting'.

Table of Contents

Chapter 1: Introduction

Learning a second language (L2) is extremely difficult for adults, in part, because it places great demands on diverse memory and attentional mechanisms (Doughty & Long, 2003) and perceptual abilities (Sebastián-Gallés & Díaz, 2012). To enhance language learning, these mechanisms and abilities have commonly been targeted with behavioral training paradigms (e.g., Colflesh et al., 2016; Ingvalson et al., 2014), and to a lesser extent with neurostimulation techniques including transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS; Meinzer et al., 2014; Mottaghy et al., 1999). Another neurostimulation technique, vagus nerve stimulation (VNS), has long been studied among patient populations for its therapeutic benefits and has recently gained attention for its ability to improve memory, attention, and auditory processing (Borland et al., 2016; Vonck et al., 2014), but, until recently, it has not been systematically tested as a language learning intervention. Here I propose to analyze portions of a large corpus from a large-scale recent study testing the pairing of non-invasive transcutaneous auricular VNS (taVNS) with a behavioral training paradigm in which English speakers learned lexical tone contrasts in L2 Mandarin.

Phonology is widely regarded as the linguistic domain that presents the greatest difficulty for adult L2 learners (Moyer, 2014). The notorious difficulty of acquiring novel phonological features, or sound patterns, and mapping them to words in a new language is well documented, and accurately perceiving novel suprasegmental contrasts such as lexical tone presents a persistent challenge for even advanced learners (Pelzl et al., 2019b) and is the subject of a substantial literature (see Pelzl, 2019a, for a recent summary). A consistent finding among the

many studies that have sought to improve L2 lexical tone acquisition via behavioral training is that certain domain-general abilities and aptitudes largely determine the success of training. The recent findings that VNS can enhance several of these mechanisms make L2 lexical tone learning an ideal case for testing the efficacy of pairing non-invasive taVNS with a behavioral training paradigm.

I begin in Chapter 2 by providing a broad introduction of neurostimulation techniques in applied linguistics followed by a more detailed review of (ta)VNS and the use of pupillometry as a means of measuring the primary mechanism thought to be targeted by the intervention to enhance learning. I then describe the large double-blind study and its corpus of data from which I formed the dissertation study in Chapter 3. Chapter 4 describes a set of analyses to show the practical utility of two types of taVNS interventions, priming and peristimulus taVNS, on behavioral lexical tone learning outcomes and validate them with a physiological index of the primary taVNS mechanism of action. Chapter 5 describes a set of behavioral and physiological analyses to (1) investigate if the taVNS-facilitated improvements in lexical tone learning arise at the phonological level in addition to the lexical level, and (2) tease apart whether any taVNS-facilitation for phonological learning arises for all tones equally or differs among tone contours. Chapter 6 takes this a step further and reports a set of exploratory analyses to investigate whether there are aptitude- or experience- by-treatment interactions at the phonological and/or lexical level supporting the observed learning effects, using individual differences commonly associated with lexical tone learning: non-linguistic tone aptitude and musicality. Each of Chapters 4, 5, and 6 examine not only the behavioral outcomes of accuracy and response time, but also

pupillometry as an index of cognitive effort. Chapter 7 is a final discussion of all three sets of results together as well as the conclusions and implications for the dissertation as a whole.

Chapter 2: Literature Review

2.1 Neurostimulation in Language Science: A Broad Overview

Neurostimulation involves the application of stimulation (e.g., electrical, magnetic, tactile) that modulates the activity of the nervous system. There are a variety of neurostimulation techniques that target different neurocognitive mechanisms, many of which support language learning. The techniques most commonly used in language research, TMS and tDCS, involve placing neurostimulators (e.g., electrodes) on or above the surface of the scalp in order to affect cortical activity and have been evaluated for improving language and cognitive performance (Miniussi et al., 2008; Naeser et al., 2012; Reis et al., 2008).

TMS uses a strong magnetic field to induce electrical current in the brain under its position on the head. A pulse of current can temporarily disrupt neural activity and TMS has been used to simulate lesions to localize brain regions necessary for a given task, including several regions necessary for language processing (Pascual-Leone et al., 2000; Walsh & Pascual-Leone, 2003). TMS provides a high degree of accuracy in identifying where task-critical regions are in the brain (i.e., spatial localization), especially when combined with structural magnetic resonance imaging (MRI). TMS pulses can also repeated over an extended period of time (repetitive TMS; rTMS) to facilitate or inhibit neural activity (Miniussi et al., 2008). It has been used in people with aphasia to promote better language recovery (e.g., Finocchiaro et al., 2006), and healthy individuals to facilitate picture naming and other language tasks (e.g., Mottaghy et al., 1999; Sakai et al., 2002). Downsides of this method include a loud clicking noise associated with pulses (particularly problematic for auditory stimuli), the potential for a distracting

sensation across the scalp, and a lack of an ability to stimulate deeper cortical structures that are further from the scalp.

tDCS uses electrical currents applied at low intensities (1–2 milliamps (mA)) to facilitate or inhibit cortical excitability (DaSilva et al., 2015; Miniussi et al., 2008). tDCS has inferior spatial localization to TMS but is able to penetrate deeper brain structures (DaSilva et al., 2015) and is a silent intervention (Miniussi et al., 2008). Stimulation has been found to facilitate working memory (e.g., Ohn et al., 2008), long-term memory for word pairs (Marshall et al., 2004), and vocabulary learning (e.g., Meinzer et al., 2014). tDCS has also been found to promote language comprehension when, for example, used to inhibit cortical activity in the right Wernicke's area in subacute stroke patients (You et al., 2011).

Rather than targeting specific cortical areas directly, peripheral nerve stimulation (PNS) involves stimulating the peripheral branches of a cranial nerve to modulate cortical function more broadly. Cranial nerves are sensory and motor neurons that project from the brainstem and supply nerves to (i.e., innervate) the body, especially the head and neck. Stimulation of their peripheral branches leads to changes in the activity of neuromodulatory systems, which regulate nervous system activity via neurotransmitters, such as changes in attention with the release of norepinephrine (NE) throughout many areas of the cortex. Vagus nerve stimulation (VNS) is a type of PNS that has, until recently, required surgical implantation, limiting its use to clinical populations. However, recent innovations have led to user-friendly, non-invasive transcutaneous VNS (tVNS) technologies that stimulate the vagus by passing electrical current on the skin.

These technologies allow for a wider range of VNS applications with neurotypical populations, including the use of tVNS to support L2 learning.

2.2 Stimulation of the Vagus Nerve

2.2.1 Mechanisms of action

VNS has been investigated invasively (iVNS) in clinical populations since the mid 1980s for its efficacy as an antiepileptic and antidepressant (George & Aston-Jones, 2010; Vonck et al., 2014). Recently, its effects on auditory processing, memory, and cognition have also been studied (for a review, see Colzato & Beste, 2020). VNS involves electrical stimulation applied at low levels to branches of the vagus nerve located in the ear (auricular: inner ear, tragus, or cymba conchae) or the neck (cervical) that carry nerve impulses back to the brain. The vagus nerve is the tenth cranial nerve and originates from the medulla in the brainstem. Stimulation to the vagus nerve projects along nerve fibers to a brainstem nucleus called the nucleus tractus solitarii (NTS). The most well-studied mechanism underlying VNS benefits for memory and cognition (e.g., George & Aston-Jones, 2010; Vonck et al., 2014) involves the NTS's innervation of the locus coeruleus (LC) brainstem nucleus, though other mechanisms are also under investigation (e.g., Manta et al., 2009). The LC produces all of the neocortex's supply of the neurotransmitter norepinephrine (NE; Samuels & Szabadi, 2008), which plays a critical role in attention modulation (Aston-Jones & Cohen, 2005).

tVNS-related benefits may be due in part to the LC-NE system's role in optimizing behavior by controlling the trade-off between scanning and focused states of attention. Peak task

performance is associated with moderate tonic patterns of LC neuron firing (slow, baseline activity indicative of one's general arousal level), and high levels of phasic patterns of LC neuron firing (fast, task-evoked activity, indicative of one attention to a stimulus; Aston-Jones & Cohen, 2005). In addition, NE facilitates cortical long-term potentiation, a form of synaptic plasticity that may be the major cellular mechanism behind memory formation (Vonck et al., 2014). LC activity has been directly associated with left-ear taVNS in humans with neuroimaging (Frangos et al., 2015; Zhang et al., 2019).

Additional neurotransmitters, including gamma-aminobutyric acid (GABA), acetylcholine, dopamine, and serotonin have also been implicated in the effects of VNS (Van Leusden et al., 2015; Manta et al., 2009; Öztürk et al., 2020). For example, the production of GABA by VNS has been linked to iVNS's efficacy in treating epilepsy (Ben-Menachem et al., 1995; Walker et al., 1999), and fMRI studies have found evidence for the production of GABA from tVNS as well (Dietrich et al., 2008; Frangos et al., 2015; Kraus et al., 2013; Yakunina et al., 2017). Interestingly, unlike NE and acetylcholine, VNS's targeting of GABA appears to be lateralized, specifically contralateral: Capone et al. (2015), using a TMS paradigm that is sensitive to GABA activity, found a significant effect in right motor cortex after active tVNS in the tragus of the left ear. In regard to acetylcholine, this neurotransmitter only appears to be affected by transcutaneous cervical VNS (tcVNS), when tVNS is applied to a branch of the vagus nerve on the side of the neck (Colzato & Beste, 2020). While taVNS only stimulates afferent fibers of the vagus nerve, tcVNS stimulates both afferent and efferent fibers, acetylcholine only being implicated in the latter.

While the precise mechanism(s) of action of VNS are still under active investigation in both humans and animal models, distinguishing between these is not the aim of the present dissertation. Indeed, Aston-Jones and Cohen (2005) note that effects of neurotransmitters may be cascading given that that NE, dopamine, serotonin, and acetylcholine don't directly affect excitation or inhibition of postsynaptic neurons but instead modulate that activity through other neurotransmitters, like glutamate and GABA. What is evident thus far and relevant to the present dissertation is that VNS does appear to have a clear, positive effect on arousal and/or memory, regardless of additional neurotransmitter(s) involved beyond the LC-NE system.

2.2.2 Effects on learning and memory

The few studies that have investigated the effects of non-invasive tVNS on cognitive function in humans have shown improvements in learning and memory. In Jacobs et al. (2015), 30 older adults participated in a single-blind within-subjects study comparing active and sham taVNS conditions, which performed a face-names association memory task. In an encoding phase, participants saw 60 neutral faces with names for five seconds each and then rested for 10 minutes (consolidation phase). During a subsequent retrieval phase, participants saw old and new faces, decided if they had seen each before and, if so, selected the correct name. Active taVNS was applied to the auricular branch of the vagus nerve within the outer ear canal while sham taVNS was applied to the earlobe. Conditions were counterbalanced within participants across two sessions, and stimulation was delivered during both encoding and consolidation (17 minutes total). Accuracy significantly improved for the active over the sham condition, although no effects on reaction times (RTs) were observed. Jacobs et al. (2015) also presented data collected

from a standard neuropsychological test of episodic memory before and after the faces-names association task, measuring both immediate memorization and delayed recall for 15 monosyllabic words. Performance declined over time in the sham taVNS condition but was maintained for the active taVNS condition.

Also relevant to the present study, VNS has been observed to cause long-lasting changes in auditory processing, which may have implications for linguistic tone learning. VNS has been associated with plasticity in primary auditory cortex during pure tone learning (e.g., Borland et al., 2018; Kilgard, 2012), which has been shown in animal models to persist at least one day after treatment (Engineer et al., 2011).

There are multiple ways in which tVNS may be implemented, including priming (i.e., conditioning) and peristimulus stimulation. Priming involves applying stimulation for a specified number of seconds or minutes prior to starting a critical learning task, presumably inducing tonic shifts in arousal and thus cortical excitability that prepare the individual to be in an optimal state for learning throughout the task. For minutes to hours after even 30-second VNS pulse trains (or sequence of pulses), studies have observed an increase in the firing rates of neurons in the LC (Groves et al., 2005), activity in LC and related brain structures (Frangos et al., 2015), and concentrations of norepinephrine in the cortex and hippocampus (Follesa et al., 2007). Peristimulus stimulation involves delivering a pulse train of stimulation just prior to the presentation of critical stimuli, presumably inducing phasic changes in task-related attention to and consolidation of specific to-be-learned information. Work exploring peristimulus VNS has shown effectiveness at improving low-level auditory processing (Engineer et al., 2011; Kilgard,

2012). Because both tVNS approaches are hypothesized to impact LC function, we predict that either may be beneficial to language learning, and thus both are included in this study. Similar approaches have also been studied in the context of TMS and tDCS (Klooster et al., 2016).

The topic of language learning is still new to exploration with tVNS interventions, but the cited research provides preliminary evidence that tVNS could impact language learning and tone learning more specifically. The effects of tVNS on attention and memory consolidation could promote more effective language learning, and the results of Jacobs et al. (2015) suggest it could enhance retention rates. Assessing the benefits of any new intervention on language-learning outcomes is non-trivial. Does the intervention serve to increase phoneme and/or word recognition accuracy overall? Speed the overall learning rate? Reduce the mental load associated with learning an individual item, freeing up mental resources for other aspects of learning? Particularly challenging for tVNS is that, due to a paucity of established research, expected effect sizes have not been established, and thus it is possible that tVNS-induced changes in neural function might be subtle or very focused, and thus primarily observable for only a subset of possible language learning outcomes. Given the range of possibilities, this dissertation's assessment of tVNS-driven language-learning benefits includes multiple outcome measures at multiple timescales (i.e., trial-level accuracy, trial-level reaction time, and moment-by-moment deployment of mental resources as assessed with pupillometry).

2.3 Using Pupillometry to Assess the Impact of tVNS

While accuracy and reaction time measures have traditionally been used to assess language learning outcomes, individuals who achieve the same level of proficiency in terms of phoneme discrimination accuracy or vocabulary size may have exerted vastly different degrees of effort to achieve that level of performance. Thus, there has been increasing interest in objectively measuring the mental effort that individuals deploy throughout the course of learning, above and beyond measures of accuracy. Effort has been broadly defined in terms of the mental resources that are allocated to meet the demands of a task (Pichora-Fuller et al., 2016). According to models like the Framework for Understanding Effortful Listening (FUEL), the allocation of effort to a task is driven by not only the demands of a task (e.g., the difficulty of parsing an acoustic stimulus), but also individual differences in auditory and cognitive capacities, and motivation and arousal levels. Importantly, differences in cognitive effort have been observed even when performance is high or when variation in performance or reaction is otherwise matched or controlled for across individuals (e.g., Kuchinsky et al., 2013). This suggests that measuring effort in addition to performance metrics may be important for comprehensively assessing the challenges that learners face.

Pupil dilation, measured with an eyetracker, has been used as a physiological marker of changes in effort in a number of cognitive and sensory domains (e.g., Zekveld et al., 2018) in part because it has been associated with the well-studied LC-NE system that modulates states of attention (Aston-Jones & Cohen, 2005) and, as described above, is the primary mechanism through which tVNS is purported to operate. LC activity is thought to influence pupil size

through NE receptors in both the muscle that controls iris dilation and the Edinger-Westphal brainstem nucleus, which innervates the iris sphincter muscle (Loewenfeld, 1999). Baseline, tonic pupil diameter has been investigated as an index of general arousal while changes in phasic pupil diameter have been linked to stimulus-dependent changes in attention and effort (Gilzenrat et al., 2010). As described, the relationship between task demands and effort is not one-to-one, and indeed pupil dilation has been shown to track this nonlinear relationship: Low demand is typically associated with low effort and a smaller dilation response, moderate load is associated with high effort and a relatively larger response, and high cognitive load may result in fatigue or overexertion associated with less effort and a smaller pupil size (e.g., Ohlenforst et al., 2017). Because of this nonlinear relationship, the predicted impact that an intervention or training program may have on the pupil response depends on participants' performance level (Kuchinsky & Vaden, 2020). If the task is so difficult that people give up, training may serve to improve performance at the cost of increased effort. Training for tasks on which performance is moderately good to high may instead result in decreased effort and either better or maintained performance.

Though studied extensively in the domains of auditory (Zekveld et al., 2018) and cognitive processing (van der Wel & van Steenbergen, 2018), pupillometry is a metric newly applied to the field of SLA (see Schmidtke, 2018 for a review). However, it may be especially useful for providing insights into the mechanisms by which tVNS supports second language learning due to its reliable link to the LC-NE system (Eckstein et al., 2017). Linking pupillometry and tVNS during word learning can triangulate outcomes from behavioral training with tVNS to generate important insights into the mechanism connecting tVNS to second

language learning, providing both a more detailed picture of learning processes in real time and a validation of the attentional mechanisms purported to be enhanced with tVNS.

2.4 Lexical Tone Training as an Intervention Target

The present dissertation takes this neurostimulation approach to determine how taVNS may support native speakers of English naïve to tone languages as they learn novel words distinguished by Mandarin lexical tone contrasts. Lexical tone contrasts are notoriously challenging for speakers of non-tonal languages, like English, which does not distinguish between multiple meanings of a word based solely on pitch. Unlike in English, pitch in Mandarin Chinese is contrastive and changing the tone of a word changes its meaning: /ma/ with a high flat tone means 'mother', /ma/ with a rising tone means 'hemp', /ma/ with a dipping tone means 'horse', and /ma/ with a falling tone means 'scold'. Mandarin has five tones, a high flat tone (tone 1; T1), a mid-rising tone (tone 2; T2), a low dipping tone (tone 3; T3), a high-to-low falling tone (tone 4; T4), and a fifth neutral tone (Wong, 1953). Many words in Mandarin comprise minimal sets with two or more tones. This feature makes comprehension difficult for monosyllabic words and even more difficult for disyllabic words, which comprise a majority of words in Mandarin and where syllable position can further affect the tone contour and perceptual complexity. Thus, accurate perception and categorization of a word's tone is essential for achieving proficiency.

Even advanced second language learners of Mandarin show persistent difficulties in the learning and processing of lexical tone. In recent work, Pelzl et al. (2019b; 2021a; 2021b)

explored the problem space as not one of simple tone identification but one of combining tonal and lexical information into a phonolexical unit. For example, Pelzl et al. (2019b) found that strengthening the connection between a given word and its tone category is a much slower process than strengthening the tone category itself, as there are only a few contrastive tones versus thousands of tone words. Thus, central to these studies (Pelzl et al., 2019b; 2021a; 2021b) is instead the question of whether the primary cause of these learner deficits is a problem of encoding—incomplete knowledge of the relationship of tonal and lexical information resulting in lower quality lexical representations, a problem of retrieval—poorer access to those units during language processing, or some combination of both. Both the results in Pelzl et al. (2021a) and Pelzl et al. (2021b) reveal potential issues with both encoding and retrieval but not perception, and couch their findings in light of the fuzzy lexical representations hypothesis (Cook, 2012; Cook & Gor, 2015; Gor & Cook, 2020; Gor et al., 2021), which posits that L2 learners have lower resolution phonolexical representations (i.e., phonological forms mapped to words in the L2 mental lexicon) for less familiar words (Hayes-Harb & Masuda, 2008; Cook, 2012; Cook et al., 2016; Darcy et al., 2013). This hypothesis has become foundational to the recently developed Ontogenesis Model of the L2 Lexical Representation (OM; Bordag et al., 2021; forthcoming). The OM provides a framework for the learning of individual words, such that fuzziness in the lexical representation (falling short of optimum acquisition for a given word) can manifest at multiple levels, including the phonological level, orthographic level, semantic level, and the mapping between these three levels and in turn how these whole representations network within a larger lexicon. In the tone learning literature, the locus of interest is focused on fuzziness the phonological level, semantic level, and/or the phonolexical mapping between the two.

Pelzl et al. (2021a) notes that fuzzy representations for tone words may be unique in that, rather than retooling the native language's existing phonological space, learners are required to develop sensitivity to F_0 cues in an entirely new way, and thus metalinguistic tone knowledge may have an additional role in L2 tone word recognition. In an experiment, Pelzl et al. (2021b) conclude that tone words in learners' mental lexicons likely have fuzzy representations from evidence that as many as a staggering 25% of those word representations for even advanced L2 Mandarin learners have missing, incorrect, or uncertain tone information. This fuzziness not only has ramifications for their performance on those words, but a lack of sensitivity to these phonolexical cues is also likely to affect how efficiently they can learn additional tone words. In Pelzl et al. (2021a), native speakers of Mandarin were compared to native English advanced learners of Mandarin on behavior and event-related potentials (ERPs) on a Mandarin word-picture matching task, and learners were found to have clear difficulties in encoding tones into long-term memory. However, even when properly encoded, some deficits in retrieval during word recognition also occurred. Logically, if even advanced learners have fuzzy representations, encoding and retrieval of L2 tone words for native speakers of non-tonal languages are a persistent problem regardless of proficiency level.

2.4.1 Differential performance of tones

Processing differences for English NSs between individual tones have been observed repeatedly. In his review, Pelzl (2019a) notes that tone 2 has been consistently reported as the most difficult tone in isolated syllables for Mandarin learners across at least eight different studies. However, when tonal categories are pitted against each other, tone 1 and tone 4 may be

more confusable than tone 1 and tone 2 (e.g., Hao, 2012; Wang et al., 1999). Tone 1, the high flat tone in Mandarin, appears to be the easiest tone to learn among tones 1, 2, and 4 (e.g., Llanos et al., 2020; Lu et al., 2015; and see Pelzl, 2019a), however the story becomes more complex in the context of differentiating between specific tone pairings and/or with multisyllabic and sentential contexts. For example, So and Best (2010) found English NSs had an easier time with tone pairings that did not share similar features, such as T1-T2 and T1-T4 vs those that share similar phonetic features, like T1-T3 and T2-T4. Wang et al. (1999) additionally found that the T1-T4 pairing was the hardest pairing to learn in their study. However, in contrast, the AX task in Lu et al. (2015) looked at discrimination of tone pairs based on the order in which they were heard and produced results showing that the T2-T1 pairing (in that order) was the hardest to discriminate, followed by T2-T4 and then T4-T1 and T1-T4 equally.

Kaan et al. (2007) investigated English, Mandarin, and Thai NSs' processing of Thai tonal contours in an oddball task paradigm using a mid-level flat tone as standard and both low falling tone and high rising tone deviants. The authors concluded that different NS groups attend to different physical properties of tonal contours: critically, while Mandarin NSs were found to be more sensitive to the pitch contour over time, English NSs were more sensitive to early starting pitch differences in height (see also Maddox et al., 2013). In the context of English NSs learning Mandarin tone, this conclusion would generate a hypothesis competing with the above literature, that participants would more easily learn tone 2, which has the largest initial F0 distance from tones 1 and 4, versus tone 4, which is relatively more similar in initial F0 to tone 1. Yet it is worth noting that realistic Mandarin tones instantiate a larger start-to-end contour for tone 4, a high-to-low falling tone, than tone 2, a mid-to-high rising tone. This difference may in

turn lend more acoustic salience to the learning of tone 4 vs tone 2. While the literature is mixed on differing levels of difficulty tones in certain scenarios, what is clear is that there is an abundance of fuzzy phonological representations observed in the tone learning literature if non-native speakers of tone languages are confusing tones to a degree that native speakers do not. Indeed, in some respects the difficulty may reflect learner individual differences in how well they've resolved their new phonological tone representations more than some quality of the tones in and of themselves. Regardless, if taVNS assists with lexical tone learning, it would be prudent to investigate any differential effects of taVNS on specific tones, whether it may more specifically enhance perceptual learning for easier- and/or harder-to-learn tones.

2.4.2 Individual difference predictors of tone learning

Importantly, there is noticeable individual variability in learning trajectory for lexical tone. For example, in Chandrasekaran et al. (2010), many learners (out of a total of 16) did approach ceiling on a lexical tone word learning task ($k = 24$) after nine days of tone training, n = 8 at or above 80% accuracy, but the other half of the learners performed below 80% accuracy, n = 5 of those below 50%. In light of the OM, these results could be interpreted as only half of participants achieving their optima for tone word learning in this study while the other half plateaued, resulting in fossilized, persistently fuzzy lexical representations for the newly learned words. Because of this wide variability in ultimate outcomes, many studies have taken an individual differences approach, investigating factors such as music experience, non-linguistic tone aptitude, and executive function (e.g., Bowles et al., 2016), or even exploring individual differences in attended acoustic cues (e.g., Chandrasekaran et al., 2010). While this tactic does

help explain the origins of this variability in ultimate performance in multisession training studies, it does not directly speak to what types of interventions can help those with lower aptitude levels in those areas, such as tone aptitude, overcome their obstacles in acquiring tone as a lexical feature. Therefore, additional types of lexical tone learning interventions should be explored. Ingvalson et al. (2011) note in their review that understanding learner individual differences will create more reliable predictions of performance to be, in turn, used to tailor training to provide the greatest benefit to learners. The authors specifically call for the assignment of learners to training paradigms on the basis of their pre-training performance so that the impact of different types of training can be assessed.

While short-term interventions such as sound-perception training have helped native English speakers perceive lexical tone differences, naïve trainees still regularly fall short of consistently and accurately categorizing lexical tones, even taking as many as 18 training sessions to reach consistent performance (e.g., Bowles et al., 2016; Chandrasekaran et al., 2010; Li & DeKeyser, 2017; Wong & Perrachione, 2007). Native speakers of English who have attained high levels of proficiency in Mandarin often continue to perform below native Mandarin speakers on some measures of tone discrimination (e.g., Pelzl et al., 2019b). Perrachione et al. (2011) is one study to examine variability among individual learning trajectories, and they found that learners with poorer performance on a pre-training non-linguistic tone perception measure made the greatest performance gains with low variability perception training while those with better performance on the pre-training task saw the greatest gains from high variability perception training. Another avenue to make language learning more accessible is

neurostimulation, which would not necessarily require modifying the training or input learners receive, and could potentially help give a boost to those with lower aptitude.

While many other factors predict language learning generally, non-linguistic tone aptitude and musicality are uniquely positioned for an ATI study with tVNS on Mandarin tone learning given their well-documented relationship to linguistic tone learning and processing. Non-linguistic tone aptitude is a general cognitive ability that relates to an individual's ability to perceive changes in tone contour. One popular measure of non-linguistic tone aptitude is the pitch contour identification task (PCID, see 3.2.3.2; Bent et al., 2006), in which participants have a 3-alternative forced choice task of responding to whether a given pure tone is flat, rising, or falling. In this task, the flat tone varies by pitch height (200-350 Hz) and the rising and falling tones vary by steepness of the contour (5-50 Hz difference from start to finish). This ability to distinguish differences in tone contour (rather than strict differences in tone height) has been found to relate to both more accurate lexical tone learning outcomes and faster lexical tone learning for native English speakers in a number of studies (e.g., Bowles et al., 2016; Li & DeKeyser, 2017; Perrachione et al., 2011; Wong & Perrachione, 2007).

Within musicality, two main variables are worth noting: music aptitude and music experience. Music aptitude is a general cognitive ability whereas music experience is a direct result of musical training. Importantly, these are related, but separable constructs that have both not only been shown to relate to better language learning outcomes generally (e.g., Dittinger et al., 2017; Slevc & Miyake, 2006), but also relate specifically to lexical tone identification and learning. For example, regarding music aptitude, Delogu et al. (2006; 2010) found that higher music aptitude scores were associated with a greater ability to discriminate tone. Cooper and

Wang (2012) found music aptitude to predict success in Cantonese tonal word learning for native speakers of English. One of the most well-known indices of music aptitude is the Wing Music Aptitude Test (Wing, 1968). In Li & DeKeyser (2017), they found that measures from all three subtests of the Wing loaded well on only one musical ability component. They found in their Mandarin tone word training study that musical ability strongly predicted error rate for tone perception, but did not predict reaction time for perception or either error or reaction time for tone production.

As for music experience, musicians have been repeatedly found to outperform non-musicians in pitch identification (e.g., Bidelman et al., 2013; Gottfried, 2007; Gottfried & Ouyang, 2005; 2006; Gottfried et al., 2004; Lee & Hung, 2008; Wayland et al., 2010) and tone word learning (e.g., Cooper & Wang, 2012; Dittinger et al., 2016; Wong & Perrachione, 2007). Similar to the music aptitude results for Li & DeKeyser (2017) that are typical of music aptitude outcomes, musicians are also generally found to make significantly fewer errors than non-musicians with no reliable difference in RTs.

Interestingly, a few tone word learning studies have included measures of both tone aptitude and musicality (e.g., Bowles et al., 2016; Cooper & Wang, 2012), and have found tone aptitude to play a larger role than musicality in predicting learning outcomes. Additionally, results from Bowles et al. (2016) and Wong & Perrachione (2007) show support for an early role for musicality in learning lexical tone contrasts while tone aptitude appears to show a larger role, both predicting early learning and predicting performance when learners are given generalization stimuli with new, untrained talkers. These individual differences are well-established, strong predictors of lexical tone learning and performance, and the most supported effect of taVNS is an

increase in the production of a neurotransmitter closely tied to attention, encoding, and learning. Thus, those with less aptitude and experience may benefit from a boost from taVNS. In the long run, a combination of insights gained from neurostimulation and aptitude-by-treatment interaction research could maximize the potential of both. Research in neurostimulation is still nascent, but selective, purposeful neurostimulation could lead to efficient and practical gains in learning.

Based on the above literature, this dissertation aims to (a) determine whether taVNS may be a practical intervention for tone word learning outcomes, (b) examine whether benefits of taVNS also arise at the phonological level for easier and/or harder tones, and (c) investigate the potential for the effects of taVNS to interact with individual differences known to affect tone learning outcomes. It will do so with the following research questions, examining results not only at the behavioral level, but also the physiological level:

1) Does active (priming and/or peristim) taVNS versus sham taVNS improve behavioral learning outcomes for Mandarin lexical tone?
2) Does active (priming and/or peristim) taVNS produce a differential deployment of cognitive effort versus sham taVNS during lexical tone learning and support the role of the LC-NE as a mechanism connecting taVNS to learning?
3) Does active (priming and/or peristim) taVNS versus sham taVNS improve behavioral learning outcomes for Mandarin phonological tone?
    - Are easy and hard tones differentially impacted?

- Do any effects generalize to untrained speakers and/or untrained segmental contrasts?

4) Does active (priming and/or peristim) taVNS versus sham taVNS produce a differential deployment of cognitive effort during phonological tone processing at test?
    - Are easy and hard tones differentially impacted?
    - Do any effects generalize to untrained speakers and/or untrained segmental contrasts?
5) Are effects of (priming and/or peristim) taVNS-facilitated learning moderated by individual differences previously known to affect Mandarin tone learning?
    - Non-linguistic tone aptitude
    - Musicality
6) Do individual differences moderate effects of active (priming and/or peristim) taVNS versus sham taVNS in eliciting a differential deployment of cognitive effort during phonological tone processing and/or lexical tone processing?
    - Non-linguistic tone aptitude
    - Musicality

Chapter 3: The Corpus

3.1 Introduction

The body of data used for this dissertation comes from an ongoing set of multi-year efforts to investigate the potential for tVNS to facilitate language learning across a multi-institution DARPA Targeted Neuroplasticity Training grant. The suite of research under the grant includes investigations into grammar learning, vocabulary learning, and lexical tone leaning with both humans and animal models. Relevant for the present dissertation is the research agenda for lexical tone learning in humans, and the aspects of that design and data collection as pertain to this dissertation are described below.

3.2 Methods

3.2.1 Participants

Eighty-three participants completed this study. Participants were recruited from the University of Maryland and surrounding community, provided informed consent prior to enrolling in this study, and were paid $20/hour overall for their time ($10 for 1.25-hour session 1, $10 for the 3-hour session 2, $125 for the 3-hour session 3). This study was approved by the University of Maryland's Institutional Review Board and the U.S. Department of Navy Human Research Protection Program (DoN HRPP).

Participants were aged 18-35 native speakers of American English recruited from the University of Maryland and surrounding community who reported no prior exposure to any tonal

language and no significant exposure to another language before age 12[1]. All participants had self-reported normal or corrected-to-normal vision and unimpaired use of their right (dominant) hand, no hearing impairments, learning disabilities, history of neurological or psychiatric disorders, or ocular disorders that would affect eyetracking, and had not taken any psychoactive medications within two months of testing. Participants reported having none of the following conditions: being pregnant or nursing, history of cardiac or vascular disease, diabetes, epilepsy, fainting, head or face injuries, pain, or pain disorders, recent hospitalizations, or implanted electronic or metallic devices including non-removable facial piercings.

Participants were pseudorandomly assigned to a taVNS group in order to balance groups on two variables known to influence lexical tone training outcomes: musicianship and non-linguistic pitch discrimination ability (e.g., Bowles et al., 2016; Chandrasekaran et al., 2010; Dittinger, et al., 2016; Wong & Perrachione, 2007). The tasks used to measure non-linguistic tone aptitude (the Pitch Contour Identification Task (PCID)) and musicianship (from an item in the Ollen Musical Sophistication Index) are described in Materials and procedures (3.2.3).

3.2.2 Design

This study comprised two lexical tone-training sessions that occurred on consecutive days (n = 4 completed two days apart, 1 priming and 3 sham participants) and a pre-training

[1] Participants who reported experience learning a language before age 12 were admitted into the study if the experience before age 12 was limited to class in a non-immersion school.

session that occurred before the first training session. The pre-training session lasted 1.25 hours and each training session lasted 3 hours. During the pre-training session, participants completed computerized tasks measuring several aspects of cognitive and musical ability, demographics, and language history. These tasks were administered to confirm participant eligibility in the study, collect measures used for group balancing, and use as other covariates outside of the scope of the present analysis.

Both training sessions included the same training tasks and tests, and included the collection of behavioral data (accuracy and reaction times), pupillometry, and electroencephalography (EEG). The EEG methods and results are beyond the scope of this dissertation and are omitted here. All tasks and tests were administered via E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA) with a 24" LCD monitor positioned 65 cm from the participant's chinrest and all sounds were presented at 70 dB SPL (decibels, sound pressure level) through a set of first-generation Neuvana earbuds (Neuvana, Boca Raton, FL) with embedded electrodes that deliver taVNS. Pupillary data were collected with an EyeLink 1000 Plus eyetracker (SR Research, Ltd., Ontario, Canada) positioned below the monitor and behavioral responses were collected with a Chronos button box.

At the start of both training days, participants completed a sound check followed by a self-paced introduction to the concept of lexical tone that gave examples of naturally produced monosyllables featuring Mandarin tones 1, 2, and 4 along with visual depictions of the tone contours. Participants then inserted earbuds modified by placing Hydrogel (Axelgaard Manufacturing Co., Ltd, Fallbrook, CA) over the electrodes on the left earbud to create a stable

conductive bridge to the skin of the outer ear canal. After another sound test, participants completed the sequence of tone-word training tasks and tests in the following order: a passive paired-associates word learning task, a match/mismatch lexical recognition test, and a learned-word lexical recall test.[2] The ten-minute priming task was administered three times each training day, once before every 20 minutes of task or test time. This involved watching a ten-minute silent animated video, *Inscapes*, which is designed to keep participants awake, engaged, and still during extended resting scans for MRI research (Vanderwal et al., 2015). During the video, participants in the active priming taVNS group received continuous taVNS at 0.2 mA below an individualized, sub-perceptual threshold (determined via a calibration procedure describe below) for the full ten minutes while the sham priming taVNS participants received no stimulation excepting a short 7 s ramp-up as part of blinding procedures. The *Inscapes* video was not administered to the peristim taVNS participants; instead, the peristim participants in the active taVNS group received 500-ms bursts of taVNS preceding each trial during all training tasks and tests except the learned word recall test due to the nature of the test. For both active and sham peristim taVNS participants, this 500 ms is in addition to the task timings described in Section 3.2.3 and—due to an audible noise artifact occasionally caused by taVNS at higher stimulation levels—was overlaid with both 60 dB SPL pink noise and a recording of the taVNS sound artifact in order to mask any actual artifact produced by the intervention. Before each task that

[2] Participants in all three conditions also completed several tasks and tests that address research questions beyond those considered in this paper, for example affect and anxiety surveys.

involved taVNS, all participants completed taVNS calibration and ramping, described below, and taVNS was administered during both sessions.

This study implemented a double-blind design—participants and session proctors were unaware of taVNS group assignments. A member of the research team not involved in data collection or analysis determined a participant's group assignment based on their non-linguistic tone aptitude and self-rated musicianship scores (collected during the pre-training session) and assigned a new number to the participant for use during training and testing. The computerized training tasks and tests were programmed to reference a pre-loaded taVNS-group list so that entering the participant number for an experimental task triggered the correct taVNS delivery, thus allowing proctors to administer the tasks and tests without knowledge of taVNS group assignment. Before each task or test involving tVNS, all participants calibrated taVNS intensity (described below), providing their perceptual threshold.

The task and procedure order are shown in Table 1 and relevant portions described in the next section (3.2.3). The training and testing tasks in the corpus were largely designed after Dittinger et al. (2016), which analyzed behavior and ERPs of musicians and nonmusicians in relation to Thai word learning, including lexical tone (Thai 0 vs. 1) and other lexically contrastive features: voicing (/b/ vs. /p/), aspiration (/p/ vs. /p^h/), and vowel length (/a/ vs. /a:/). The Dittinger et al. (2016) design allowed for examination of word learning, word recognition, and phonological categorization of their nine trained words across a group manipulation in which the groups were hypothesized to perform differentially. Likewise, the study in this corpus examined the training of nine words that varied minimally across tone, but also voicing and

vowel: /ba1/, /ba2/, /ba4/, /bi1/, /bi2/, /bi4/, /pi1/, /pi2/, and /pi4/. In this way, the tone learning results can be more generalizable across these additionally contrastive features, albeit in just a small to-be-learned vocabulary. The entire design below was piloted to ensure that the two-day training resulted in neither floor nor ceiling performance consistently across pilot subjects.

Table 1. Task and procedure order for the training sessions. Bolded items are analyzed in the present dissertation.

Training Session 1	**Training Session 2**
Tone Introduction	EEG Application & Impedance Check
Phonological Tone Categorization - Pretest [P]	Pre-taVNS Mood Questionnaires
Dynamic Pupil Range [P]	Tone Introduction
EEG Application & Impedance Check	Video & taVNS Priming [S-Priming]
Pre-taVNS Mood Questionnaires	Phon. Tone Categorization – Training [S-Peristim]
Video & taVNS Priming [S-Priming]	**Word Learning – Passive** [E, **P,** S-Peristim]
Oddball Task - Pretest [E, P, S-Peristim]	Word Learning – Active [S-Peristim]
Video & taVNS Priming [S-Priming]	Video & taVNS Priming [S-Priming]
Phon. Tone Categorization – Training [S-Peristim]	**Lexical Recognition Test** [E, **P,** S-Peristim]
Word Learning – Passive [E, **P,** S-Peristim]	Video & taVNS Priming [S-Priming]
Word Learning – Active [S-Peristim]	Oddball Task – Posttest [E, P, S-Peristim]
Video & taVNS Priming [S-Priming]	**Phonological Tone Categorization – Posttest** [P]
Lexical Recognition Test [E, **P,** S-Peristim]	Dynamic Pupil Range [P]
Post-taVNS Mood/Comfort Questionnaires	Post-taVNS Mood/Comfort Questionnaires
Lexical Recall	**Lexical Recall**
	Post-Experiment Questionnaire

[E] EEG data collected, [P] pupillometry data collected, [S-Priming] taVNS for the priming group only (sham also saw the video, but without taVNS). [S-Peristim] taVNS for the peristim group only (sham also had extra time per trial, but without taVNS).

taVNS involved calibration and ramping (for active and sham of both kinds) and stimulation during the task (for active only).

3.2.3 Materials and procedures

3.2.3.1 taVNS calibration

taVNS originated from a Digitimer DS8R Constant Current Stimulator (Digitimer North America, LLC, Fort Lauderdale, FL), which was set to deliver square waves with a 50 μs pulse and 350 μs interphase dwell with alternating polarity and a 100% recovery phase ratio. All participants (active and sham) completed the same calibration procedure, which consisted of administering 2000 ms taVNS pulses at random 1000–3000 ms intervals that increased from 2 to 10 mA in 0.5 mA steps until participants indicated they could feel the stimulation by pressing a button. taVNS intensity was then reduced by 1.0 mA (or to 2.0 mA, if the level was below this threshold) and then slowly ramped up in 0.1 mA steps until participants again pressed a button to indicate they felt the stimulation. At the start of the following task or test, all participants received a brief sequence of taVNS pulses that ramped up from 2.0 mA to perceptual threshold, while only the active taVNS groups received taVNS during the task or test at 0.2 mA below their perceptual threshold.

3.2.3.2 Non-linguistic tone aptitude

Non-linguistic pitch discrimination ability was measured with an abbreviated version of the pitch contour identification task (PCID) task used in Bowles et al. (2016; originally from Bent et al., 2006). Participants were presented with a pure tone and identified the tone as flat, rising, or falling by pressing a button. Stimuli varied by initial pitch height (200-350 Hz) for the flat tone and pitch contour difference (5-50 Hz) for rising and falling tones. There were 12 practice trials. The version used in this study was shortened from the original in that, instead of eight repeat trials for every unique item (i.e., each of 42 unique F0 start/end combination

resulting in 336 total trials), there were only three repeat trials for every unique item (126 total trials, 42 items each for flat, rising, and falling tones). Overall accuracy on this task was used for group balancing purposes.

3.2.3.3 Musicianship

Self-rated musicianship was assessed from a questionnaire that consists of items pertaining to an individual's experience playing an instrument and listening to music (Ollen, 2006). Self-rated musicianship was measured with the question: *Which title best describes you?* Possible responses were: 1 = non-musician, 2 = music-loving non-musician, 3 = amateur musician, 4 = serious amateur musician, 5 = semiprofessional musician, and 6 = professional musician.

3.2.3.4 Music aptitude

Part 3 of the Wing Music Aptitude Test (Wing, 1968) was administered. This portion indexes memory for pitch in a melody, and was included in this study as a measure of perceptive tonal short-term memory. Participants listened to a short melodic piano phrase twice and indicated which note (if any) changed in the second instance. Phrases increased in length (3 to 10 notes) as the test progressed. Participants had 4.1 seconds to respond. After three practice trials, 30 items were presented. This task produced a single measure of accuracy per participant.

3.2.3.5 Tone Introduction

Stimuli and Procedure. At the start of training days 1 and 2, lexical tone training began with participants introduced to the three Mandarin tones that are the focus of this study: tone 1 (high flat), tone 2 (rising), tone 4 (falling). Tone 3 was not included in this study due to a creaky quality in the speech tokens that is common for this tone and which leads to easier discrimination from the other tones. Additionally, depending on task context, tone 3 can also be the most confusable tone for both non-native speakers and native Mandarin speakers (Kirkham et al., 2011). The purpose of this task was to familiarize participants with the tone differences that are the focus of the subsequent training tasks and outcome measures and no data were collected for this task. During this task, participants were shown descriptions of each tone, which were drafted by a second language researcher and speaker of Mandarin. Following the description of each tone, participants were shown visual representations of each tone (the images in Figure 1) and listened to recorded examples of each tone with vowel /a/, which were produced by a male speaker of Mandarin. These recordings were used in a previous pitch contour perception task (Wong & Perrachione, 2007).

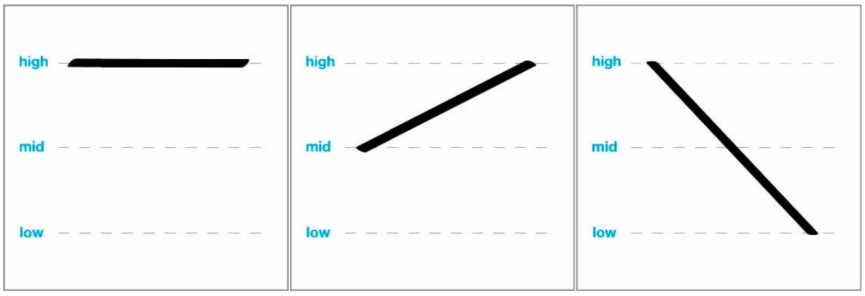


Figure 1. The visual aid shown to learners in the tone introduction for the high flat (tone 1), rising (tone 2), and falling (tone 4) tones.

3.2.3.6 Word learning tasks

3.2.3.6.1 Passive paired-associates word learning task

Stimuli. Recordings of /ba/, /bi/, and /pi/ spoken by a male and female native Mandarin speaker with tones 1 (high flat), 2 (mid rising), and 4 (high-to-low falling) spoken in carrier sentences were taken with permission from a previous tone-learning study (Bowles et al., 2016). This training can be considered low variability due to the low number of speakers and tokens per speaker. These Mandarin syllables were paired with nine English words: TRAY, OVEN, VASE, GOWN, RAFT, SOFA, MENU, LENS, and COIN. The words were all four letters long in order to control for screen luminance. Word frequency (logSUBTLEX: 2.29-2.71; Brysbaert & New, 2009) and concreteness (4.61-5.00; Brysbaert et al., 2014) were controlled. Three counterbalances were used to minimize any potential idiosyncrasies of learning a particular English word with a particular Mandarin syllable or tone. As a result, across lists, each English word appeared once with each tone, and once with each segment. Each participant encountered only one list, which was the same for training session 1 and training session 2. These nine words were each spoken by one male and one female native speaker of Mandarin, and were recorded by the same speakers, and at the same time, as stimuli in a previous study (Bowles et al., 2016). After recording, the audio files were root-mean squared (RMS) normalized to a consistent sound level (70 dB SPL). Given the above, this study design is considered a low-variability training study of lexical tone.

Procedure. Participants were instructed to learn the meanings of nine foreign language words, which would vary in sound (consonant, vowel) and tone. The importance of trying to memorize the words was stressed as they would be tested later. Every trial had a 750 ms baseline

period in which an English word was presented in the middle of the screen with the visual contour of its tone above (a flat, rising, or falling line). Then there was a 1750 ms period in which a Mandarin syllable was presented auditorily as the written English word and contour remained on the screen. Participants were not required to respond to the stimuli, but pupillometry was collected during this task. Each English word was presented a total of ten times (five times each per male and female speaker) for a total of 90 trials. Stimulus lists were pseudorandomized to avoid blocking by tone, segment type, or speaker.

3.2.3.6.2 Active word learning task

Stimuli. For each participant, the active word learning task employed the same counterbalance of the nine visual English word stimuli and their accompanying Mandarin pairings used in the passive word learning task. Each Mandarin pseudoword was presented a total of four times for a total of 36 trials. Stimulus lists were pseudo-randomized across tone, segment type, and speaker and each participant encountered only one list, which was the same across training days.

Procedure. At the beginning of this training task, participants read a set of instructions that was presented on screen. Like the passive word learning task, there was no practice block for this task. The instructions indicated that the goal of this task was for participants to choose the correct written English translation of the Mandarin word that was aurally presented during each trial by pressing one of two buttons. The instructions described the sequence of events in each trial: (1) a 500 ms baseline period, in which two English words were presented in the middle of the screen above the number for their corresponding button (1 or 2); (2) a 2,500–4,250 ms period

that began with the aural presentation of the Mandarin pseudoword while the English words and button numbers remained on screen and ended once participants pressed a button to indicate their response; (3) a 1,500 ms feedback period in which a box appeared on screen around the correct English translation and a word (“YES!”, “NOPE”, or “SLOW”) appeared in the middle of the screen above the English words, indicating participant performance, before the Mandarin word was presented a second time while an image of the corresponding tone contour appeared above the correct English translation. The box that appeared around the correct English translation was blue if the participant pressed the correct button and it was red if the participant pressed the incorrect button or did not respond within the allotted time.

For each Mandarin pseudoword, each of its four presentations in this task appeared with a different distractor word, which was drawn from the same set of nine English translations. The English distractors were chosen based on the relationship between their correct Mandarin pseudoword translation and the Mandarin pseudoword that was presented during the trial. Distractor words fell into three categories based on the combination of initial consonant segment and tone: same tone different segment, same segment different tone, and different segment and tone. Across the four presentations of each Mandarin pseudoword, at least one trial appeared with each distractor type. Across all trials, each tonal confusion pair (e.g., tone 1 is correct, tone 2 is distractor) and segmental confusion pair (e.g., /ba/ is correct, /bi/ is distractor) was presented 3–5 times and there was an equal number of correct answers that corresponded to each button (1 vs. 2).

3.2.3.7 Phonological tone categorization

3.2.3.7.1 Phonological tone categorization training task

Stimuli. The purpose of this task was to train participants in discriminating between the three Mandarin tones introduced in the Tone Introduction task as they occur in the nine phonologically plausible Mandarin pseudowords described above. From these nine pseudowords, a list of 126 pseudorandomized items was created, which included seven of each combination of segment, tone, and speaker (e.g., /ba1/, female speaker).

Procedure. At the beginning of this task, participants read a set of instructions that was presented on screen and completed a practice block with feedback that was identical in structure to the rest of the task but included only the sounds used in the tone introduction task as stimuli. The instructions indicated that the goal of this task was for participants to listen to the list of Mandarin pseudowords and to press one of three buttons to indicate the tone of each pseudoword as quickly and accurately as they could. The instructions also described the sequence of events for each trial, which consisted of: (1) a 500 ms baseline period during which a four-character mask ("####") appeared in the middle of the screen above the three tone images (high on 1, rising on 2, falling on 3) and the numbers 1–3 indicating the corresponding buttons; (2) a 1,750 ms period when the tone word was aurally presented and participant responses were collected while the mask and tone images remained on screen; and (3) a 1,500 ms feedback period consisting of a word that replaced the mask, indicating the participant's performance on the trial ("YES!", "NOPE", or "SLOW"), as well as a box that appeared around the correct tone image and button number.

3.2.3.7.2 Phonological tone categorization test

Stimuli. For each participant, this task employed the same nine Mandarin pseudowords used in the phonological categorization training task as well as nine novel pseudowords that did not appear in any of the training tasks and were included in this task to examine the generalizability of taVNS-facilitated learning. The generalization items were similar to those that appeared in the training tasks but contained novel combinations of consonant-vowel sequences and speakers: two consisted of new consonant-vowel sequences /fa/ and /ti/ spoken by the same male voice they heard producing the nine pseudowords used in training; two consisted of the pseudowords /ba/ and /pi/ used in the training tasks spoken by one new male speaker; and two consisted of the new consonant-vowel sequences /fa/ and /ti/ produced by the new speaker. From these items, a list of 108 trials was created, which included 54 tokens of the trained pseudowords (three tokens of each item produced by the original male speaker and three tokens of each item produced by the original female speaker) and 54 tokens of the generalization pseudowords (three tokens of each generalization item type and tone). The list was pseudorandomized so that the same tokens did not appear in consecutive trials.

Procedure. This task was administered as a pretest, at the beginning of the first training session, and a post-test, at the end of the second training session. At the beginning of this task, participants read a set of instructions that was identical to that used in the phonological categorization training task and then completed two practice blocks: one with feedback, identical to the practice block in the phonological categorization training task, and one without feedback. The instructions also described the sequence of events for each trial, which was identical to that of the phonological categorization training task but without the feedback period. Accuracy and

RT were recorded for each trial, starting at the onset of the aural presentation of the tone word, and pupillometry data were collected over the duration of each trial.

3.2.3.8 Lexical recognition test

Stimuli. In this test, each Mandarin pseudoword was presented 24 times, split by speaker, for a total of 216 trials. These trials were split into two 108-trial blocks with one break in between. Trials were pseudorandomized within block to avoid repetition of the same Mandarin pseudoword in consecutive trials. Half the trials were matches and half the trials were mismatches via tone only, not segment. No feedback was given.

Procedure. There was a 750 ms baseline period in which a visual English word appeared in the center of the screen. The tonal contour never appeared with the word in this test, unlike in the passive paired-associates word learning task. A subsequent 2,000 ms period began with an aurally presented Mandarin syllable while the English word remained on the screen. During this time, participants indicated whether or not the pairing was a correct translation by pressing a button. Finally, there was a 1,000 ms period in which a four-character visual mask of 'XXXX' replaced the written word on the screen.

3.2.3.9 Lexical recall test

Stimuli and Procedure. This test consisted of 9 trials. For each trial, participants were presented one of the nine Mandarin pseudowords produced by the female speaker and were given unlimited time to listen to each item as many times as they liked. Participants were instructed to type the correct English translation of the Mandarin syllable on a keyboard. There

was no word bank. Responses to this test were reviewed and the only hedge cases in determining accuracy were a limited number of instances where the participant had pluralized the English word (e.g., typed 'COINS' instead of 'COIN'). These responses were accepted as correct. There were no other synonyms or misspellings.

3.2.3.10 Post-experiment awareness questionnaire

This questionnaire gathered information about participants' awareness of their stimulation condition. Participants were asked to indicate to which condition they believed they were assigned when they received stimulation (answer options described priming, peristimulus, sham, sham without ramping, and other), their confidence in their answer, and whether the stimulation helped them perform better on study tasks (rating from 1-9).

3.2.4 Group balancing and double-blinding procedures

taVNS-group means for the ID measures collected during the pre-training session were compared using two-tailed *t*-tests. These results indicate that group balancing procedures were successful in balancing active and sham tVNS groups on PCID and self-rated musicianship (*p*s > .10). Participant responses to post-experiment questionnaire items probing their awareness of their assigned taVNS group were analyzed and the results indicate that the taVNS calibration procedures were successful in blinding participants to their taVNS group (*p*s > .10).

During this pre-training session, non-linguistic tone aptitude via the PCID task (Bent et al., 2006) and self-rated musicianship (Ollen Musical Sophistical Index; Ollen, 2006) were

collected, two variables that have been shown to be predictive of linguistic tone learning in previous research (e.g., Bowles et al., 2016; Chandresekaran et al., 2010; Dittinger, et al., 2016; 2017; Wong & Perrachione, 2007). After this pre-training session, participant data was sent to a researcher for group balancing and blinding; this researcher was not involved with proctoring the study or analyzing the data. This researcher took the participant scores on these two measures and put the participant in either the active or sham (no stimulation) taVNS group in order to keep the scores between the two groups as balanced as possible. The researcher then gave the proctors a new number for the participant to be used in the following sessions, which, when entered into E-Prime, triggered a preloaded group assignment list so that the participants received active stimulation or not without the participant or proctors being aware of which condition. In this way, along with the procedure described below to deliver taVNS below a participants' perceptual threshold in the active taVNS condition, the study was double-blinded. These procedures were successful in balancing active and sham groups on PCID and self-rated musicianship ($ps > .05$) and in blinding participants to condition ($ps > .05$ on a post-study survey).

As an additional step, the research team "triple-blinded" the data by sending it back to the outside team member to assign yet another participant number before data could be analyzed and group assignment revealed. In this way, proctors that also analyzed data are still prevented from knowing which participant they ran was in which taVNS group.

Chapter 4: taVNS-facilitated Lexical Tone Learning[3]

4.1 Introduction and Motivation

No previous research has directly investigated the potential impact of tVNS (priming or peristim) on language learning. Both priming and peristim tVNS deliveries may be impactful for language learning for different reasons: priming tVNS may put a learner in the optimal attentional state for learning, while peristim tVNS may assist with more robust encoding of individual to-be-learned stimuli.

As outlined above, Mandarin lexical tone is a well-studied but persistently difficult feature for English NSs. Therefore, it is a productive first use-case for the potential practical application of this non-invasive neuromodulatory intervention. This Chapter outlines analyses to examine some practical outcomes of *ab initio* language learning to gauge the practical utility of this type of intervention: *i.e.*, is it worth further investigation? Recognition and recall of newly learned vocabulary with lexical tone is examined across two days of learning, as well as pupillometric analyses to examine the potential mechanistic differences between the two modes of taVNS delivery during passive paired-associates learning.

Especially relevant to the pupil analyses of passive word learning in the present Chapter, changes in phasic pupil dilation (the task-evoked pupillary response; TEPR) has been linked to

---

[3] This chapter is largely based on the following peer-reviewed publication, with minor modifications: Pandža, N. B., Phillips, I., Karuzis, V. P., O'Rourke, P., & Kuchinsky, S. E. (2020). Neurostimulation and pupillometry: New directions for learning and research in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 56-77. https://doi.org/10.1017/S0267190520000069

an event-related potential (ERP) component that has been well-studied in the field of SLA: the N400. The N400 has been used to track lexical learning and has been shown in passive word-learning tasks (in which objective performance cannot be measured) to index the formation of semantic representations (e.g., Dittinger et al., 2016). Kuipers and Thierry (2011) found smaller N400 amplitudes to be associated with larger pupil dilation (more phasic LC-NE activity) in a passive picture-word semantic association task indicating that less effort (less phasic LC-NE activity/smaller pupil diameter) was associated with better integration of the word in the lexicon (larger N400 amplitude/more negative deflection); likewise more effort (larger pupil dilation) was exerted on unfamiliar words (with weaker lexical representations/smaller N400 amplitudes). This was observed despite there being no behavioral response required of participants. Thus, in a lexical tone learning study with both taVNS and pupillometry, one could expect smaller pupil dilation to reflect a more robust learning of new words. Importantly, pupillometry may allow us to observe the effect of taVNS as stimulus perception and lexical integration processes unfold, even in the absence of differences in traditional performance metrics of word learning.

4.1.1 Research questions

The present study uses taVNS and investigates its effects on lexical tone learning across multiple outcome measures that are sensitive to changes at varying timescales (across sessions, across trials, and across milliseconds). Priming taVNS, in which taVNS is applied for a continuous period preceding some learning or performance task, and peristimulus (peristim) taVNS, in which short bursts of taVNS are time-locked to individual stimuli in some learning or performance task, were utilized and contrasted with sham taVNS in a double-blind study with

tone word-learning tasks and lexical recognition (3.2.3.8) and recall (3.2.3.9) tests. Behavioral outcomes on recognition and recall tests were analyzed as indices of learning. Pupillometry was collected during a passive word learning task (3.2.3.6.1) and was analyzed as an index of cognitive effort.

The research questions of the present chapter are:

1) Does active (priming and/or peristim) taVNS versus sham taVNS improve behavioral learning outcomes for Mandarin lexical tone?
2) Does active (priming and/or peristim) taVNS produce a differential deployment of cognitive effort versus sham taVNS during lexical tone learning and support the role of the LC-NE as a mechanism connecting taVNS to learning?

The results presented here are from a larger endeavor (described in Chapter 3) showing positive effects of priming and peristim taVNS interventions compared to a sham stimulation condition in a double-blind study of lexical tone learning. These analyses represent an initial foray into studying the impact of taVNS on language learning, with a two-day training paradigm for native English speakers naïve to tone languages tasked with learning words with lexical tone. We conclude by discussing how the observed taVNS-related improvements here show promise for further neurostimulation research in the field of second language acquisition.

4.2 Results

4.2.1 taVNS improves tonal language learning performance

A total of 69 participants (46 female) ages 18–34 years (M = 21.56, SD = 3.16) were analyzed, after 14 were excluded for missing data, noncompliance throughout learning tasks, and/or incorrectly interpreting task instructions for at least one of the tasks. There were 17 participants in the priming taVNS group, 17 participants in the peristim taVNS group, and 35 participants in the sham taVNS group. Descriptives for behavioral tasks and tests are available in Table 2. At first glance, it can be seen that average scores across all groups improved from Day 1 to Day 2, suggesting that the training effected learning. It also appears that ceiling performance was not achieved for any group or session, suggesting that more training days would be necessary to achieve mastery over these new tone words, and also that there is good variability in training outcomes for analysis.

Table 2. tVNS group means (standard deviations) for lexical recognition and recall tests.

Word Learning Outcome Measure	Peristimulus (n = 17)		Priming (n = 17)		Sham (n = 35)	
	Day 1	*Day 2*	*Day 1*	*Day 2*	*Day 1*	*Day 2*
Accuracy (% correct)						
Recall	52.8 (37.3)	69.4 (33.3)	54.2 (29.9)	84.3 (24.2)	34.8 (28.7)	65.0 (30.3)
Recognition - Match	73.6 (14.5)	84.1 (11.2)	78.0 (13.2)	90.2 (8.35)	70.4 (12.4)	84.5 (12.2)
Recognition - Mismatch	65.4 (19.1)	76.8 (18.4)	61.7 (18.9)	80.4 (18.7)	55.8 (16.9)	73.7 (18.9)
Reaction Time (ms)						
Recognition - Match	922 (113)	846 (144)	955 (128)	835 (117)	932 (157)	888 (138)
Recognition - Mismatch	1001 (125)	917 (143)	1069 (98)	940 (129)	1016 (162)	988 (155)

To answer the first research question, whether taVNS improves behavioral performance on lexical recognition or recall, the priming, peristim, and sham taVNS groups were compared with binomial logistic mixed-effects models (MEMs) for accuracy (recognition and recall) and a linear MEM for RT (recognition only; participants were not given a response deadline for the recall task). The MEM for recognition RT analyzed only correct trials, with spurious responses excluded (responses <60 ms, <1% of the data). All MEMs were run with the lme4 (Bates et al., 2015) package in R (R Core Team, 2019), and model testing to arrive at the models of best fit for random and fixed effects (including covariates for musicianship and PCID) was performed with the buildmer package (Voeten, 2019), using the Satterthwaite approximation for degrees of freedom for linear MEM *p*-values. Final models of best fit are reported below.

Accuracy results for the lexical recognition test are shown in Table 3 and plotted in Figure 2 for the effects of interest. There was a positive effect of peristim tVNS over sham on

mismatch trials (est. = 0.581, *SE* = 0.192, *p* = .002), although not for match trials (est. = 0.267, *SE* = 0.195, *p* = .171 when releveling model baseline to match trials). There was no effect of priming tVNS over sham for mismatch trials (est. = 0.184, *SE* = 0.194, *p* = .343), but there was a significant effect of priming tVNS over sham on match trials (est. = 0.403, *SE* = 0.198, *p* = .041 when releveling model baseline to match trials) and, when releveling, a marginal difference from priming to peristim for mismatch trials (est. 0.397, *SE* = 0.229, *p* = .083) and no difference for match trials. All of the effect sizes were consistent from training day 1 to 2, as everyone improved at the same (logarithmic) rate. Musicianship (est. = 0.190, *SE* = 0.088, *p* = .032) and PCID (est. = 0.349, *SE* = 0.090, *p* < .001) were both significant.

Table 3. Logistic MEM for lexical recognition accuracy.

Fixed effects	**Estimate**	***SE***	***p***		
Intercept (Mismatch, Training Day 1, Sham)	-0.128	0.230	.579		
Task Condition (Match)	0.718	0.141	<.001		
Training Day (Day 2)	1.072	0.118	<.001		
tVNS Condition (Peristim)	0.581	0.192	.002		
tVNS Condition (Priming)	0.184	0.194	.343		
Task Cond. (Match) X Day (Day 2)	0.012	0.095	.897		
tVNS Cond (Peristim) X Task Cond (Match)	-0.314	0.176	.073		
tVNS Cond (Priming) X Task Cond (Match)	0.219	0.178	.219		
Musicianship	0.190	0.088	.032		
Non-Linguistic Tone Aptitude	0.349	0.090	<.001		
Random effects	**Variance**	***SD***	**Correlation**		
Intercepts | Participant	0.432	0.657			
Task Cond. (Match) | Participant	0.371	0.610	-.47		
Day (Day 2) | Participant	0.613	0.783	.19	.24	
Task Cond. (Match) X Day (2) | Participant	0.259	0.509	.17	-.51	-.77
Intercepts | Item (presented sound)	0.076	0.276			
Task Cond (Match) | Item	0.144	0.379	-.21		
Day (Day 2) | Item	0.046	0.213	.27	.19	

Number of obs.: 29,794; Participants: 69; Items (unique presented sound files): 18

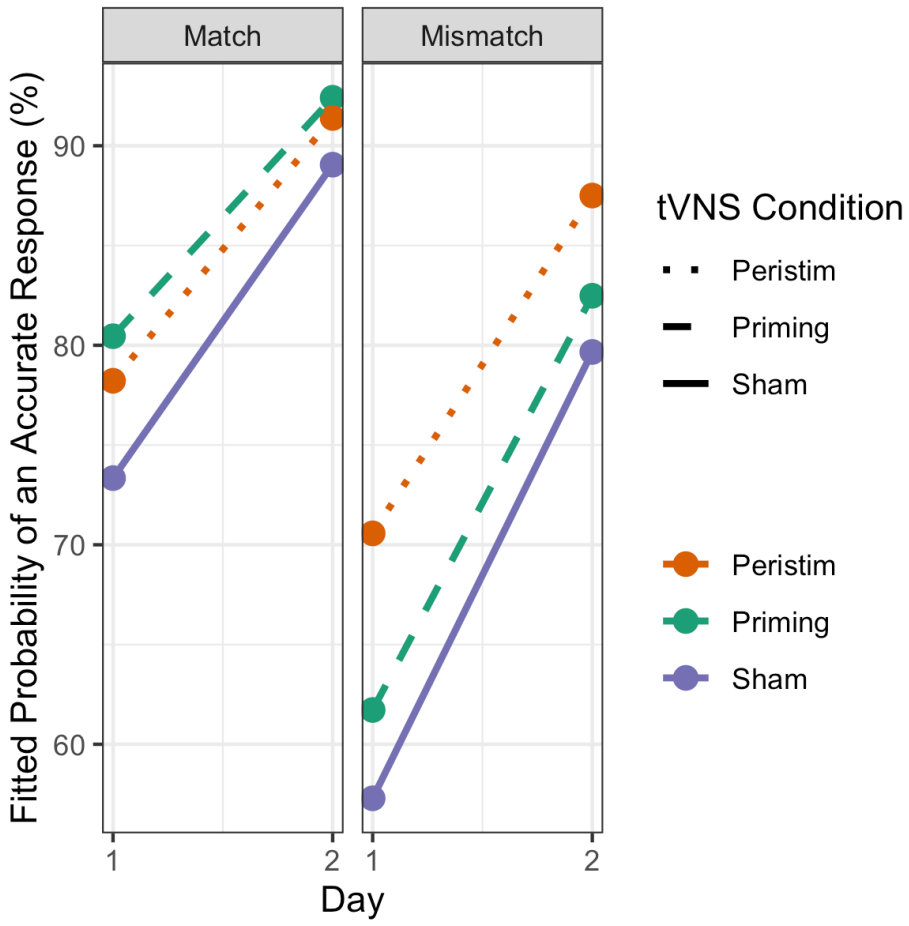


Figure 2. Modeled effects of tVNS on lexical recognition test accuracy.

RT results for the lexical recognition test related to our research questions are shown in Table 4 and plotted in Figure 3. No differences in priming or peristim tVNS over sham were observed at day 1 or day 2 (*p*s > .10), but there was a significant interaction for priming tVNS and training day indicating that the priming tVNS sped up from day 1 to day 2 significantly more than the sham group (est. = -0.107, *SE* = 0.036, *p* = .004). For the peristim group, there was no difference from either priming or sham (*p*s > .10). The observed effect for priming tVNS was

consistent across match and mismatch trials. The covariates of musicianship and PCID were not significant (*p*s > .10).

Table 4. Linear MEM for lexical recognition RTs.

Fixed effects	**Estimate**	***SE***	***p***		
Intercept (Mismatch, Training Day 1, Sham)	6.878	0.025	<.001		
Task Condition (Match)	-0.108	0.019	<.001		
Training Day (Day 2)	-0.028	0.022	.204		
tVNS Condition (Peristim)	-0.015	0.041	.715		
tVNS Condition (Priming)	0.050	0.041	.225		
Task Cond. (Match) X Day (Day 2)	-0.005	0.017	.793		
tVNS Cond (Peristim) X Day (Day 2)	-0.048	0.036	.182		
tVNS Cond (Priming) X Day (Day 2)	-0.107	0.036	.004		
Random effects	**Variance**	***SD***	**Correlation**		
Intercepts | Participant	0.019	0.138			
Task Cond (Match) | Participant	0.009	0.096	-.17		
Day (Day 2) | Participant	0.014	0.117	-.32	-.03	
Task Cond (Match) X Day (Day 2) | Participant	0.006	0.077	.21	-.66	-.16
Intercepts | Item (presented sound)	0.001	0.036			
Task Cond (Match) | Item	0.004	0.060	-.42		
Day (Day 2) | Item	0.001	0.025	-.16	-.35	
Task Cond (Match) X Day (Day 2) | Item	0.002	0.048	.03	.51	-.38
Residual	0.090	0.301			

Number of obs.: 21,949; Participants: 69; Items (unique presented sound files): 18

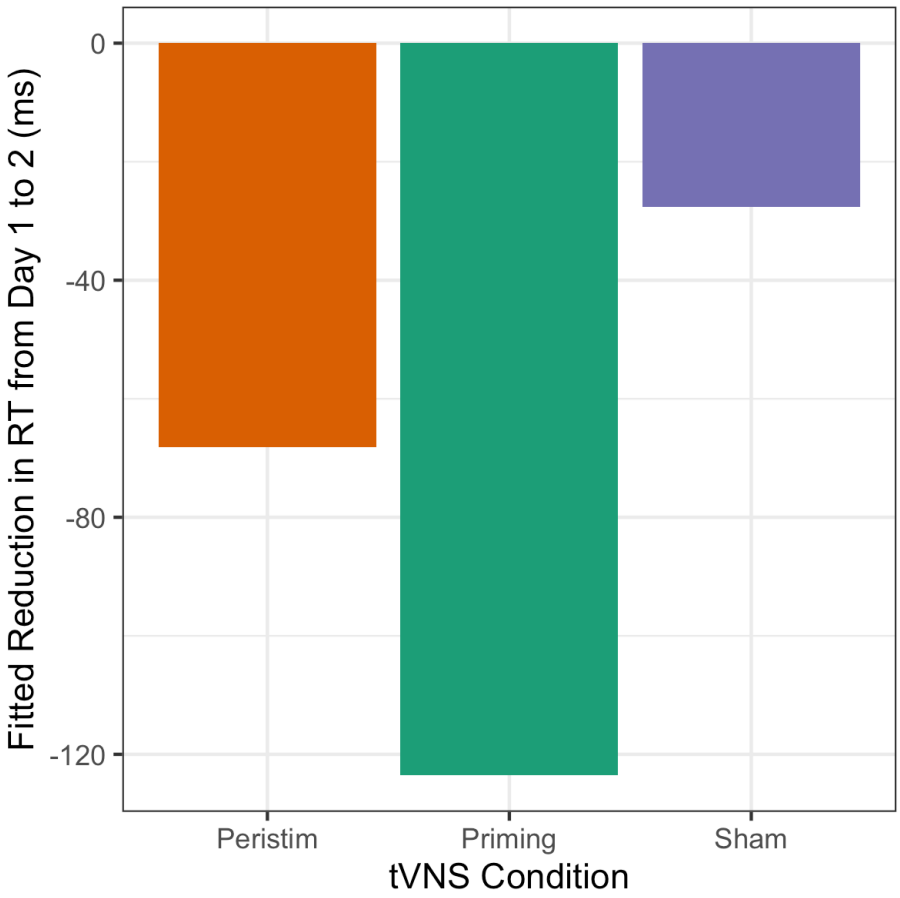


Figure 3. Modeled effects of tVNS on reductions in lexical recognition test RT.

Accuracy results for the lexical recall test are shown in Table 5 plotted in Figure 4. Priming tVNS was associated with better recall performance over sham (est. = 1.179, *SE* = 0.500, *p* = .018). Accuracy was marginally better with peristim versus sham tVNS (est. = 0.912, *SE* = 0.510, *p* = .072), and no difference was observed between priming and peristim (*p* > .10 when releveling the model baseline). These group effects were consistent from training day 1 to day 2 as everyone improved from day 1 to 2 at the same (logarithmic) rate. The covariate of

musicianship was not significant ($p > .10$) but PCID was significant (est. = 0.676, $SE = .211$, $p = .001$).

Table 5. Logistic MEM for lexical recall accuracy.

Fixed effects	**Estimate**	***SE***	***p***
Intercept (Training Day 1, Sham)	-0.867	0.309	.005
Training Day (Day 2)	1.872	0.244	<.001
tVNS Condition (Peristim)	0.917	0.510	.072
tVNS Condition (Priming)	1.179	0.500	.001
Non-Linguistic Tone Aptitude	0.676	0.211	.018
Random effects	**Variance**	***SD***	**Correlation**
Intercepts | Participant	2.402	1.550	
Day (Day 2) | Participant	1.531	1.237	-.23
Intercepts | Item	0.076	0.276	

Number of obs.: 1,236; Participants: 69; Items: 9

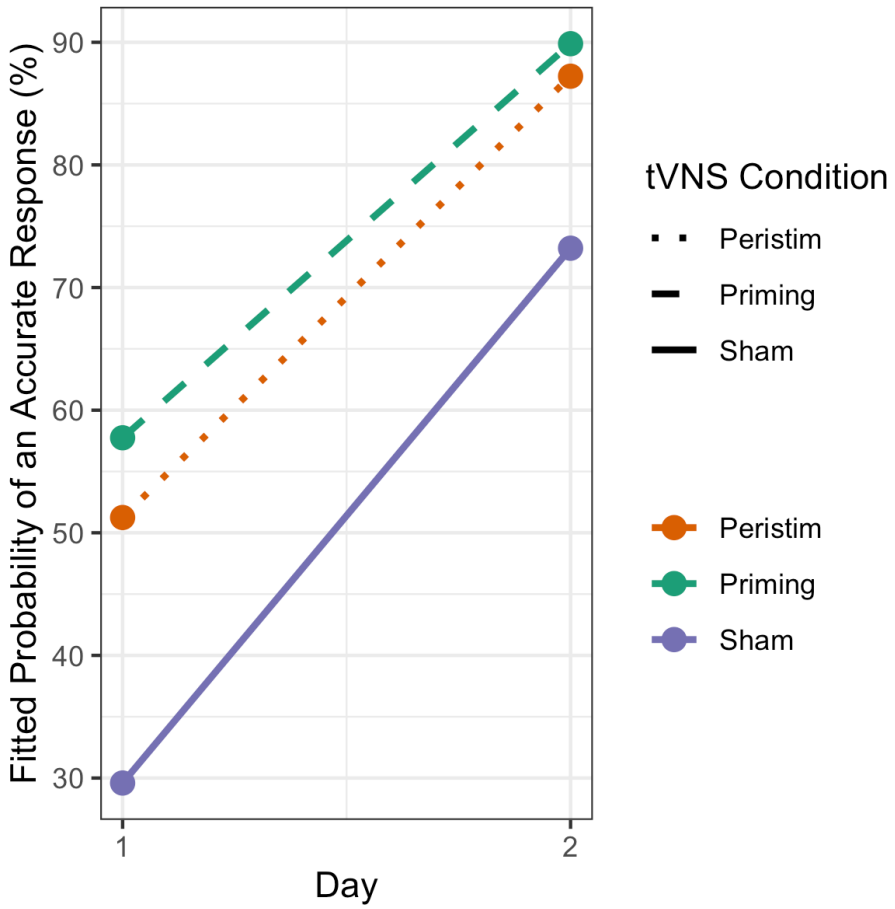


Figure 4. Modeled effects of tVNS on lexical recall test accuracy.

4.2.2 Pupillometry reveals differences in effort by taVNS group during learning

After finding differential benefits of stimulation on behavioral performance, the impact of stimulation on pupillometry was examined to try to tease apart the mechanistic differences between groups during learning to answer our second research question. In pupillometry analyses, the entire pupil response over the course of a trial is evaluated, as group differences for the pupil response can be seen in three ways after controlling for variation in both participants

and trials: (1) whether one group has an earlier peak in the pupil response than the other (quicker deployment of effort), (2) whether one group has a more peaked (effortful) response than another group, and (3) whether one group's response drops off more quickly over the time course of a word learning trial (less sustained effort over time). Pupillometry data from the passive word learning task were analyzed with generalized additive mixed modeling (GAMM), a processor-intensive analysis in which each time point from every trial from every participant can be analyzed.

Data were preprocessed in three steps. (1) Data were downsampled to 50 Hz (one datapoint every 20 ms) as recommended for GAMMs by van Rij et al. (2019), since above that the added detail does not significantly change the results but does significantly increase the time it takes a computer to calculate the model. (2) The 750 ms baseline period before each trial was subtracted from the trial for each person so that any observed differences between the groups are due to tVNS impacting the encoding of information in a specific trial rather than conflating it with any effects of tVNS on general arousal. (3) Any trials for which more than 33% of the data were missing (due to blinks, saccades, looking offscreen, etc.) were rejected from analysis. This last step resulted in fewer than 15 usable trials for 2 participants on one training day, who were then excluded from this analysis, resulting in 35 sham, and 15 priming tVNS participants across two training days for this analysis. Importantly, the number of removed trials was not associated with any one particular condition, and thus should not impact the pattern of observed results. GAMMs were implemented with the mgcv package (Wood, 2017) following previous recommendations in applying GAMMs to pupillometry data and language science data (Sóskuthy, 2017; van Rij et al., 2019), including an autoregressive model and random smooths

for participants and items. GAMMs provide a more appropriate analysis than growth curve analysis for pupillometry data in particular because they can deal with the issue of autocorrelation (that the position of the pupil at one time point is correlated with its position at the next time point, increasing Type I error if not controlled for) as well as the fact that pupil size may be influenced by the pupil's position relative to the eyetracker camera (van Rij et al., 2019). Through a computationally intensive algorithm, GAMMs objectively find the number of inflections for the pupil response curve that can support the data. Unlike more traditional analyses, the summary table usually has no utility for inferring statistical significance due to the complexity of the smooth terms for each curve, but significance can be determined by first testing a model with and without the parametric (traditional predictors that one would include in MEMs or regression models) and smooth terms (predictors specific to GAMMs that allow the penalized estimation of a non-linear relationship) of interest followed by visual inspection of difference curves (subtracting one curve from another) to inspect whether a difference curve and its confidence interval are distinct from zero (Sóskuthy, 2017; van Rij et al., 2019).

Model testing determined that including parametric and smooth terms for tVNS condition, training day, and their interaction significantly improved model fit ($\chi^2(14) = 131.14, p < .001$). The final model's summary table with all terms is presented in Table 6 and the estimated TEPRs (task-evoked pupil responses) are depicted in Figure 5, which shows the time course of a trial on the x-axis (0 to 1750 ms), pupil size on the y-axis, and different descriptive TEPR curves for each of the three conditions on each day.

Table 6. GAMM summary table for passive word learning pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Training Day 1, Sham)	0.934	6.154	.879
Day 1 X Peristim	-3.020	9.556	.752
Day 1 X Priming	-9.081	10.197	.373
Day 2 X Sham	2.422	3.462	.484
Day 2 X Peristim	-9.348	9.533	.327
Day 2 X Priming	-3.165	10.151	.755
Approx. significance of smooth terms	**Edf**	**Ref.df**	***p***
s(Time):Day 1 X Sham	8.307	8.620	<.001
s(Time):Day 1 X Peristim	3.018	3.744	.431
s(Time):Day 1 X Priming	1.068	1.097	.323
s(Time):Day 2 X Sham	8.282	8.603	<.001
s(Time):Day 2 X Peristim	7.436	8.324	<.001
s(Time):Day 2 X Priming	5.441	6.825	.104
s(X gaze position, Y gaze pos.)	28.836	28.998	<.001
te(Time, Non-Ling. Tone Aptitude)	60.109	67.753	<.001
s(Time, Participant)	289.765	331.000	<.001
s(Time, Item)	371.879	449.000	<.001

Number of obs.: 719,951; Participants: 67; Items: 90

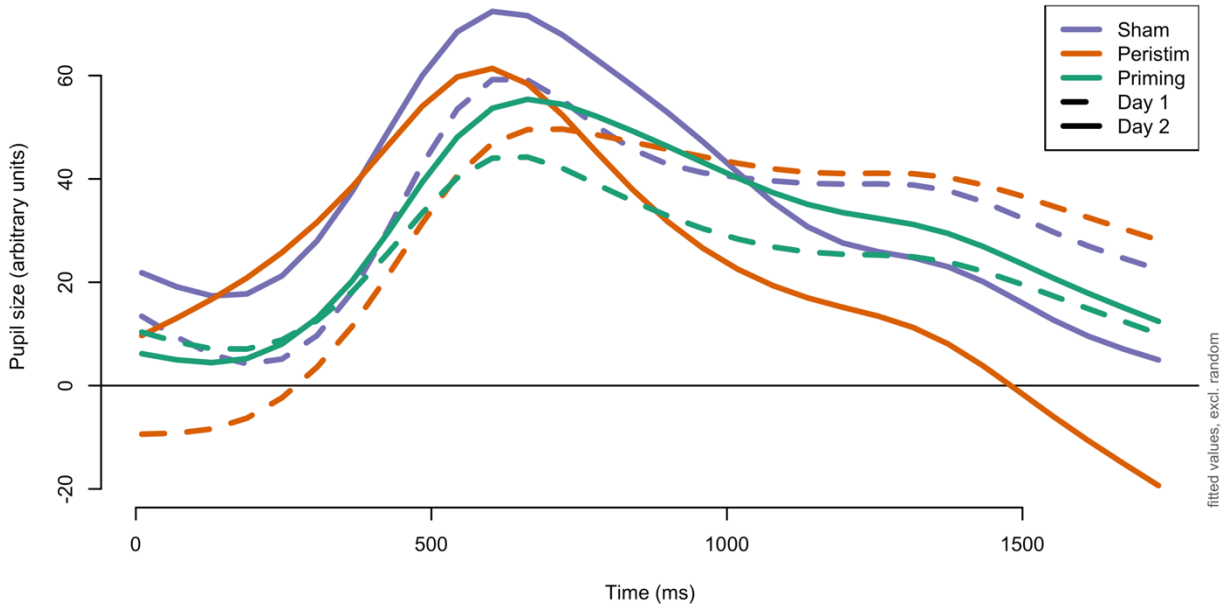


Figure 5. Descriptive model curves from the passive word learning pupillometry GAMM.

The most robust group differences were observed in changes from training day 1 to 2 and can be observed in Figure 6, which shows the difference curves from day 1 to day 2 for each of the three groups. Statistical significance is supported by inspecting where each of the three curves is different from zero. For the peristim and sham groups, the TEPR increased during the early part of a trial from day 1 to day 2, reflecting an earlier deployment of effort and a less sustained response on the second day of training compared to the first. The timecourse of the TEPR did not change overall from day 1 to day 2, but there was a larger response on day 2, reflecting more engagement of cognitive effort from day 1 to 2. Comparing groups—inspecting where each of the three curves separated from each other, a clear effect for the peristim group

emerged such that, from day 1 to day 2, there was less sustained effort during a learning trial compared to both priming tVNS and sham tVNS. Sham tVNS also appeared to have a significantly less sustained response than priming tVNS, but this effect was much less robust. Peristim tVNS also showed evidence of more effort being recruited earlier during a learning trial than priming.

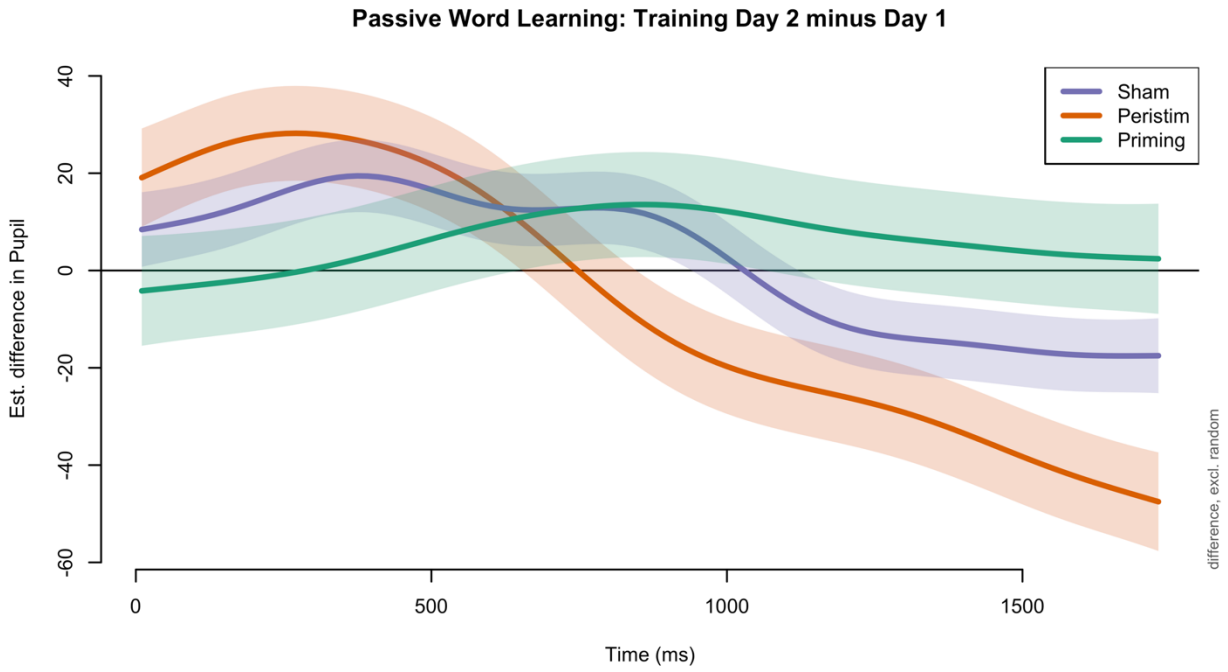


Figure 6. GAMM difference curves and their confidence intervals showing significant differences within and between groups for the change in TEPR from training day 1 to training day 2.

4.3 Discussion

Two different taVNS interventions were observed to elicit performance improvements over a sham control in a double-blind study of learning novel pseudowords featuring lexical

tone. For lexical recognition, peristim taVNS—neurostimulation time-locked to stimulus presentation—showed an advantage over sham in lexical recognition accuracy of mismatch trials by about 5-10% but not in RT, while priming taVNS—neurostimulation 10 minutes continuously prior to a task or test—showed an RT advantage of about 100 ms and a significant effect on accuracy over sham by about 3-6% on match trials. For lexical recall (which included a small number of test items), priming had a positive effect on accuracy over sham by about 15-30% (depending on day) and peristim showed a marginal effect on accuracy over sham by about 15-20% and was not significantly different from priming.

While these results indicate learning advantages for taVNS recipients, the effect sizes are not easily compared to those found in other tone word learning studies since these typically involve more training sessions and lexical items and focus on interactions between learner characteristics and training design (e.g., Perrachione et al., 2011) or stimulus manipulations (e.g., Antoniou & Wong, 2016). The passive word learning task and recognition test used here were based on Dittinger et al.'s (2016) study of professional musicians and non-musicians learning nine Thai words that included tonal contrasts. Although analyses of N400s elicited during these tasks suggested more efficient word learning for musicians, there were no corresponding differences in accuracy or RT scores. However, musicians were more accurate on a semantic relatedness task involving the same words, which does provide one important benchmark: Word learning advantages attributed to a single session of taVNS in the present study emerge as early in training as the learning advantages attributed to years of musical training in Dittinger et al. (2016).

There are a few reasons that priming and peristim taVNS may have had differential impacts on accuracy and RT for recognition and recall tests. One is that the recall test was short; given that there are only nine words to learn and nine items on the recall test at each session, 'large' improvements only reflect a relatively small difference in the number of items correctly recalled. Another explanation may be that taVNS may facilitate different aspects of word learning: number of items that can be learned (via peristim) and speed with which learned items can be accessed (via priming). Additionally, the observation that stimulation type differently impacted changes in match vs. mismatch accuracy suggest a need for future research regarding the relative benefits of priming and peristim on attention to relevant information versus inhibition of distracting information.

Given that taVNS is hypothesized to affect production of NE and effort allocation, we used pupillometry to investigate the allocation of cognitive effort for each group during the passive word learning task of the experiment. Supporting our expectation that a smaller TEPR reflects a better integration of newly learned words (particularly when the task is not so difficult that people give up), we observed a significantly faster drop off in the TEPR for peristim than for sham from training day 1 to day 2. This suggests that the peristim group required less sustained effort for a given learning trial than sham while memorizing words, and later performed better on accuracy for those words on both days. For priming, the results are less clear, as there were not robust differences from sham. Overall within the priming group from day 1 to day 2, there was a slightly more peaked TEPR. This weaker effect may have arisen if, as previous work suggests, priming taVNS impacts task-evoked effort indirectly compared to peristim. Priming may alter the tonic firing pattern of the LC, which in turn allows for participants to be in the optimal

arousal state to exert mental effort (Aston-Jones & Cohen, 2005). The task in this paper as designed is not appropriate for a tonic analysis, but future experiments should align aspects of this work with invasive animal models of auditory learning with taVNS to explore this possibility.

Even after only one day of training, taVNS had positive effects on lexical tone word learning as measured on lexical recognition and recall tests. These advantages persisted into the second day of training. Moreover, the two types of taVNS interventions, priming and peristim, resulted in different types of benefits for learning. Despite peristim having less total stimulation duration, it resulted in as good or better accuracy than priming. This suggests that the total amount of stimulation may be less relevant than the nature of the stimulation, i.e., time-locked to a given stimulus or primed continuously before a task or test.

4.4 Conclusions and Next Steps

The current Chapter's results suggest a promising future for taVNS as fast and effective language learning support. Improvements were observed almost immediately and coincided largely with pupillary changes that reflected a predicted influence of taVNS on the LC-NE system. While this double-blind study design strengthens causal inferences about taVNS effects on tonal word learning, future work should establish taVNS parameters and protocols that optimize efficacy for different learning tasks and learner characteristics. The present results showed taVNS benefits between groups balanced on pitch aptitude and musicianship, however

future work would benefit by more directly examining how taVNS efficacy interacts with these variables known to predict L2 tone learning success.

As a first attempt to apply tVNS to language learning, this study administered priming and peristimulus taVNS during training and test phases. At this juncture, these results and the previous literature suggest that neurostimulation, when paired with behavioral language learning approaches may provide a much-needed practical boost for adult language learners to overcome the inherent difficulty in learning a second language. However, the origin of the benefit of taVNS in this chapter is still unclear: while it examined practical outcomes of recognition and recall of new vocabulary, it is possible that taVNS also facilitated learning at the phonological level that in turn resulted in better lexical learning outcomes. Spending less time on low-level features of language comprehension in the classroom such as L2 phonology, improved speech segmentation, or rote vocabulary learning means more instructional time can be spent on learning higher level linguistic features, like pragmatics, which are equally vital to developing advanced language proficiency. More work lies ahead to fully understand how taVNS can best support language learning and to maximize its benefits to language learners at different stages of learning and with different learning strengths and background.

Chapter 5: taVNS-facilitated Phonological Tone Learning

5.1 Introduction and Motivation

The results from Chapter 4 have revealed promising practical utility for a taVNS intervention to enhance Mandarin lexical tone learning outcomes at the lexical level. The next step toward optimizing this training intervention is to help tease apart at what level it is operating: do taVNS-facilitated outcomes only arise at the lexical level or do they arise earlier, at the phonological level? While it has long been found that lexical tone learning is more than simply phonological tone categorization (Pelzl, 2019a), some research has connected the dots from training benefits on tone categorization leading to improvements on tone word learning (e.g., Cooper & Wang, 2013; Ingvalson et al., 2013). To complement these findings, previous research in humans and with animal models show VNS effects on low-level auditory learning (e.g., Borland et al., 2018; Engineer et al., 2011; Kilgard, 2012; Llanos et al., 2020). Thus, enhanced phonological tone learning leading to improved outcomes for lexical tone learning is another potential explanation worth exploring for the taVNS-facilitation observed in Chapter 4.

In addition, it is well-established that not all tonal contours are processed as similarly difficult (see Pelzl, 2019a, for a review). The present training study corpus uses Mandarin tones 1, 2, and 4. The taVNS facilitation effect sizes in Chapter 4 are essentially averaged across performance on those tones. While it's possible that both easier and harder tones would be affected by taVNS (Llanos et al., 2020), by ignoring differences in tonal contour we are sidestepping this empirical question and may be obscuring larger effect sizes for some tones and smaller ones for others. As noted in Chapter 2, tone 1 will likely be the easiest tone, but it is

unclear if tone 2 or tone 4 will be harder than the other. English NSs may be more sensitive to early starting pitch differences in height (e.g., Maddox et al., 2013) making tone 2 more of a stand-out (easier) than tone 4, which has a closer starting pitch height to tone 1. However, natural Mandarin tonal contours (like the ones used in the present corpus) instantiate a larger start-to-end contour change for tone 4, the high-to-low falling tone, than tone 2, the mid-to-high rising tone. This difference may in turn lend more acoustic salience to the learning of tone 4 vs tone 2. Chapter 5 will address the taVNS-facilitation at the phonological level as well as differences between tonal contours head-on.

5.1.1 VNS and auditory learning

In addition to the studies cited in Chapter 2 supporting a link between VNS and auditory cortex plasticity in animal models (Borland et al. 2018, Engineer et al., 2011, Kilgard, 2012), and the results of Chapter 4 (Pandža et al., 2020) supporting a link between taVNS and auditory learning in humans at the lexical level, there is also new work (Llanos et al., 2020) supporting a link between taVNS and auditory learning in humans at a more basic phonological level; that is, category learning of tones in the absence of word learning.

Since the publication of the results in Chapter 4 (Pandža et al., 2020), a more recent study by Llanos et al. (2020) investigated the impact of peristim taVNS on the learning of specific Mandarin tones, that is, whether the difficulty of a specific tone contour mattered for taVNS efficacy on (non-lexical) tone categorization. Per Llanos et al. (2020), Native English learners of tonal languages are more sensitive to differences in pitch height (i.e., as related to tones 1 and 3

in Mandarin) rather than pitch direction (tones 2 and 4). The authors *a priori* acknowledged two possibilities: (1) taVNS would facilitate learning easier-to-learn tones (T1 and T3) by enhancing arousal, benefitting the stimuli with the greatest perceptual salience, and/or (2) taVNS would facilitate learning of harder-to-learn tones (T2 and T4) by increasing sensitivity to pitch direction and increasing their perceptual salience. Llanos confirmed the first hypothesis only, as taVNS only enhanced tonal category learning when taVNS was paired with easier-to-learn tones.

The Llanos et al. (2020) training paradigm consisted of five Mandarin syllables spoken with each of the four tones by two speakers. Each of the 40 tokens was presented in each of six training blocks in a one-session study, in which participants indicated with a button press the tone category of each stimulus and were given yes/no feedback. A seventh speaker-generalization block was also administered, using all stimuli spoken by two different speakers and without feedback. Thirty-six participants were split into three (n = 12) groups: (1) tVNS-easy, in which participants only received taVNS paired with easy tones 1 and 3, (2) tVNS-hard, in which participants only received taVNS paired with hard tones 2 and 4, and (3) a sham control, in which participants still did calibration but did not receive stimulation for the purposes of single blinding of participants. It is also worth noting the authors briefly mention in their discussion the testing of a fourth taVNS group, a peri-response (vs peristim) condition in which participants only received stimulation after indicating a correct response. However, this peri-response group showed no gains over the control, either because taVNS does not operate via reward-related neuromodulatory signals or because there were too few trials as the peri-response group received 30% less total stimulation on average compared to their active peristim taVNS stimulation groups. The tVNS in their study targeted the auricular branch of the vagus nerve at

the cymba concha and cymba cavum sites of the left ear (rather than the inner ear as in the present corpus) using a staircase calibration procedure to find a participant's perceptual threshold, then delivering stimulation during the task 0.2 mA below their perceptual threshold (3.0 mA maximum). The peristim pulse train began 300 ms before the onset of each targeted auditory stimulus (depending on group assignment) and lasted 250 ms through roughly half the duration of the auditory stimulus.

In a single-blind design, Llanos et al. (2020) found an accuracy advantage for taVNS only for the tVNS-easy group on the easy (T1 and T3) tones, and this effect persisted into the generalization block, with a new speaker, no feedback, and no peristim taVNS. The tVNS-easy group improved across all tone categories (averaged across easy and hard) by about 26% by the third block, the same improvement the control group reached only in the sixth training block. Curiously, while the effect size appeared similar between tVNS-easy and tVNS-hard groups, this comparison was not reported. Tandem EEG analyses did not find evidence of any taVNS-induced changes in the representation of the tonal auditory stimuli. The authors interpret their results as peristim taVNS-facilitation of perception and memory consolidation of perceptually salient categories (easier tone categories).

There are a number of differences between Llanos et al. (2020) and the present corpus that are worth highlighting. One important consideration of Llanos et al. (2020) is that their Mandarin tone learning paradigm, while a low-variability type of tone training overall, was nonetheless a higher-variability training than the present corpus, using 40 stimuli (5 syllables x 2 talkers x 4 tones) for training, to contrast with the lower-variability training of the present corpus under consideration using only 18 training stimuli (3 syllables x 2 talkers x 3 tones), and so their

training paradigm may have had at least a slightly higher overall difficulty than the present training and comparisons of results may need to be interpreted in that light.

Additionally, the Llanos et al. (2020) training only occurred in one session rather than the present dissertation's two training sessions on consecutive days, allowing for a more nuanced investigation here. Further, all testing of the stimuli occurred during one training task for Llanos et al. (2020), and there was no lexical component to the study; it was purely phonological learning using yes/no feedback during training with a generalization block at the end for speaker. This task is very similar to the phonological tone categorization training and tests in the present corpus (3.2.3.7), although feedback here, in addition to yes/no, also indicated the correct tone in the event of an incorrect response in the training task. The Llanos et al. (2020) task also had a generalization block, with new speakers and no feedback, which is similar to the phonological tone categorization pretest and posttest in the present corpus, which does not have feedback, has speaker *and* syllable generalization stimuli, and provides a more structured look at outcomes and generalization after a two-day training program, separate from the phonological tone categorization training itself.

Another difference with the present study's corpus is that, while Llanos et al. (2020) primarily investigated peristim taVNS and, to a lesser extent, peri-response taVNS, the present study directly compares peristim and priming taVNS with equal footing. Additionally, the present study's peristim condition is agnostic to tone difficulty. In this way, participants overall receive more total stimulation and all tones receiving stimulation is in line with the two *a priori* hypotheses in Llanos et al. (2020): in theory, taVNS could positively affect both easier and more difficult tones, but for potentially different reasons. By having a more within-subjects design in

regard to tone difficulty, we can make more generalizable claims when analyzing effects of taVNS on specific tones. Both priming and peristim taVNS showed promising, though interestingly separable effects in Chapter 4, so Chapter 5 will further examine whether these differences in efficacy are also reflected at the phonological level of tone learning.

One larger limitation of Llanos et al. (2020) is that, due to their particular taVNS setup, all audio was only delivered monoaurally through an insert earphone in the right ear. Not only is this less than ideal from an ecological validity perspective for auditory language learning, but this may have additionally affected their behavioral and electrophysiological results. There are strong contralateral links between the right ear and left hemisphere auditory cortex and also between the left ear and right auditory cortex. While language is primarily processed in the left hemisphere for a vast majority of individuals, lexical tone relies heavily on tonal processing capabilities in the right hemisphere. For example, Wang et al. (2001) found with their dichotic listening task that American English speakers naïve to lexical tone process Mandarin tone using both hemispheres. Plus, a recent study (Shao & Zhang, 2020) used a dichotic listening task and found that native speakers of Cantonese had poorer discrimination accuracy and longer reactions times when hearing Cantonese tones through the right ear as opposed to the left ear. The present corpus uses binaurally presented audio and does not have this potential confound.

In sum, while Llanos et al. (2020) was an important step forward, there are a number of gaps in their investigation of the effects of taVNS on mandarin tone learning that the present corpus can fill. This study is in a position to examine priming in addition to peristim taVNS, an additional consecutive day of training and testing, and the generalizability of the effects to

additional syllables in addition to additional speakers with binaurally presented Mandarin tone training and testing.

5.1.2 Research questions

The study reported in this chapter uses taVNS and investigates its effects on Mandarin tone learning across multiple outcome measures that are sensitive to changes at varying timescales (across sessions, across trials, and across milliseconds). Priming taVNS and peristim taVNS will again be contrasted with sham taVNS using data from the corpus, looking at the phonological tone categorization test (3.2.3.7.2) as a pretest, prior to any taVNS or training on Day 1, and as a posttest, after all phonological and lexical tone training on Day 2. Behavioral outcomes of accuracy and reaction time on phonological tone categorization tests, pre and post around a two-day phonological and lexical tone training, will be analyzed as indices of learning. Pupillometry was collected during the tests and pupil size will be analyzed as an index of cognitive effort. Pretest and posttest outcomes will be analyzed for T1, T2, and T4 separately to investigate tone difficulty for both trained stimuli and untrained generalization stimuli, consisting of untrained speaker, untrained syllable, and untrained speaker+syllable stimuli.

The research questions of the present study build on the first two in Chapter 4:

3) Does active (priming and/or peristim) taVNS versus sham taVNS improve behavioral learning outcomes for Mandarin phonological tone?

    a. Are easy and hard tones differentially impacted?

b. Do any effects generalize to untrained speakers and/or untrained segmental contrasts?

4) Does active (priming and/or peristim) taVNS versus sham taVNS produce a differential deployment of cognitive effort during phonological tone processing at test?
    a. Are easy and hard tones differentially impacted?
    b. Do any effects generalize to untrained speakers and/or untrained segmental contrasts?

Based on Llanos et al. (2020), it is predicted that taVNS will have a positive impact on phonological training outcomes from this corpus as well. From previous literature, it is expected that T1 will appear the easiest to learn, with T2 and T4 being more difficult. Due to the conflicting literature on tone difficulty for native English speakers, it is unclear which of T2 and T4 should be more difficult. Going on the results of Llanos et al. (2020), it may be that taVNS could only affect the easier T1; however, their design was quite different from the current corpus, mainly in that it was a somewhat higher variability training than the current design. Thus, it may be that the easier tones in their study were the easier of two already more difficult training conditions, and thus there is still no clear hypothesis for which of the tones that taVNS may affect in this study. As for the pupillometric outcomes for this task, based on the Llanos et al. (2020) results for behavioral data and the results of Chapter 4, it is predicted that taVNS will at least affect cognitive effort deployed in the peristim taVNS condition, such that there is less sustained effort than sham. No previous studies could be found exploring Mandarin tone category learning with pupillometry, but differences by individual tone may reflect less effort for

easier-to-learn tones, and perhaps less sustained effort from day-to-day for harder-to-learn tones in the active taVNS conditions.

5.2 Results

5.2.1 taVNS improves behavioral tone learning difficulty and generalizability of learning

A total of 81 participants (56 female) ages 18–34 years (M = 21.51, SD = 2.98) were analyzed, after 1 was excluded for noncompliance throughout learning tasks. There were 21 participants in the priming taVNS group, 20 participants in the peristim taVNS group, and 40 participants in the sham taVNS group. Descriptives for behavioral tasks and tests are available in Table 7. At first glance, it can be seen that average scores across all groups improved from Day 1 to Day 2, suggesting that the training effected learning. It also appears that ceiling performance was not achieved for any group or session, suggesting that more training days would be necessary to achieve mastery over these new tone words, and also that there is good variability in training outcomes for analysis.

Table 7. taVNS group means (standard deviations) for the phonological categorization test.

Phonological Categorization	**Peristimulus (n = 20)**		**Priming (n = 21)**		**Sham (n = 40)**	
	Pre Day 1	***Post Day 2***	***Pre Day 1***	***Post Day 2***	***Pre Day 1***	***Post Day 2***
Accuracy (% correct)	52.5 (16.7)	68.9 (16.2)	57.3 (18.3)	67.9 (16.9)	53.2 (14.6)	65.5 (18.0)
Tone 1: Critical	71.9 (21.7)	76.1 (20.4)	78.8 (19.0)	80.7 (18.6)	74.8 (22.8)	76.9 (18.4)
Tone 2: Critical	80.3 (21.9)	86.4 (16.4)	77.0 (22.5)	82.0 (19.0)	78.7 (19.7)	80.8 (19.5)
Tone 4: Critical	23.3 (24.6)	72.2 (31.1)	27.3 (26.0)	68.5 (23.2)	19.9 (20.5)	66.0 (27.5)
Tone 1: Generalization	68.6 (19.5)	63.9 (25.1)	72.1 (23.3)	71.7 (19.5)	72.8 (27.1)	63.7 (22.5)
Tone 2: Generalization	58.3 (25.5)	62.5 (25.7)	62.4 (25.2)	58.5 (27.0)	58.1 (18.4)	59.9 (24.2)
Tone 4: Generalization	12.5 (15.3)	52.3 (27.1)	25.9 (28.0)	46.0 (23.8)	15.0 (16.5)	45.6 (25.9)
Reaction Time (ms)	902 (147)	879 (162)	931 (108)	915 (155)	959 (135)	977 (160)
Tone 1: Critical	917 (142)	922 (184)	947 (136)	906 (171)	991 (174)	980 (168)
Tone 2: Critical	890 (209)	805 (162)	928 (159)	848 (198)	936 (140)	893 (198)
Tone 4: Critical	1035 (214)	912 (194)	1081 (223)	955 (244)	1078 (212)	1033 (222)
Tone 1: Generalization	904 (185)	914 (160)	898 (130)	949 (155)	947 (181)	1011 (194)
Tone 2: Generalization	913 (222)	824 (214)	876 (91)	856 (139)	926 (139)	949 (155)
Tone 4: Generalization	903 (226)	963 (244)	1092 (209)	1024 (180)	1111 (250)	1086 (193)

As for Chapter 4, binomial logistic and linear mixed-effects models were used to analyze accuracy and reaction time data, respectively. The MEM for phonological categorization RT analyzed only correct trials, with spurious responses excluded (responses <60 ms, <1% of the data). All MEMs were run with the lme4 (Bates et al., 2015) package in R (R Core Team, 2019).

Model testing to arrive at the models of best fit for random and fixed effects (including covariates for musicianship and PCID) was performed with the buildmer package (Voeten, 2019), using the Satterthwaite approximation for degrees of freedom for linear MEM *p*-values. Final models of best fit are reported in Table 8 and Table 9. Independent variables of interest included Session (Pretest vs. Posttest), taVNS Group (Priming vs. Peristim vs. Sham), Tone (1, 2, 4), Condition (Trained Stimuli vs. Generalization Stimuli), and their interactions. Potential covariates included Musicianship (centered at music-loving non-musician) and Non-Linguistic Tone Aptitude (centered via z-score).

Table 8. Logistic MEM for phonological categorization test accuracy.

Fixed effects	**Estimate**	***SE***	***p***			
Intercept (Tone 1, Critical, Day 1, Sham)	1.685	0.272	<.001*			
Tone 2	-0.254	0.319	.426			
Tone 4	-3.409	0.335	<.001*			
Generalization Items	-0.673	0.232	.004*			
Training Day (Day 2)	0.085	0.211	.688			
taVNS Group (Peristim)	-0.238	0.250	.341			
taVNS Group (Priming)	0.048	0.247	.847			
Tone 2 X Day 2	0.309	0.245	.208			
Tone 4 X Day 2	2.521	0.251	<.001*			
Generalization X Day 2	-0.506	0.186	.007*			
Day 2 X taVNS Group (Peristim)	0.172	0.198	.385			
Day 2 X taVNS Group (Priming)	0.233	0.202	.248			
Tone 2 X taVNS Group (Peristim)	0.370	0.309	.232			
Tone 4 X taVNS Group (Peristim)	0.106	0.352	.762			
Tone 2 X taVNS Group (Priming)	0.091	0.306	.767			
Tone 4 X taVNS Group (Priming)	0.482	0.341	.157			
Day 2 X Tone 2 X taVNS Group (Peristim)	0.061	0.232	.793			
Day 2 X Tone 4 X taVNS Group (Peristim)	0.406	0.249	.103			
Day 2 X Tone 2 X taVNS Group (Priming)	-0.384	0.233	.099^			
Day 2 X Tone 4 X taVNS Group (Priming)	-0.743	0.239	.002*			
Musicianship	0.248	0.103	.016*			
Non-Linguistic Tone Aptitude	0.537	1.111	<.001*			
Random effects	**Variance**	***SD***	**Correlation**			
Intercepts | Participant	1.037	1.018				
Tone 2 | Participant	0.881	0.938	-.26			
Tone 4 | Participant	1.178	1.085	-.45	.63		
Generalization | Participant	0.293	0.541	-.66	.03	.17	
Day 2 | Participant	0.205	0.453	-.23	-.17	-.01	.45
Intercepts | Item (presented sound)	0.421	0.649				
Day 2 | Item	0.253	0.503	-.60			

Number of obs.: 17,485; Participants: 81; Items (unique presented sound files): 36

The accuracy results for the phonological categorization test show a lack of taVNS group differences on any tone on Day 1, which makes sense given the pretest was before any taVNS stimulation for the active groups. Stimulus condition (critical vs. generalization) did not interact with Tone or taVNS Group; thus, the remaining effects for Tone and Group are consistent across

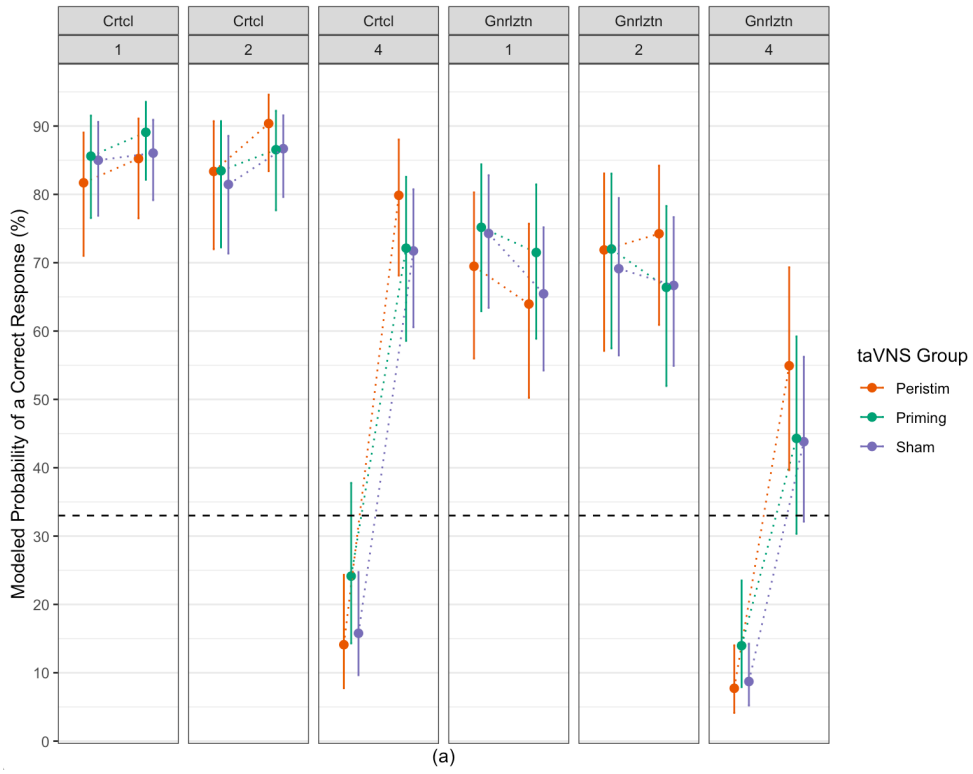
stimulus conditions. The covariates of Musicianship and Non-Linguistic Tone Aptitude were both significant and remained in the model. Reported results are averages across these effects.

At pretest for all taVNS groups, Tone 1 and Tone 2 were not significantly different from each other in difficulty (~ 80-85% probability of a correct response) while Tone 4 was harder than both (~15-25% probability of a correct response; releveled for Tone 2 vs 4: *Est.* = -3.155, *SE* = 0.319, $p < .001$) by a large magnitude. None of the three groups improved from pre to post on Tone 1. Sham and Priming did not improve on accuracy for Tone 2 from pre to post, but Peristim did (~84 to 90% improvement; releveled to Tone 2, Peristim, *Est.* = 0.627, *SE* = 0.246, $p = .011$). For all taVNS groups, accuracy on Tone 4 significantly improved from pre to post. The Peristim group improved more than sham (~14 to 80%; releveled to Tone 4 baseline: *Est*. = 0.578, *SE* = 0.225, $p = .010$), while the Priming group improved (~25 to 72%) to a lesser degree than Sham (~ 15 to 72%; releveled to Tone 4 baseline: *Est.* = -0.511, *SE* = 0.206, $p = .013$).

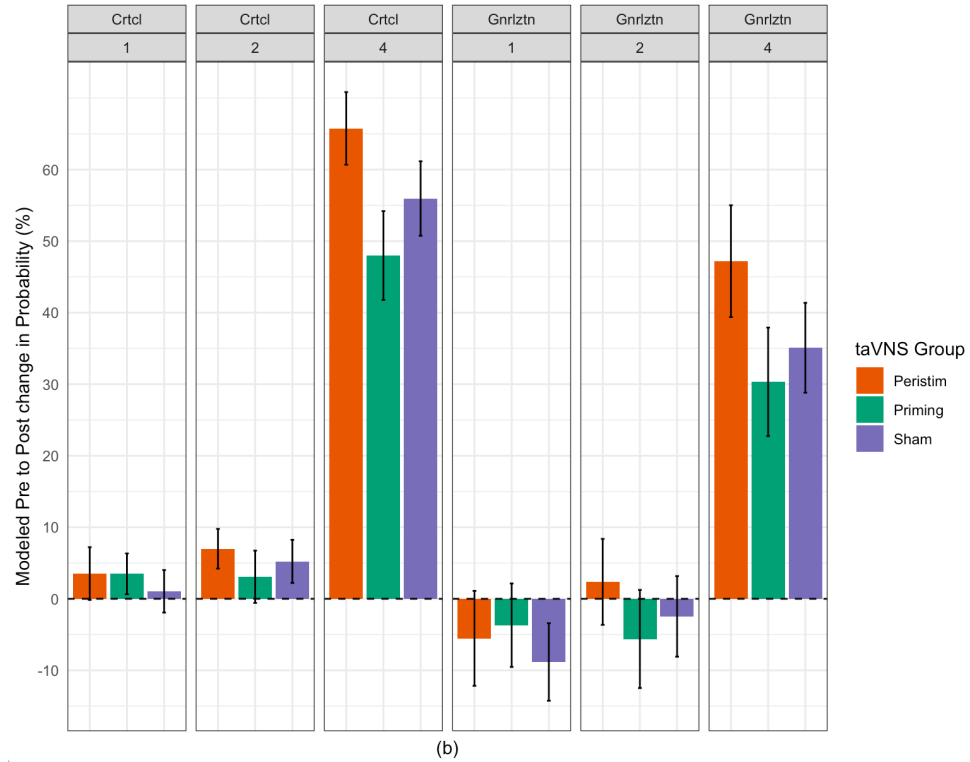
All participants performed equally worse on Generalization items across the board compared to Critical items. Given that this is true at pretest before training, it appears there is something inherently more difficult about these stimuli. Regardless, the negative Generalization X Day 2 interaction shows that any improvements for Critical items were mitigated for Generalization items. Indeed, while the Peristim, Priming, and Sham groups respectively improved by about 65%, 47%, and 56% probability on Critical Tone 4 items, they only improved roughly 47%, 30%, and 35% on Generalization Tone 4 items. On Tone 1, the Sham group in fact did *worse* day-to-day by about 5% (releveled to Generalization baseline: *Est.* = -0.422, *SE* = 0.209, $p = .043$). Peristim's day-to-day improvement on Tone 2 Critical items also disappears for Tone 2 Generalization items.

Model estimates for all effects are visualized in Figure 7a, with pre-to-post changes highlighted in Figure 7b. In sum, the results show, for Critical items, (1) Tone 4 was the most difficult tone to learn; (2) even after training, Tone 4 remained harder than Tones 1 and 2; (3) Peristim improved the most on Tone 4, more than the other groups; (4) while Sham improved to greater degree than Priming on Tone 4, they ended up at roughly the same performance (~72% probability); (5) participants across the board did worse and improved less on Generalization items.

Figure 7. Modeled accuracy results split by taVNS group, tone (1, 2, 4), and stimulus condition (critical, generalization): (a) Probability model estimates for logistic MEM of phonological categorization accuracy at pre and post. Dotted line represents chance performance at 33% percent probability. (b) Pre-to-post change in probability estimates for logistic MEM of phonological categorization accuracy to show improvement from training. Dotted line at 0% change pre to post.


Crtcl
1
Crtcl
2
Crtcl
4
Gnrlztn
1
Gnrlztn
2
Gnrlztn
4
Modeled Probability of a Correct Response (%)
0
10
20
30
40
50
60
70
80
90
taVNS Group
Peristim
Priming
Sham

(a)


Crtcl
1
Crtcl
2
Crtcl
4
Gnrlztn
1
Gnrlztn
2
Gnrlztn
4
Modeled Pre to Post change in Probability (%)
-10
0
10
20
30
40
50
60
taVNS Group
Peristim
Priming
Sham

(b)

Table 9. Linear MEM for phonological categorization test reaction times of correct items.

Fixed effects	**Estimate**	***SE***	***p***
Intercept (Tone 1, Critical, Day 1, Sham)	6.864	0.031	<.001*
Tone 2	-0.057	0.035	.107
Tone 4	0.068	0.040	.091^
Training Day (Day 2)	-0.014	0.024	.548
taVNS Group (Peristim)	-0.087	0.040	.032*
taVNS Group (Priming)	-0.045	0.039	.250
Tone 2 X Day 2	-0.050	0.020	.013*
Tone 4 X Day 2	-0.048	0.028	.087^
Day 2 X Peristim	0.013	0.042	.753
Day 2 X Priming	-0.036	0.034	.346
Tone 2 X Peristim	0.015	0.035	.663
Tone 4 X Peristim	0.059	0.047	.206
Tone 2 X Priming	0.032	0.034	.346
Tone 4 X Priming	0.078	0.044	.079^
Tone 2 X Day 2 X Peristim	-0.048	0.035	.172
Tone 4 X Day 2 X Peristim	-0.099	0.047	.033*
Tone 2 X Day 2 X Priming	-0.002	0.034	.956
Tone 4 X Day 2 X Priming	-0.069	0.044	.120
Generalization Items	-0.070	0.032	.036*
Generalization X Day 2	0.102	0.021	<.001*
Generalization X Tone 2	0.079	0.046	.093^
Generalization X Tone 4	0.123	0.053	.025*
Generalization X Peristim	0.033	0.026	.202
Generalization X Priming	0.018	0.025	.473
Generalization X Tone 2 X Day 2	-0.023	0.030	.447
Generalization X Tone 4 X Day 2	-0.054	0.041	.194
Generalization X Day 2 X Peristim	-0.069	0.037	.060^
Generalization X Day 2 X Priming	0.006	0.035	.869
Generalization X Tone 2 X Peristim	-0.023	0.037	.536
Generalization X Tone 4 X Peristim	-0.121	0.060	.044*
Generalization X Tone 2 X Priming	-0.059	0.036	.097^
Generalization X Tone 4 X Priming	-0.102	0.052	.052^
Generalization X Tone 2 X Day 2 X Peristim	-0.006	0.052	.904
Generalization X Tone 4 X Day 2 X Peristim	0.136	0.071	.057^
Generalization X Tone 2 X Day 2 X Priming	0.012	0.051	.811
Generalization X Tone 4 X Day 2 X Priming	0.110	0.065	.088
Non-Linguistic Tone Aptitude	-0.074	0.013	<.001*
Random effects	**Variance**	***SD***	**Correlation**
Intercepts | Participant	0.017	0.129	
Tone 2 | Participant	0.008	0.090	-.35
Tone 4 | Participant	0.008	0.090	-.12 .56

Day 2 | Participant	0.015	0.123	-.38	.26	.1
Intercepts | Item (presented sound)	0.002	0.050			
Residual	0.056	0.237			

Number of obs.: 10,583; Participants: 81; Items (unique presented sound files): 36

The reaction time differences for phonological categorization test accuracy in Table 9 show no taVNS group differences at pretest for Tones 2 and 4, but for Tone 1, Peristim starts out significantly faster than Sham. The four-way interaction of taVNS Group X Tone X Day X Stimulus Condition was significant and remained in the model. The covariate of Non-Linguistic Aptitude was significant, while the covariate of Musicianship was not and dropped out of the model.

At pretest, responses to Tones 1 and 2 were equally fast, while responses to Tone 4 were significantly slower for Peristim and Priming groups, and marginally slower for the Sham group. None of the groups improved significantly on Tone 1 pre to post. All of the groups equally improved (faster correct responses) on Tone 2 pre to post. For Tone 4 pre to post, Sham significantly sped up (releveled to Tone 4: *Est.* = -0.063, *SE* = 0.031, *p* .044), Peristim marginally sped up even faster than Sham (releveled to Tone 4: *Est*. = -0.086, *SE* = 0.052, *p* = .099), and Priming significantly sped up even faster than Sham (releveled to Tone 4: *Est.* = -0.105, *SE* = 0.050, *p* = .037). At posttest overall, all participants were responding equally quickly to Tones 1 and 4 and more quickly to Tone 2.

On Generalization items at post, all groups are slowest on Tone 4 compared to Tones 1 and 2. The Sham group is significantly faster for both Tones 1 and 2 (releveled to Tone 4, Day 2, Generalization: Est. = -0.089, *SE* = 0.0366, *p* = .018; Est. = -0.141, *SE* = 0.036, *p* < .001); the Priming group is significantly faster on Tone 1 and marginally even faster on Tone 2 (releveled

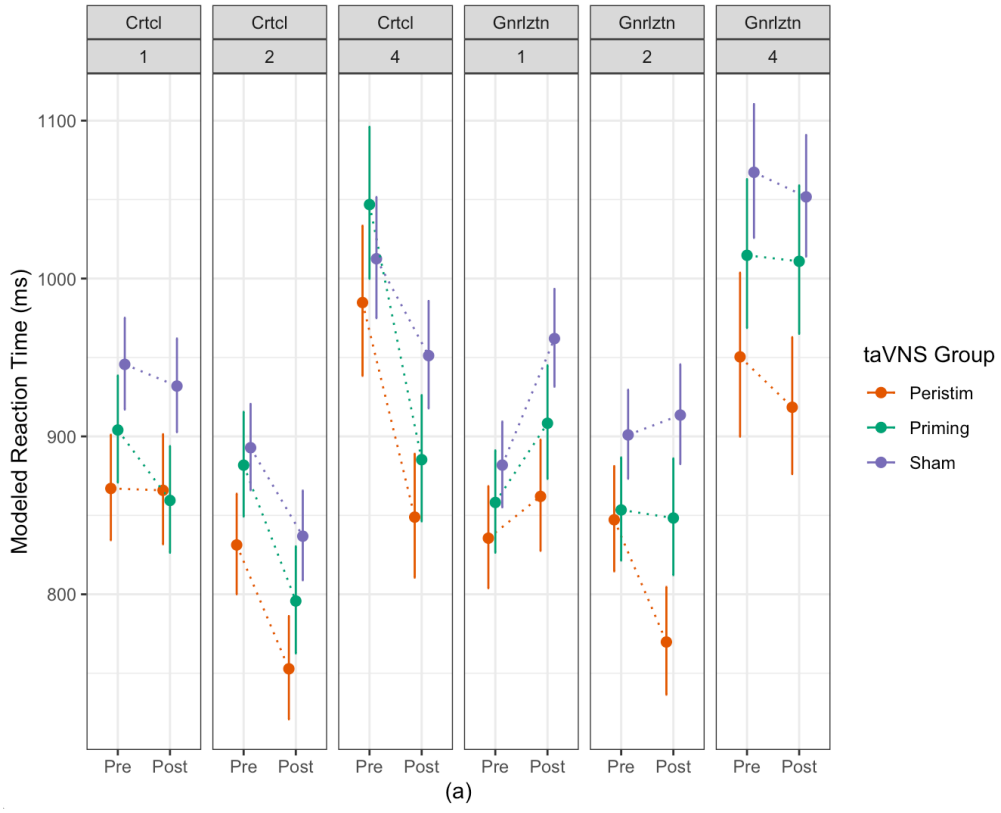
to Priming, Day 2, Generalization: Est. = 0.107, *SE* = 0.042, *p* = .013; Est. = -0.068, *SE* = 0.041, *p* = .099); the Peristim group is not significantly faster on Tone 1 but is significantly faster on Tone 2 (Releveled to Peristim, Day 2, Generalization: Est. = 0.063, *SE* = 0.043, *p* = .140; Est. = -0.113, *SE* = 0.042, *p* = .008).

For pre-to-post differences on Generalization items, the Sham group was a little faster on Tone 1 on Day 1, and significantly slower on these items from pre to post. The Priming group showed only a marginal effect for a slowdown (Releveled to Priming, Generalization: Est. = 0.057, *SE* = 0.034, *p* = .095), and Peristim was not significantly faster or slower day to day on Tone 1 (Releveled to Peristim, Generalization: Est. = .031, *SE* = .035, *p* = .378), but Priming and Peristim's differences were also not significantly different from Sham's slowdown (releveled to Sham, Generalization: Est. = -0.030, *SE* = 0.042, *p* = .473; Est. = -0.056, *SE* = 0.043, *p* = .198). For Tone 2, there were no day-to-day differences for Sham and Priming on Generalization items (releveled to Tone 2, Generalization: Est. = 0.014, *SE* = 0.026, *p* = .587; Est. = -0.020, *SE* = 0.043, *p* = .647), but there were significant improvements in RT for Peristim (Releveled to Tone 2, Generalization: Est. = -0.110, *SE* = 0.044, *p* = .014). For Tone 4, there were no significant differences pre-to-post for any taVNS group.

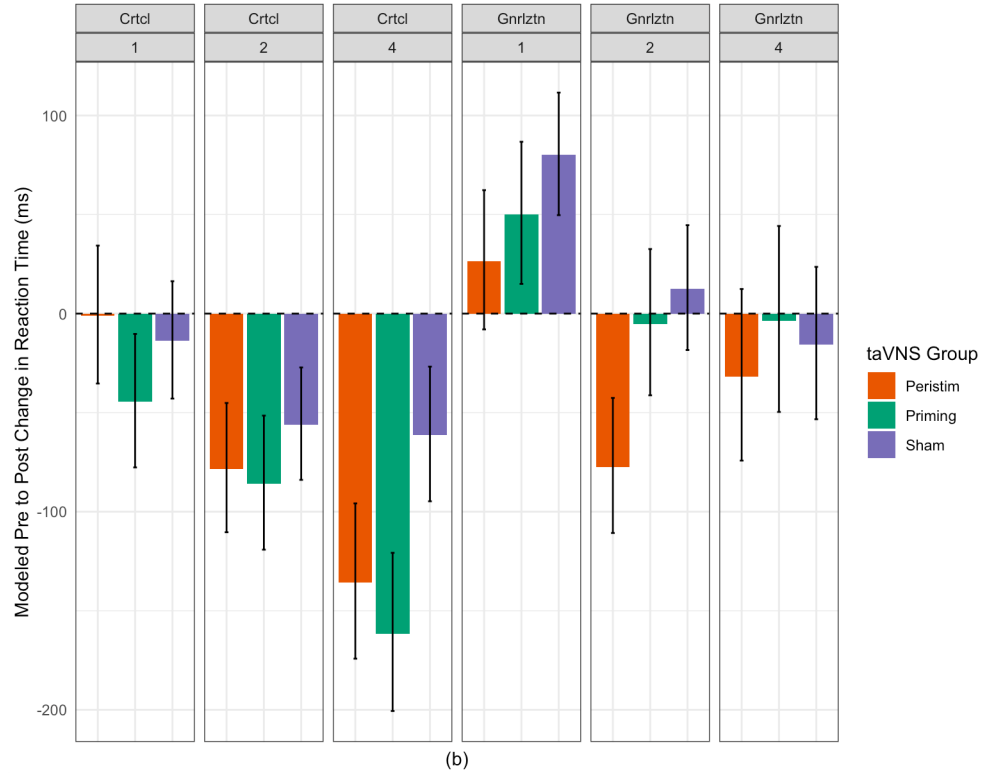
Model estimates for all effects are visualized in Figure 8a, with pre-to-post changes highlighted in Figure 8b. In sum, the results show, for Critical items, (1) Tone 4 was initially the slowest to be responded to correctly, but matched speed with Tone 1 at post; (2) after training, participants responded to Tone 2 the fastest; (3) there were no significant improvements day-to-day on Tone 1; (4) all groups improved equally on Tone 2; (5) Priming improved the most on Tone 4 vs. Sham. For Generalization items, we see (1) Sham (but not the active stimulation

groups) got slower on Tone 1 day-to-day; (2) There were no day-to-day differences for any group on Tone 4; (3) only Peristim showed any day-to-day improvements on Tone 2.

Figure 8. Modeled accuracy results split by taVNS group, tone (1, 2, 4), and stimulus condition (critical, generalization): (a) Predicted RT model estimates for linear MEM of phonological categorization RTs of accurate responses at pre and post. (b) Pre-to-post change in RT estimates for linear MEM of phonological categorization RT to show improvement from training. Dotted line at 0% change pre to post.


Crtcl
1
Crtcl
2
Crtcl
4
Gnrlztn
1
Gnrlztn
2
Gnrlztn
4
Modeled Reaction Time (ms)
1100
1000
900
800
Pre
Post
taVNS Group
Peristim
Priming
Sham

(a)


Crtcl
1
Crtcl
2
Crtcl
4
Gnrlztn
1
Gnrlztn
2
Gnrlztn
4
Modeled Pre to Post Change in Reaction Time (ms)
100
0
-100
-200
taVNS Group
Peristim
Priming
Sham

(b)

5.2.2 taVNS affects physiology for tone learning difficulty and generalizability of learning

Generalized additive mixed modeling was again used to analyze pupillometry data. Pupillometry data were preprocessed in three steps. (1) Data were downsampled to 50 Hz (one datapoint every 20 ms) as recommended for GAMMs by van Rij et al. (2019), since above that the added detail does not significantly change the results but does significantly increase the time it takes a computer to calculate the model. (2) The 500 ms baseline period before each trial was subtracted from the trial for each person. (3) Any trials for which more than 33% of the data were missing (due to blinks, saccades, looking offscreen, etc.) were rejected from analysis. GAMMs were implemented with the mgcv package (Wood, 2017) following previous recommendations in applying GAMMs to pupillometry data and language science data (Sóskuthy, 2017; van Rij et al., 2019), including an autoregressive model and random smooths for participants and items. Model testing was conducted using ordered factors (Wieling, 2018) for the parametric smooths and reference/difference smooths for the random effects structure (Sóskuthy, 2021) to arrive at the model of best fit. Using ordered factors allows direct interpretation of the GAMMs summary table and *p*-values, unlike the modeling procedure used in Chapter 4. The basic smooth of time becomes a reference smooth (similar to the intercept in a multiple regression or mixed effects model), and all other smooth terms with an ordered factor represent a difference smooth between that reference smooth and when the value of the ordered factor is set to true. For example, the reference smooth s(Time) is for Tone 1 on Day 1. The ordered factor smooth s(Time):IsTone2 would be the difference curve between Tone 2 and Tone 1 on Day 1 over the time course of a trial, and a significant *p*-value for this second term would

indicate that the difference between Tone 1 and 2 on Day 1 was significant. Plotting is still necessary to determine the direction of the effect (when the difference between tones occur and the direction of the effect).

The results of the final GAMM model are presented in Table 10. First, because the peristim group is not represented in the model, this means that the pupil response for the peristim group was not different from the sham group, and so all rows without 'Priming' in them represent the estimated pupil responses for *both* the sham and peristim groups. Given the ordered factor model specification, the parametric coefficients represent an overall intercept difference for the pupil response of a given factor.

Table 10. GAMM summary table for phonological categorization test pupillometry analysis.

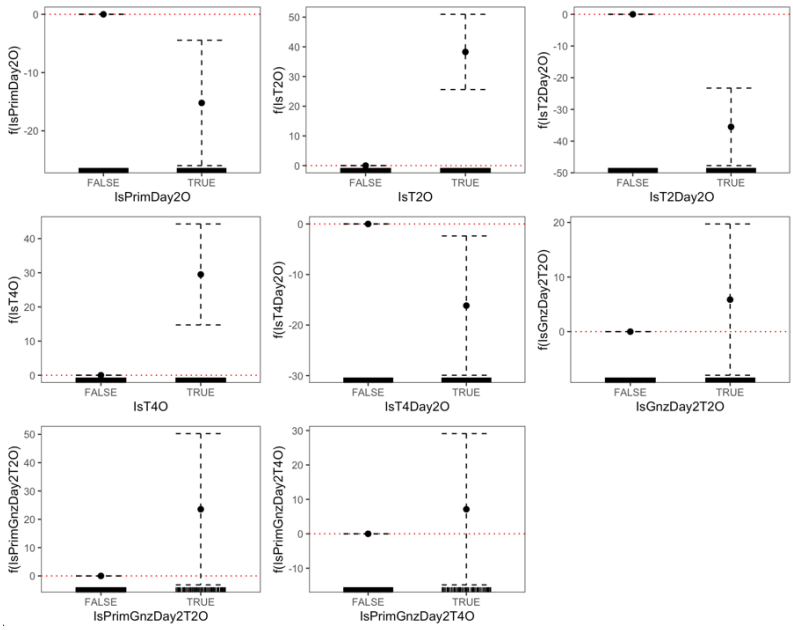
Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Tone 1, Day 1, Sham/Peristim)	-18.191	5.567	.001*
IsDay2Priming	-15.207	5.489	.006*
IsTone2	38.279	6.464	<.001*
IsTone2Day2	-35.497	6.237	<.001*
IsTone4	29.503	7.541	<.001*
IsTone4Day2	-16.148	7.029	.022*
IsGeneralizationDay2Tone2	5.868	7.069	.406
IsGeneralizationDay2Tone2Priming	23.548	13.631	.084^
IsGeneralizationDay2Tone4Priming	7.131	11.209	.525
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Tone1, Day 1, Sham/Peristim)	17.771	18.476	<.001*
s(Time):IsDay2Priming	9.296	12.190	<.001*
s(Time):IsTone2	12.024	14.605	<.001*
s(Time):IsTone2Day2	11.264	14.007	<.001*
s(Time):IsTone4	4.839	6.456	<.001*
s(Time):IsTone4Day2	5.694	7.611	<.001*
s(Time):IsGeneralizationDay2Tone2	4.576	1.905	.004*
s(Time):IsGeneralizationDay2Tone2Priming	7.579	10.098	.004*
s(Time):IsGeneralizationDay2Tone4Priming	3.705	4.975	<.001*
s(X gaze position, Y gaze pos.)	77.814	78.954	<.001*
s(Musicianship)	1.436	1.479	.809
ti(Musicianship, Time)	13.390	16.459	.003*
Random smooth terms	**edf**	**Ref.df**	***p***
s(Time, Participant)	365.022	405.000	<.001*
s(Time, Participant):IsTone2	311.329	405.000	<.001*
s(Time, Participant):IsTone2Day2	308.025	405.000	<.001*
s(Time, Participant):IsTone4	244.687	405.000	<.001*
s(Time, Participant):IsTone4Day2	266.791	405.000	<.001*
s(Time, Participant):IsGnrlztnDay2Tone2	255.425	388.000	<.001*
s(Time, Item)	142.311	179.000	<.001*
s(Time, Item):IsDay2Priming	95.120	177.000	<.001*

Number of obs.: 708,132; Participants: 81; Items: 36

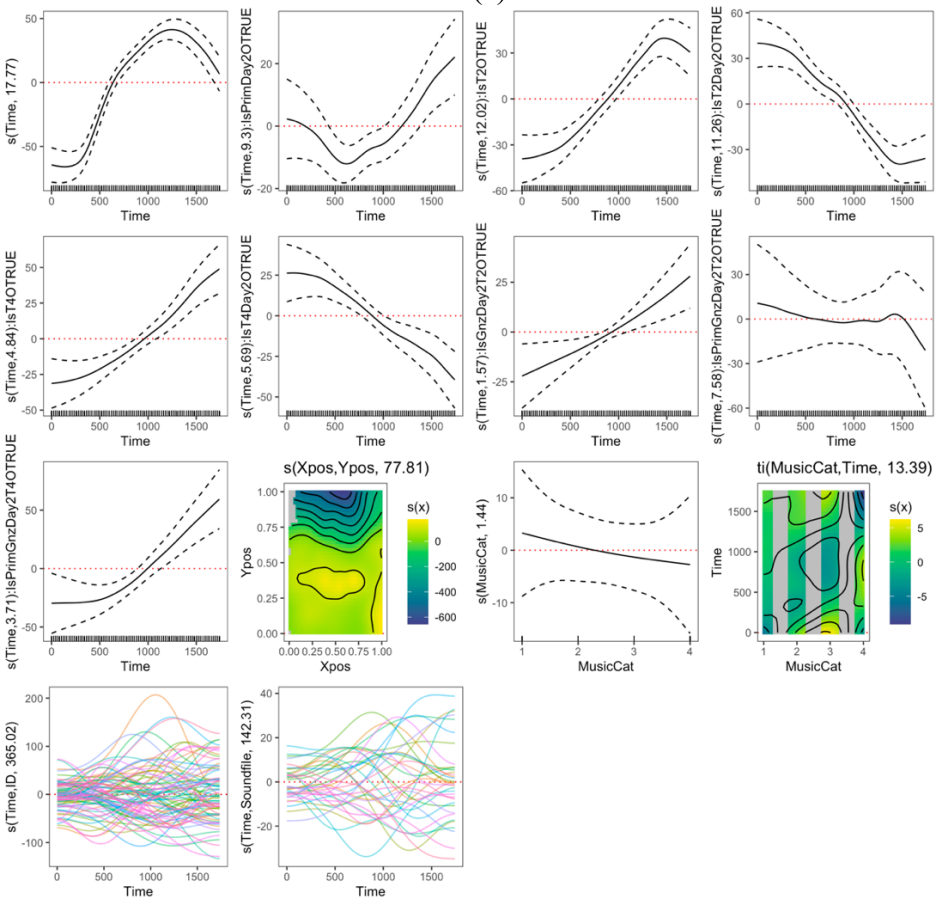
These pupillometry results are visually represented in Figure 9. In Figure 9a, the intercept differences are plotting, representing an overall up or down shift in the smooth terms represented in Figure 9b. In Figure 9b, the top left smooth is the reference smooth, and smooths two through

nine are the fixed difference smooths compared to that reference smooth. The tenth, two-dimensional smooth controls for eye position on the screen, and smooths 11 and 12 are covariate terms for musicianship that remained significant. Plots 13 and 14 are the random smooths by participant and by item. For better ease of interpretation, these intercept differences and difference smooths from Figure 9 are added together below to show estimated complete effects for specific conditions of interest.

Figure 9. Modeled pupil dilation results split by taVNS group, tone, and stimulus condition with reference levels of Sham/Peristim, Tone 1, Day 1, Critical: (a) Predicted pupillometry parametric effects for GAMM of phonological categorization pupil response for accurate responses at pre and post. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. (b) Predicted pupillometry fixed smooth effects for GAMM of phonological categorization pupil response for accurate responses at pre and post. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point.


f(IsPrimDay2O)
IsPrimDay2O
f(IsT2O)
IsT2O
f(IsT2Day2O)
IsT2Day2O
f(IsT4O)
IsT4O
f(IsT4Day2O)
IsT4Day2O
f(IsGnzDay2T2O)
IsGnzDay2T2O
f(IsPrimGnzDay2T2O)
IsPrimGnzDay2T2O
f(IsPrimGnzDay2T4O)
IsPrimGnzDay2T4O
FALSE
TRUE

(a)


s(Time, 17.77)
s(Time,9.3):IsPrimDay2OTRUE
s(Time,12.02):IsT2OTRUE
s(Time,11.26):IsT2Day2OTRUE
s(Time,4.84):IsT4OTRUE
s(Time,5.69):IsT4Day2OTRUE
s(Time,1.57):IsGnzDay2T2OTRUE
s(Time,7.58):IsPrimGnzDay2T2OTRUE
s(Time,3.71):IsPrimGnzDay2T4OTRUE
s(Xpos,Ypos, 77.81)
s(MusicCat, 1.44)
ti(MusicCat,Time, 13.39)
s(Time,ID, 365.02)
s(Time,Soundfile, 142.31)
Time
Xpos
Ypos
MusicCat
s(x)

(b)

Before any taVNS for the stimulation groups or any tone training (Figure 10), on Day 1 we see that Tone 2 (*Est*. = 38.279, *SE* = 6.464, $p < .001$) and Tone 4 (*Est*. = 29.503, *SE* = 7.541, $p < .001$) elicited overall significantly higher pupil responses than Tone 1, indicating larger processing effort. Additionally, in their respective smooth terms, we see parallel differences such that Tones 2 (s(Time):IsTone2, edf = 12.024, $p < .001$) and 4 (s(Time):IsTone4, edf = 4.839, $p < .001$) both show a smaller pupil response compared to Tone 1 at the beginning of a trial and a larger response toward the end of a trial. These combined parametric and smooth effects are shown in Figure 10 and show a larger pupil response for Tones 2 and 4 in the later part of the trial, showing larger, more sustained effort compared to Tone 1.

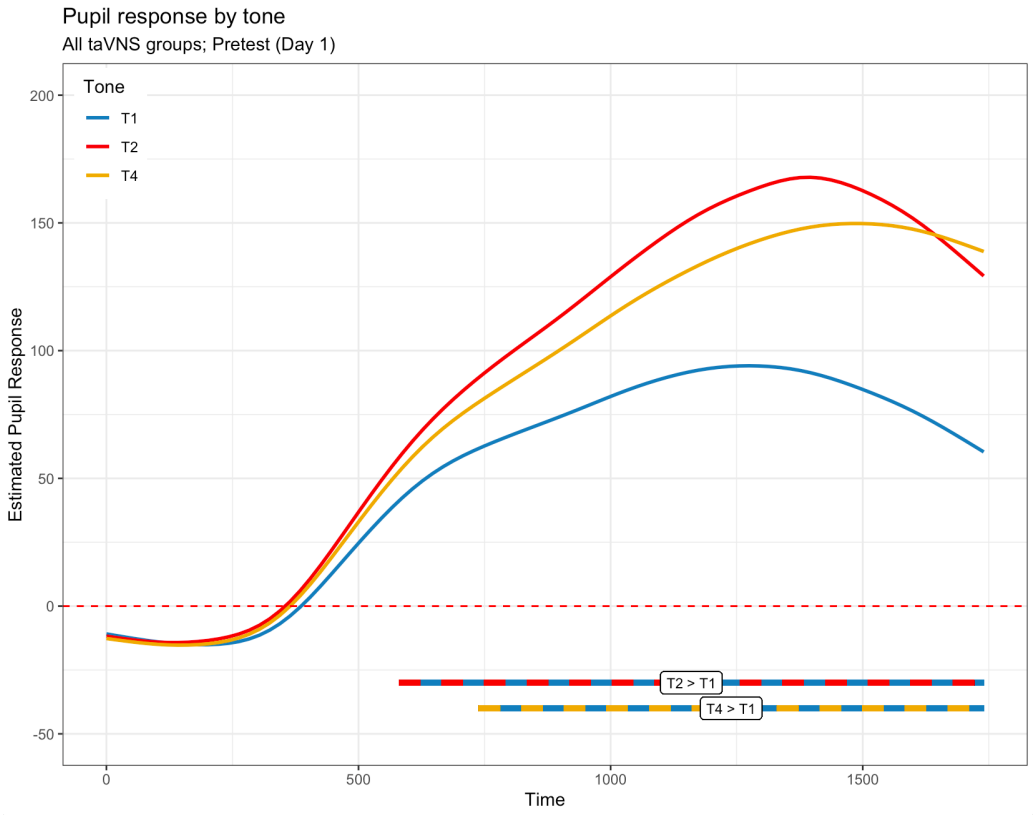


Figure 10. Estimated pupil responses for Tones 1, 2, and 4 on Day 1 (pretest) for all participants. Horizontal lines show locations of significant differences where T2 and T4 have a larger pupil response from T1.

It can be seen that the pupil response for the priming group is significantly lower overall from Day 1 to Day 2 than the peristim and sham groups (*Est.* = -15.207, *SE* = 5.489, $p = .006$), and there is also a significant smooth difference for priming vs sham and peristim day to day (edf = 9.296, $p < .001$) such that there is a lower pupil response toward the middle of the trial and a larger response toward the end. Additionally, there are significant parametric and smooth terms for Tones 2 (*Est.* = -35.497, *SE* = 6.237, $p < .001$; edf = 11.264, $p < .001$) and 4 (*Est.* = -16.148, *SE* = 7.029, $p = .022$; edf = 5.694, $p < .001$) on Day 2 that are a mirror image of the effects on

Day 1. Taken together, these effects for Critical items on Day 2 are represented in Figure 11. The larger pupil response for Tones 2 and 4 compared to Tone 1 on Day 1 was heavily reduced on Day 2 back toward the lower Tone 1 response such that pupil response for Tone 2 essentially overlaps with Tone 1 and Tone 4 still shows a little larger response toward the end of the trial. The overall significant effects for the priming taVNS group results in a reduced pupil response for all Tones compared to the Sham and Peristim groups, from about 100 ms into the trial up until Priming overlaps with the other taVNS groups again at about 1,200 ms.

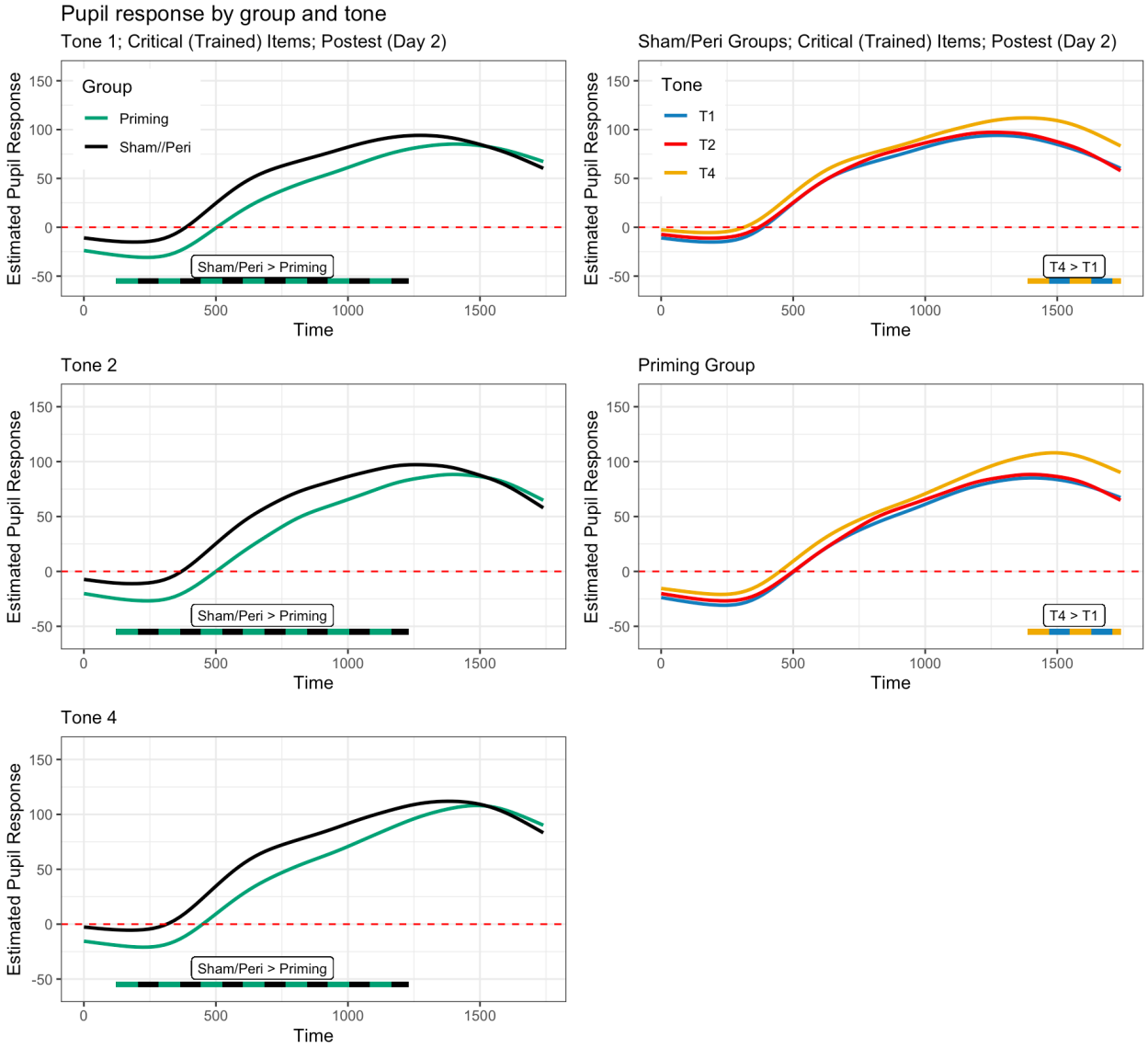


Figure 11. Estimated pupil responses for Tones 1, 2, and 4 on Day 2 (posttest) for peristim and sham taVNS (no difference between groups) and priming taVNS. Horizontal line shows location of significant differences of a smaller pupil response across all tones for Priming compared to Sham and Peristim.

Regarding Generalization items that weren't trained in the two-day study, there are no significant parametric differences from Critical items, but there are smooth differences for Tone 2 for Sham and Peristim (edf = 4.576, p = .004) with an additional effect for Priming (edf = 7.579, p = .004), and a smooth difference for Tone 4 only for Priming (edf = 3.705, p < .001). Taken together, these results are combined in Figure 12. For Sham and Peristim, the pupil response for Tone 2 Generalization items is a little more effortful than for Critical items, patterning more similarly here to Tone 4 items than Tone 1 at the end of the trial. For Priming, the small pupil response for Tone 1 Critical items is maintained for Tone 1 and Tone 4 Generalization items, and there is now a larger pupil response after about 1,300 ms for Tone 4 compared to Sham and Peristim.

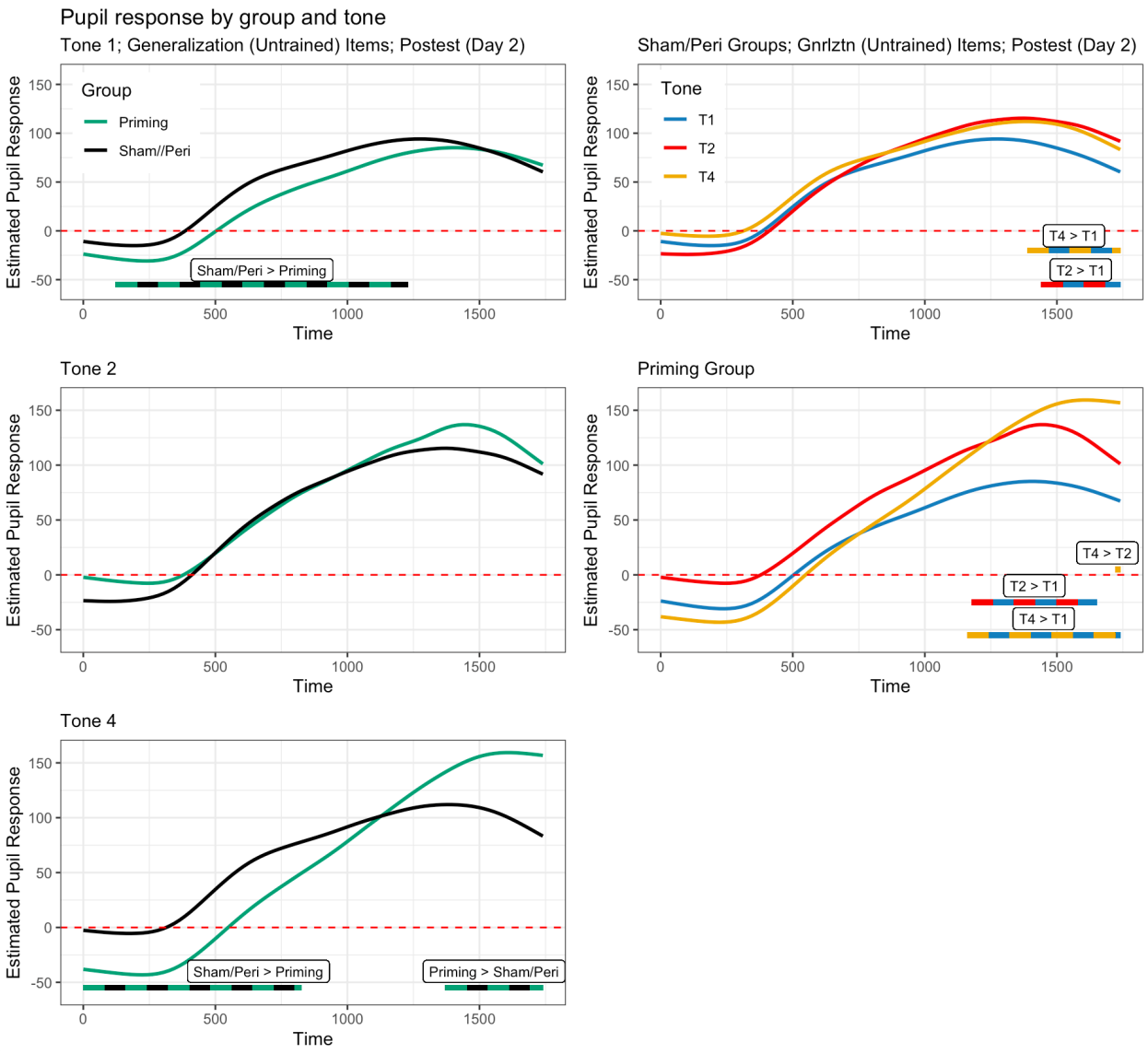


Figure 12. Estimated pupil responses for generalization items for Tones 1, 2, and 4 on Day 2 (posttest) for peristim and sham taVNS (no difference between groups) and priming taVNS. Horizontal lines at bottom indicate significant differences between taVNS groups for individual tones; line color indicates the group with the larger pupil response. Horizontal lines at top indicate significant differences within-group, between Tone.

5.3 Discussion

Two different taVNS interventions were again observed to elicit performance improvements over a sham control in a double-blind study of learning novel tonal contrasts. In order to answer the research questions, relative tone difficulty between tones 1, 2, and 4 first needed to be established. Independent of group differences, looking at pretest performance especially and taking into account all of accuracy, reaction time, and pupillometry metrics, it's clear that, in this corpus, T4 is the most difficult tone; T2 is next, with T1 being easiest. This conclusion is based on the following results. T4 accuracy at pretest was around 15-25%, below chance, indicating that responses to T1 and/or T2 were over-endorsed, while T1 and T2 were above 80% on modeled accuracy. T4 reaction times at pretest were the slowest and T2 reaction times at posttest were the fastest, indicating the least acoustic information was needed to make a correct response for T2. This tracks from previous research (e.g., Maddox et al., 2013) indicating that learners of tone previously naïve to tonal languages will pay most attention to initial height differences in the tonal contour, which makes T2 easiest to distinguish early on as T1 and T4 have more similar starting heights. Finally, from the pupillometry results, there is a very clear and robust effect at pretest for all groups to show that T2 and T4 equally showed larger cognitive effort over the course of the trial versus T1, suggesting greater processing difficulty for items responded correctly to. Interestingly, at posttest both T2 and T4 curves flattened toward the T1 curve, showing less effort after training. T2 was no different from T1 at post, but T4 still showed a little more sustained effort vs T1, further lending evidence to the T4 > T2 > T1 difficulty scale observed with the behavioral data. This difficulty scale is contrary to Pelzl's (2019a) review of the literature where he concluded that T2 is consistently found to be more difficult than T4 in

isolated syllables. This discrepancy may be due to the lack of T3 in the present study, given that T3 has a similar starting position to T2. Additionally, when looking at specific tone pairings, T1 and T4 have been found to be more confusable than T1 and T2 (Hao, 2012; Wang et al., 1999).

To answer the research question on behavioral learning outcomes by tone difficulty, peristim taVNS showed an advantage over sham in phonological categorization accuracy improvement from pretest to posttest on both trained and untrained items by about 10-15% for T4, the hardest-to-learn tone in this design. Peristim was also the only group to show significant improvement in accuracy for T2 (about 6%) from pre to post. Priming taVNS, on the other hand, curiously showed less improvement in phonological categorization accuracy from pre to post than sham (about a 5-10% difference). However, it's worth noting that priming and sham performance at post was not significantly different, and this change in improvement is likely reflecting descriptively different starting places at pretest before any stimulation (descriptively, not significantly, priming had better accuracy on T4 at pre than peristim and sham). For phonological categorization RTs, peristim showed marginally faster RTs than sham (about 70 ms difference) for T4 pre to post while priming showed significantly faster RTs vs sham (about 100 ms difference).

These phonological categorization results already show interesting parallels to the results observed in the lexical recognition task in Chapter 4. Namely, peristim is showing comparable accuracy improvements over sham (10-15% here, 5-10% in Chapter 4) and only a marginal improvement in RT for the hardest tone condition here over sham compared to no RT improvements vs sham in Chapter 4. Likewise, there is no notable accuracy difference between priming and sham here or in Chapter 4, but there is a clear advantage for priming over sham in

terms of RT improvements by about 100 ms for the hardest tone compared to about a 100 ms advantage for priming over sham in the lexical recognition test in Chapter 4. Already these results suggest that, not only does active taVNS improve behavioral learning outcomes at both the phonological and lexical levels, but also that effects may be driven specifically by improvements on the most difficult items.

To answer the research question on behavioral learning outcomes for untrained items, across the board accuracy improvements were lessened compared to trained items but taVNS group differences for T4 were maintained. The significant accuracy improvement for peristim on T2 disappeared. Interestingly, the sham group performed significantly worse pre to post on accuracy for T1 untrained items by almost 10%, which the active taVNS groups avoided. On reaction time, there were no improvements for any group pre to post for T4 untrained items. For untrained T2, again there were no improvements for sham or priming, but peristim significantly sped up in the same magnitude as it did for trained T2 items, about 75 ms. Curiously, for untrained T1, sham had a significant slowdown in RT by about 75 ms to match its worse accuracy effect. Priming had a marginal slowdown for untrained T1 of about 50 ms, while peristim showed no change pre to post.

These generalization results for phonological categorization show that, while untrained items are overall more difficult, the effects of taVNS also persist to items with untrained new speakers and syllables. The pattern for accuracy is largely maintained, while for RT there's a slightly different story. Despite the fact that priming more clearly enhanced RT improvements for lexical recognition and phonological recognition for trained items, peristim is the only group to show any clear improvements for untrained RT effects. The interpretation from the discussion

of Chapter 4 still holds relevant here, that priming taVNS's primary mechanism of action may indeed be enhanced access to already learned information while peristim taVNS's primary mechanism of action may be better encoding of new information in the moment. Thus, we would see this clearer advantage in the speedup of trained items for priming but peristim for untrained: peristim appeared to facilitate better encoding of T2, and also possibly T1 given that peristim showed the least evidence of a slowdown from pre to post. There was no advantage for peristim on T4, but it is worth noting that there were a few *a priori* group differences on RT at pretest that survived to posttest, the largest being that, at pretest for T4 untrained items, sham was the slowest and peristim was the fastest.

Together, the results for the impact of taVNS moderated by tone difficulty and moderated by generalization items expand, but also partially conflict with, the findings from Llanos et al. (2020). First, it is worth noting that the paradigm in Llanos et al. (2020) appears overall more difficult than the present corpus as the ultimate outcome for percent correct on trained items in their study for the control group was about 55% while here the sham group reached about 75% accuracy on trained items at post. They found overall benefits in accuracy for peristim taVNS only when taVNS was paired with easier tones (T1 and T3). The overall peristim improvement on accuracy for trained and untrained items was about the same magnitude as observed in this Chapter, roughly a 12.5% improvement compared to their control compared to the 10-15% improvement found here, although this magnitude of improvement was found here for the most difficult tone, T4. While Llanos et al. (2020) only found improvements for both easy and hard tones when taVNS was paired with easy tones, they did not examine which specific tones benefited the most, and this study paired peristim with all trained tones, and found the largest

measurable benefits for harder tones, which had more room to improve. Because the training paradigm in this study was easier (either because of fewer trained stimuli or because there was an additional day of training), it's possible that the easier tones in Llanos et al. (2020) had more room to show improvement and thus influence from taVNS versus this corpus. Given that taVNS paired to easy tones in their study showed improvements for all tones, it's also possible that the peristim with the easier tones facilitated better encoding of the easier tones, in turn making it easier to distinguish the harder tones by contrast. Additionally, this dissertation expands upon Llanos et al. (2020) by also examining RTs and showing a unique benefit of peristim over sham for untrained items. Even if peristim taVNS improvement was found after a slightly different type of administration, the fact that the accuracy improved at similar magnitudes is also interesting in light of the fact that Llanos et al. (2020) applied taVNS to the cymba conchae of the outer ear while this corpus applied taVNS to the inner ear, showing a potential generalizability for tVNS stimulation locations.

To answer the research question on deployment of cognitive effort by tone difficulty, we see no difference between peristim and sham taVNS groups which resulted in peristim falling out of the model. However, we do see a significant effect of priming taVNS vs peristim and sham, an equal effect across T1, T2, and T4 that priming showed less cognitive effort from about 100 ms up till about 1,200 ms in a trial. Thus, while there were overall differences in effort by tone and by priming compared to peristim and sham, these differences in effort did not interact for trained items.

To answer the research question on deployment of cognitive effort for untrained items, we do see some group differences across tones. The effort deployed for untrained T1 items is the

same as for trained items, and so that priming group difference also persists. For T2, there are no longer significant group differences but all groups have shifted up to a slightly higher peak and sustained response compared to T1 and untrained items. T4 has a similar uptick as T2 does for sham and peristim, but for priming the curve is shifted such that priming shows even less effort compared to peristim and sham than we observed for trained items, but there is a steeper increase in effort at the end of the trial such that priming shows a significantly higher peak through the end of the trial compared to sham and peristim. This result is counterintuitive; a possible explanation could be that this mirrors the accuracy results for untrained T4, where priming improved significantly less pre to post than both peristim and sham. However, as noted above, this seems unlikely as accuracy at posttest (rather than looking at improvement) was not significantly different from sham.

Contrary to my initial hypothesis, While Chapter 4 showed a limited effect of priming taVNS on the TEPR but a strong effect of peristim taVNS, for phonological categorization we see no effect of peristim compared to sham, and instead largely a benefit for the priming group. One explanation is a difference in task demands as Chapter 4 looked at a passive word learning task with pupillometry in which no behavioral responses were elicited while here we are examining a phonological recognition pretest and posttest. Another explanation is the difference in GAMM modeling techniques that were employed. While Chapter 4 only found and presented group differences from day-to-day, this Chapter split the stimuli by tone and used a more sophisticated modeling method that allows a better examination of nuance in the results. Further, in light of the diverging results for the untrained items here, where the results show a lower effect for priming for T1, no difference for T2, and ultimately a higher effect for T4, it's possible

that splitting the passive word learning task by tone could have shown nuance in the priming group's effects by tone that were erased when averaging results across tonal contour. One final important difference between the tasks is that the passive word learning task was immediately preceded by taVNS priming for that group and peristim participants received peristim stimulation before every trial in the passive word learning task. Given the nature of the phonological categorization pretest and posttest, it was neither immediately preceded by priming taVNS nor did it have peristim stimulation at every trial. Thus, it's possible that peristim differences in cognitive effort only arise in the context of active stimulation with behavioral improvements persisting even when the stimulation is not immediately present.

In total, the results in this Chapter show positive effects for both peristim and priming taVNS compared to sham taVNS, although in different ways, and greatly build upon the results from Llanos et al. (2020). We again largely see a separation in behavioral outcomes: peristim best improving accuracy while priming better improves reaction time compared to sham. We do also get new insight with the inclusion of generalization items in this phonological categorization paradigm. For items with untrained speakers and syllables, we see reaction time benefits with peristim only. We also see overall less effort deployed by the priming group for trained items, which may be a sustained effect of taVNS priming as, by this posttest on day 2, the participants had received three 10-minute stimulation intervals within a two-hour period.

5.4 Conclusions and Next Steps

The current Chapter's results expand the scope of those in Chapter 4. After a two-day training, active taVNS had positive effects not only at the lexical learning level as we saw in

Chapter 4, but, as we see in this Chapter, also earlier on the phonological learning of tone for learners naïve to lexical tone. Thus, both peristim and priming active taVNS administrations result in early benefits to phonological tone learning that persist into lexical tone learning.

Further, the tones most strongly enhanced by these active taVNS interventions were the more difficult ones, and behavioral advantages for categorizing tones on untrained items persisted for peristim only. Only priming showed a differential deployment of cognitive effort during the posttest suggesting that the mechanism of action for priming enhancing phonological categorization is related to a task-evoked early reduction of cognitive effort, regardless of tone difficulty.

Because total stimulation over the training days for priming greatly outweighed peristim, this lends further credibility to the conclusion that the total amount of taVNS stimulation given is less relevant than the nature of the stimulation administration related to the information to be learned.

Unlike the lexical recognition and recall tasks in Chapter 4, the phonological categorization test in Chapter 5 had a true pretest as a baseline to ultimate outcomes, which lends further credibility to the conclusions being able to control for differences before administering taVNS. It also had a true posttest without peristim during the posttest or priming immediately before, meaning the positive behavioral effects for both peristim and priming taVNS aren't simply due to either tightly adjacent or simultaneous stimulation administration.

The results in this Chapter on their own are suggestive of some practical considerations for eventual taVNS use outside the laboratory. Since peristim appears to most directly enhance learning accuracy as well as being able to transfer benefits more easily to related, but untrained

items, peristim is emerging as the winner for most practical benefits gained during learning between peristim and priming taVNS administrations.

However, given that the results here are yet further in line with the conclusions of Chapter 4 that peristim may enhance encoding of new information while priming may enhance retrieval of learned information, future research would benefit from a paradigm that tests a taVNS condition that employs *both* priming and peristimulus stimulation compared to only using one or the other to see if the mechanisms of action are complementary, and to investigate whether learners could see the benefit of both improved encoding and retrieval for Mandarin tone learning.

Additionally, in seeing positive outcomes for both conditions at posttest wherein neither did the participants receive peristim during the test nor priming immediately before, these initial findings are suggestive of possible longer lasting benefit of both priming and peristim administrations. Since it is entirely possible that the results were influenced by recent active taVNS even if it wasn't for the phonological test itself, additional research needs to be conducted with delayed posttests on days in which no taVNS is administered to further examine the duration of these positive taVNS effects. Future studies in this vein should also control groups for pretest scores on phonological categorization to avoid murkiness of different pretest starting places, as is did make some of the results here less clear.

While this Chapter helps clarify the effects of taVNS on Mandarin tone learning, there is still much to explore. I have shown that taVNS enhances learning not only at the lexical level, but also earlier at the phonological level; I have also shown that it is maximally effective for more difficult tones. One potential result of this research is to promote taVNS early in Mandarin

tone learning to reduce classroom time on low-level language features like L2 phonology and rote vocabulary learning so more classroom time can be spent on higher level linguistic features, optimizing the route to increased language proficiency. While this Chapter explored how taVNS impacts learning for different features of the input (difficulty and untrained sounds and speakers) while controlling for the potential covariates of non-linguistic tone aptitude and musicality, another important avenue to maximize the effectiveness of taVNS is to explore whether taVNS differentially impacts learners with different levels of preexisting ability. Are there specific populations of learners that may benefit the most—or not at all—from taVNS?

Chapter 6: taVNS-facilitated Language Learning Moderated by Individual Differences

6.1 Introduction and Motivation

The results from Chapter 5 revealed practical utility of taVNS for difficult tasks like lexical tone learning for not only the learning of difficult contrasts in word contexts but also lower-level phonological tone categorization learning. An additional step toward optimizing a taVNS training intervention is to tease apart factors that may modulate taVNS efficacy. Are taVNS-facilitated outcomes modulated by established predictors of tone learning? Could taVNS only facilitate tone learning for those with a low baseline predisposition for learning these contrasts? Or, in contrast, could taVNS only show efficacy for those already predisposed toward successful tone learning?

As has been reviewed in Chapter 2, acquiring lexical tone as a feature for non-native learners still remains difficult, with wide variability in learning trajectory even after as many as 18 training sessions (Bowles et al., 2016; Chandrasekaran et al., 2010; Li & DeKeyser, 2017; 2019; Liu & Chandrasekaran, 2013; Wang et al., 1999; Wang et al., 2003; Wong et al., 2011; Wong & Perrachione, 2007). While a standard individual differences approach does help explain the origins of this variability in ultimate performance in multisession training studies, it does not directly speak to what types of interventions can help those with lower aptitude levels in those areas, such as tone aptitude, overcome their obstacles in acquiring tone as a lexical feature. As an extension of Ingvalson et al. (2011)'s recommendation for more traditional aptitude-by-treatment interaction (ATI) studies, this Chapter explores another type of treatment to make language

learning more accessible: neurostimulation, which would not necessarily require modifying the training or input that learners receive in order to enhance learning.

Non-linguistic tone aptitude and musicality are uniquely positioned for an ATI study with tVNS on Mandarin tone learning, given the strong, established relationship between these variables and tone learning outcomes. It is worth noting that many studies exploring the effects of musicality and tone learning have specifically geared their data collection to include a representative sample from music departments (e.g., Bowles et al., 2016), but the corpus used for this dissertation chose to collect data from the general population, in which musicians are less well-represented. This choice was made in light of the fact that taVNS is a relatively new intervention, previously never applied to language learning in such a systematic way. Before investigating such a specific population, it was deemed more critical and informative to test these effects in as representative a sample of the general population as possible. While this has undoubtedly limited the range in music experience scores and will be an inherent limitation of that measure in this corpus, music aptitude still shows variation in populations where music experience is intentionally restricted (e.g., Li & DeKeyser, 2017).

A few tone word learning studies have included measures of both tone aptitude and musicality (e.g., Bowles et al., 2016; Cooper & Wang, 2012), and have found tone aptitude to play a larger role than musicality in predicting learning outcomes. Additionally, results from Bowles et al. (2016) and Wong & Perrachione (2007) show support for an early role for musicality in learning lexical tone contrasts while tone aptitude appears to show a larger role, both predicting early learning and predicting performance when learners are given generalization stimuli with new, untrained talkers.

The present Chapter analyzes data from a Mandarin lexical tone training paradigm, for which stimulation and sham control groups were *a priori* balanced on non-linguistic tone aptitude and music experience, and investigates whether individual differences modulate the efficacy of transcutaneous auricular vagus nerve stimulation (taVNS) on lexical tone learning on phonological and lexical outcomes.

6.1.1 Research questions

The study in this chapter uses taVNS and investigates its effects on Mandarin tone learning across multiple outcome measures that are sensitive to changes at varying timescales (across sessions, across trials, and across milliseconds). Priming taVNS and peristim tVNS will again be contrasted with sham taVNS using data collected in the corpus: a phonological tone categorization test (3.2.3.5) as a pretest (prior to any taVNS or training) and posttest (after all taVNS and training on Day 2) and a lexical recognition test (3.2.3.7) performed at the end of training on both Day 1 and Day 2. The individual differences in this study were also collected as part of the larger corpus: the pitch contour identification task (3.2.3.2), self-rated musicianship (3.2.3.3), and the Wing music aptitude test (3.2.3.4). Behavioral outcomes of accuracy and reaction time on phonological tone categorization and lexical recognition tests around a two-day phonological and lexical tone training will be analyzed as indices of learning phonological tone and lexical tone, respectively. Pupillometry was collected during the tests and analyzed as an index of cognitive effort. Critically, this Chapter will take an aptitude-by-treatment interaction (ATI) approach to investigate whether taVNS is more or less effective at varying levels of tone

aptitude and/or musicality. If the results of Chapter 5 show tone-specific learning effects, pretest and posttest outcomes will be analyzed for T1, T2, and T4 separately for both the phonological tone categorization and lexical recognition tests. The research questions of this study, building on top of the research questions in Chapters 4 and 5, are as follows:

5) Are effects of (priming and/or peristim) taVNS-facilitated learning moderated by individual differences previously known to affect Mandarin tone learning?
    - Non-linguistic tone aptitude
    - Musicality
6) Do individual differences moderate effects of active (priming and/or peristim) taVNS versus sham taVNS in eliciting a differential deployment of cognitive effort during phonological tone processing and/or lexical tone processing?
    - Non-linguistic tone aptitude
    - Musicality

Given the mechanism of action of taVNS, it may indeed be the case that the efficacy of taVNS may depend on individual differences in the learner. For example, if taVNS enhances attention and memory consolidation, facilitating language learning, perhaps learners with already high baseline levels of aptitude show little to no improvement with taVNS, whereas those with lower levels show more marked improvement. Thus, taVNS as a technique may help compensate for lower aptitude in the pursuit of language learning. Conversely, it may also be possible that "the rich get richer," that those with already high baseline levels of aptitude benefit more from stimulation. The third possibility is that there is no interaction, and taVNS enhances Mandarin tone learning equally for both, independent of aptitude and experience. This Chapter is

exploratory and all three hypotheses given equal weight as, to date, no known ATI studies have been conducted with VNS broadly, let alone any exploring interactions of tVNS treatment with non-linguistic tone aptitude or musicality for Mandarin tone learning. Further, no prior studies could be found investigating the role of these individual differences on tone-learning pupillometric outcomes. Thus, while previous research suggests non-linguistic tone aptitude will have an effect on both accuracy and RTs and musicality only on accuracy, it is less certain to what degree these factors may affect pupil dilation.

6.2 Results

6.2.1 taVNS moderates effects of individual differences on behavioral tone learning

A total of 72 participants (51 female) ages 18–34 years (M = 21.54, SD = 3.06) were analyzed, after 11 were excluded for missing data, noncompliance throughout learning tasks, and/or incorrectly interpreting task instructions for at least one of the tasks. There were 19 participants in the priming taVNS group, 18 participants in the peristim taVNS group, and 35 participants in the sham taVNS group. Descriptives for the behavioral tests and individual differences in this chapter are in Table 11. The IDs correlated with each other moderately, but not so highly that the variance they explained was so overlapping: non-linguistic tone aptitude (PCID) correlated with music aptitude (Wing) r = .48, PCID with music experience (MCat) r = .38, and Wing with MCat r = .35. Participants did not vary across taVNS group on PCID (linear regression model comparison p = .953) and MCat (ordinal model comparison p = .455), which

were the two variables *a priori* balanced between groups, and neither did they vary on Wing (linear regression model comparison $p = .161$).

Table 11. taVNS group descriptive statistics for the phonological categorization test and lexical recognition test.

	Peristimulus (n = 18)		**Priming (n = 19)**		**Sham (n = 35)**	
	Pre Day 1	***Post Day 2***	***Pre Day 1***	***Post Day 2***	***Pre Day 1***	***Post Day 2***
Accuracy: % correct (SD)						
Critical	60.9	81.9	62.7	77.9	57.1	73.9
(Phon. Cat.)	(18.3)	(14.3)	(18.7)	(17.2)	(16.5)	(18.2)
Generalization	48.9	62.7	55.2	60.6	48.1	54.9
(Phon. Cat.)	(15.0)	(13.4)	(19.2)	(17.8)	(14.0)	(19.2)
Match	73.0	83.8	77.6	89.2	69.8	82.5
(Lex. Recog.)	(14.3)	(10.9)	(12.6)	(8.5)	(12.8)	(14.6)
Mismatch	65.5	76.9	60.3	78.0	55.2	71.5
(Lex. Recog.)	(18.5)	(17.9)	(18.6)	(19.6)	(17.5)	(20.0)
Reaction Time: ms (SD)						
Critical	901	864	939	888	974	964
(Phon. Cat.)	(157)	(159)	(119)	(181)	(148)	(178)
Generalization	877	886	895	941	950	1002
(Phon. Cat.)	(157)	(169)	(103)	(146)	(142)	(158)
Match	929	849	962	844	939	892
(Lex. Recog.)	(113)	(140)	(124)	(116)	(155)	(141)
Mismatch	1000	918	1077	939	1029	985
(Lex. Recog.)	(122)	(139)	(100)	(121)	(159)	(150)
Individual Differences						
PCID Accuracy % correct (SD)	67.5 (10.0)		67.8 (11.5)		66.9 (9.2)	
Wing Accuracy % correct (SD)	59.3 (17.1)		53.9 (18.8)		49.0 (18.7)	
Musicianship Category (MCat frequency)	1 (6); 2 (6); 3 (4); 4 (2)		1 (2); 2 (10); 3 (4); 4 (3)		1 (9); 2 (16); 3 (7); 4 (3)	

Note. MCat possible responses were: 1 = non-musician, 2 = music-loving non-musician, 3 = amateur musician, 4 = serious amateur musician, 5 = semiprofessional musician, and 6 = professional musician; there were no participants in categories 5 or 6 in this dataset.

As for Chapters 4 and 5, binomial logistic and linear mixed-effects models were used to analyze accuracy and reaction time data for the phonological categorization test and the lexical recognition test, respectively. All MEMs began with the IVs from Chapters 4 and 5: for Session (Day 1 vs. Day 2), taVNS Group (Priming vs. Peristim vs. Sham), and Condition (Match vs Mismatch for lexical recognition test; Critical (trained) Stimuli vs. Generalization (untrained) Stimuli for phonological categorization test). The main individual differences (IDs) of interest were non-linguistic tone aptitude (PCID) and musicality, represented by music experience (self-rated musicianship category, MCat) and music aptitude (Wing). Each ID was tested with the lme4 (Bates et al., 2015) package in R (R Core Team, 2019) using buildmer (Voeten, 2019) by adding the potential for the simple effect of the variable and its potential interactions with the other listed IVs. The models contained the potential of each of the three IVs, such that each was controlling for the variance of the other. Final models of best fit are reported below. Given the number of models and effects that significantly improved model fit, as well as priori results in Chapters 4 and 5 that averaged over the effects of the IDs, the below description and discussion of results will focus specifically on the interaction effects (or lack thereof) of individual differences and taVNS group, which are the focus of the research questions in this Chapter. A further note for the interpretation of the effects of ID: PCID and Wing were continuous variables that are z-scored for all analyses and graphs. In this way, we can directly compare the effect sizes of PCID and Wing as they are on the same scale: a value of zero reflects an average score of PCID or Wing for this sample, -1 reflects one standard deviation below the average score of PCID or Wing for the sample, 2 reflects two standard deviations above, and so on. For MCat, being an ordinal variable with only four levels in this sample (1 = non-musician, 2 = music-

loving non-musician, 3 = amateur musician, 4 = serious amateur musician), it was treated as continuous and was centered at level 2, which was the most common response for this sample (thus: -1 = non-musician, 0 = music-loving non-musician, 1 = amateur musician, 2 = serious amateur musician for all the models and plots below).

The accuracy results for the phonological categorization test are shown in Table 12. In short, the picture is complicated. Four-way interactions of Stimulus Condition X Individual Difference X Day X taVNS Group survived for all three IDs. Thus, there were ATI effects (interactions of ID and taVNS Group) for all three IDs. There are situations in which all three IDs, PCID, MCat, and Wing positively predicted phonological categorization accuracy by a large magnitude, the biggest effect being for PCID on Day 1 predicting a jump from 30% at the low end of PCID to 90% on the high end across groups. By the same token, there are also situations reflected by the interactions in which all three IDs showed an essentially flat line, or null effect of the ID.

Table 12. Logistic MEM for phonological categorization test accuracy moderated by IDs.

Fixed effects	**Estimate**	***SE***	***p***
Intercept (Critical, Day 1, Sham)	0.510	0.443	.249
Training Day (Day 2)	1.045	0.340	.002*
taVNS Group (Peristim)	0.077	0.306	.802
taVNS Group (Priming)	0.296	0.312	.343
Day 2 X Peristim	0.203	0.236	.389
Day 2 X Priming	-0.308	0.247	.212
PCID (Non-Ling. Tone Apt.)	0.567	0.213	.008*
PCID X Day 2	-0.152	0.144	.290
PCID X Peristim	-0.028	0.346	.935
PCID X Priming	0.289	0.303	.342
PCID X Day 2 X Peristim	0.237	0.251	.346
PCID X Day 2 X Priming	-0.058	0.218	.790
Wing (Music Apt.)	0.550	0.196	.005*

Wing X Day 2	0.030	0.144	.833
Wing X Peristim	-0.411	0.359	.251
Wing X Priming	-0.250	0.311	.421
Wing X Day 2 X Peristim	0.004	0.266	.988
Wing X Day 2 X Priming	-0.081	0.232	.728
MCat (Music Exp.)	0.093	0.198	.640
MCat X Day 2	-0.087	0.144	.548
MCat X Peristim	0.281	0.314	.371
MCat X Priming	-0.119	0.343	.728
MCat X Day 2 X Peristim	0.106	0.240	.659
MCat X Day 2 X Priming	0.470	0.257	.067^
Generalization	-0.697	0.590	.238
Generalization X Day 2	-0.474	0.464	.307
Generalization X Peristim	-0.181	0.228	.428
Generalization X Priming	0.090	0.241	.709
Generalization X Day 2 X Peristim	0.095	0.254	.709
Generalization X Day 2 X Priming	-0.048	0.278	.863
Generalization X PCID	-0.372	0.148	.012*
Generalization X PCID X Day 2	<0.001	0.145	>.999
Generalization X PCID X Peristim	-0.020	0.220	.928
Generalization X PCID X Priming	0.017	0.194	.930
Generalization X PCID X Day 2 X Peristim	0.049	0.254	.847
Generalization X PCID X Day 2 X Priming	0.100	0.220	.650
Generalization X Wing	-0.331	0.127	.009*
Generalization X Wing X Day 2	0.193	0.147	.188
Generalization X Wing X Peristim	0.488	0.231	.035*
Generalization X Wing X Priming	0.280	0.202	.166
Generalization X Wing X Day 2 X Peristim	-0.423	0.271	.118
Generalization X Wing X Day 2 X Priming	-0.252	0.237	.289
Generalization X MCat	0.140	0.129	.277
Generalization X MCat X Day 2	-0.208	0.148	.160
Generalization X MCat X Peristim	-0.116	0.204	.571
Generalization X MCat X Priming	0.113	0.223	.614
Generalization X MCat X Day 2 X Peristim	0.046	0.245	.850
Generalization X MCat X Day 2 X Priming	0.017	0.264	.948

Random effects	**Variance**	***SD***	**Correlation**		
Intercepts | Participant	0.737	0.859			
Day 2 | Participant	0.186	0.431	-.34		
Generalization | Participant	0.107	0.326	-.62	.40	
Intercepts | Item (presented sound)	3.042	1.744			
Day 2 | Item	1.794	1.340	-.94		
PCID | Item	0.043	0.209	.70	-.64	
Priming | Item	0.201	0.448	-.95	.92	-.76

Peristim | Item	0.135	0.367	-.42	.51	-.29	.45			
Day 2 X Priming | Item	0.258	0.508	.87	-.87	.92	-.89	-.53		
Day 2 X Peristim | Item	0.108	0.329	-.16	.04	.22	.05	-.49	.17	

Number of obs.: 15,541; Participants: 72; Items (unique presented sound files): 36

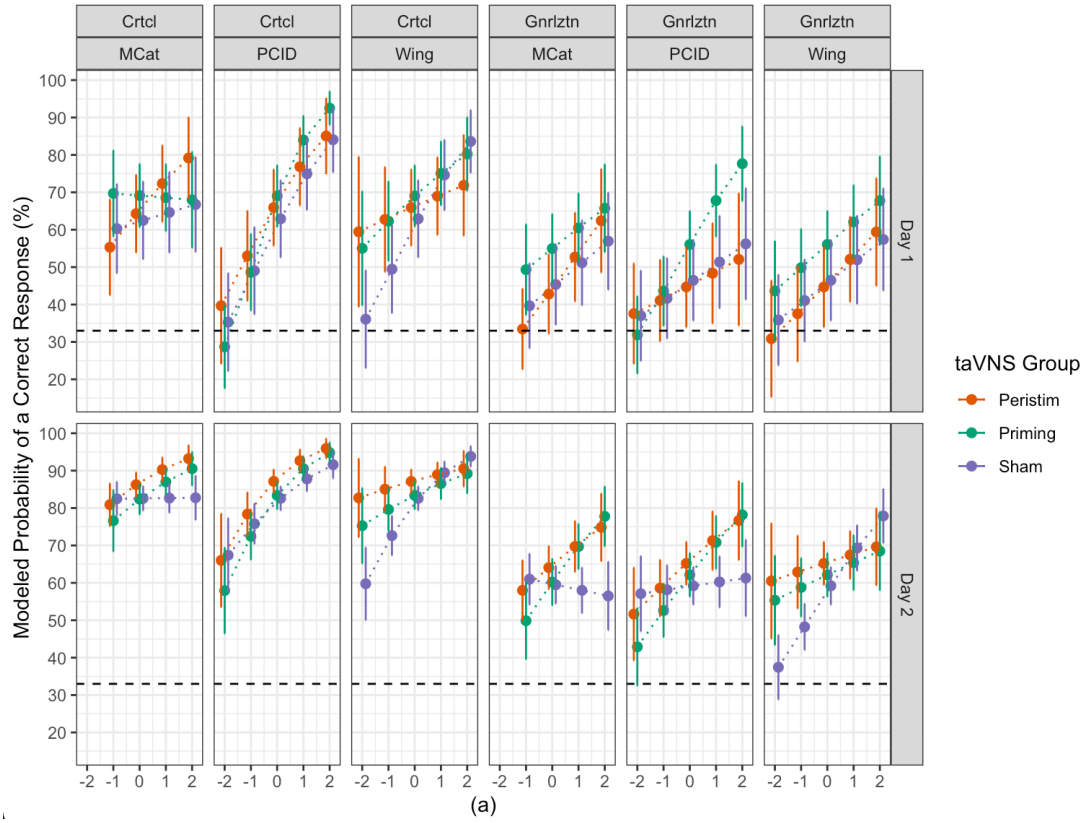
Given the complicated resulting model, the modeled estimates and standard errors are shown in Figure 13. In Figure 13a, these effects are split by day. However, given that day 1 for phonological categorization was a pretest before any stimulation, Figure 13b reflects the amount of improvement from pre to post for each group by each ID.

For Critical (trained) items, for PCID we see no aptitude-by-treatment interaction (ATI). All groups appear more or less to improve at the same rate in Figure 13b (those with the lowest ability improving the most from training), and there appears to be no meaningful separation of trajectories by group in Figure 13a. For MCat, we see that Peristim and Sham improved relatively similarly across MCat categories pre to post with some decline in improvement at higher MCat levels, but Priming differed in that those with low music experience improved less from Priming and those with more music experience improved more, although this may be to descriptively different starting points at pre. At post, there is some descriptive separation of groups to suggest that MCat is predictive for active taVNS groups but not predictive (a flat line) for Sham although this wasn't significant. For Wing, improvements from pre to post seem to parallel, but be more muted than, those for PCID. The main difference is that Peristim appears to main more equal improvement from pre to post across levels of Wing, whereas it declines more for Priming and Sham. Interestingly, it appears at post that the plot for Peristim is more or less a flat line in such a way that those with low music aptitude showed up to a 20% advantage

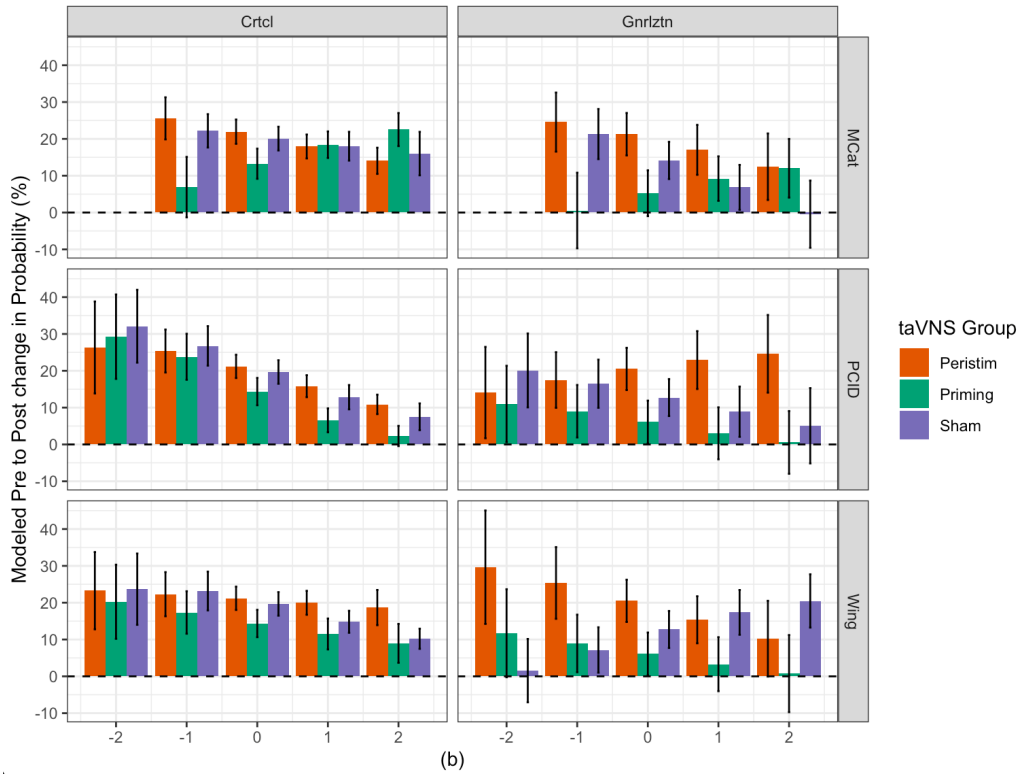
compared to those with low music aptitude in the Sham group. However, this may mostly be an artifact of starting differences at pretest.

For Generalization (untrained) items, for PCID we see an ATI such that 'the rich get richer' for Peristim vs Sham and Priming. At post, it appears Peristim and Priming are equally showing this effect vs Sham, but the Priming group had initial starting differences at pretest. For MCat, the pattern is similar to trained items but a little difference in that Peristim maintains improvements better at higher levels of MCat than Sham. Priming does not show much improvement by MCat for untrained items. For Wing, untrained items show a clear positive effect of Wing at pretest, and at posttest the curves have been flatted a bit for Peristim and, to a lesser extent, Priming, such that those with low music aptitude show a benefit from active taVNS by up to 20%.

Figure 13. Modeled phonological categorization accuracy results split by taVNS group, stimulus condition (critical, generalization), and individual difference (MCat, PCID, Wing): (a) Probability model estimates for logistic MEM of accuracy at pre and post. Dotted line represents chance performance at 33% percent probability. (b) Pre-to-post change in probability estimates for logistic MEM of accuracy to show improvement from training. Dotted line at 0% change pre to post.


Crtcl
MCat
Crtcl
PCID
Crtcl
Wing
Gnrlztn
MCat
Gnrlztn
PCID
Gnrlztn
Wing
Day 1
Day 2
Modeled Probability of a Correct Response (%)
taVNS Group
Peristim
Priming
Sham

(a)


Crtcl
Gnrlztn
MCat
PCID
Wing
Modeled Pre to Post change in Probability (%)
taVNS Group
Peristim
Priming
Sham

(b)

The accuracy results for the lexical recognition test are shown in Table 13. Four-way interactions of Stimulus Condition X Individual Difference X Day X taVNS Group survived for MCat and Wing, along with three-way interactions of PCID X Day X Group and Stimulus Condition X PCID X Group. Thus, there were ATI effects for all three IDs. There are situations in which all three IDs, PCID, MCat, and Wing positively predicted lexical recognition accuracy by a large magnitude, the biggest effect being for PCID on Day 1 predicting a jump from 30% at the low end of PCID to about 85% on the high end across Sham and Priming groups. By the same token, there are also situations reflected by the interactions in which all three IDs showed an essentially flat line, or null effect of the ID.

Table 13. Logistic MEM for lexical recognition test accuracy moderated by IDs.

Fixed effects	**Estimate**	***SE***	***p***
Intercept (Mismatch, Day 1, Sham)	0.322	0.129	.012*
Training Day (Day 2)	1.129	0.145	<.001*
taVNS Group (Peristim)	0.287	0.183	.117
taVNS Group (Priming)	0.204	0.182	.263
Day 2 X Peristim	-0.344	0.235	.144
Day 2 X Priming	0.012	0.236	.960
PCID (Non-Ling. Tone Apt.)	0.588	0.130	<.001*
PCID X Day 2	0.034	0.142	.813
PCID X Peristim	-0.408	0.218	.061^
PCID X Priming	0.052	0.190	.783
PCID X Day 2 X Peristim	-0.087	0.207	.674
PCID X Day 2 X Priming	-0.036	0.183	.845
Wing (Music Apt.)	0.185	0.122	.131
Wing X Day 2	0.411	0.153	.007*
Wing X Peristim	0.225	0.224	.314
Wing X Priming	-0.316	0.193	.101
Wing X Day 2 X Peristim	-0.277	0.272	.308
Wing X Day 2 X Priming	0.051	0.248	.838
MCat (Music Exp.)	-0.136	0.123	.269
MCat X Day 2	0.070	0.161	.665
MCat X Peristim	0.601	0.197	.002*

MCat X Priming	0.038	0.213	.860			
MCat X Day 2 X Peristim	-0.145	0.258	.574			
MCat X Day 2 X Priming	0.139	0.279	.618			
Match	0.678	0.119	<.001*			
Match X Day 2	0.018	0.134	.896			
Match X Peristim	-0.130	0.147	.378			
Match X Priming	0.123	0.148	.406			
Match X Day 2 X Peristim	-0.029	.190	.877			
Match X Day 2 X Priming	-0.023	0.197	.908			
Match X PCID	-0.588	0.099	<.001*			
Match X PCID X Day 2	0.121	0.087	.160			
Match X PCID X Peristim	0.468	0.163	.004*			
Match X PCID X Priming	0.334	0.142	.018*			
Match X Wing	0.072	0.096	.454			
Match X Wing X Day 2	-0.096	0.121	.427			
Match X Wing X Peristim	-0.624	0.176	<.001*			
Match X Wing X Priming	0.365	0.158	.021*			
Match X Wing X Day 2 X Peristim	0.392	0.201	.052^			
Match X Wing X Day 2 X Priming	-0.351	0.197	.075^			
Match X MCat	0.109	0.099	.274			
Match X MCat X Day 2	-0.407	0.131	.002*			
Match X MCat X Peristim	-0.017	0.162	.915			
Match X MCat X Priming	0.348	0.177	.049*			
Match X MCat X Day 2 X Peristim	0.418	0.212	.048*			
Match X MCat X Day 2 X Priming	0.327	0.240	.173			
Random effects	**Variance**	***SD***	**Correlation**			
Intercepts | Participant	0.292	0.540				
Day 2 | Participant	0.444	0.666	.32			
Match | Participant	0.121	0.348	-.21	.40		
Match X Day 2 | Participant	0.116	0.341	-.27	-.84	-.13	
Intercepts | Item (presented sound)	0.103	0.321				
Match | Item	0.128	0.357	-.30			
Day 2 | Item	0.044	0.210	-.03	.14		
Wing | Item	0.006	0.078	.35	.46	.73	
Match X Day 2 | Item	0.072	0.268	.37	-.01	-.31	.22

Number of obs.: 31,089; Participants: 72; Items (unique presented sound files): 18

The modeled estimates and standard errors are shown in Figure 14. In Figure 14a, these effects are split by day. Figure 14b reflects the amount of improvement from training day 1 to day 2 for each group by each ID.

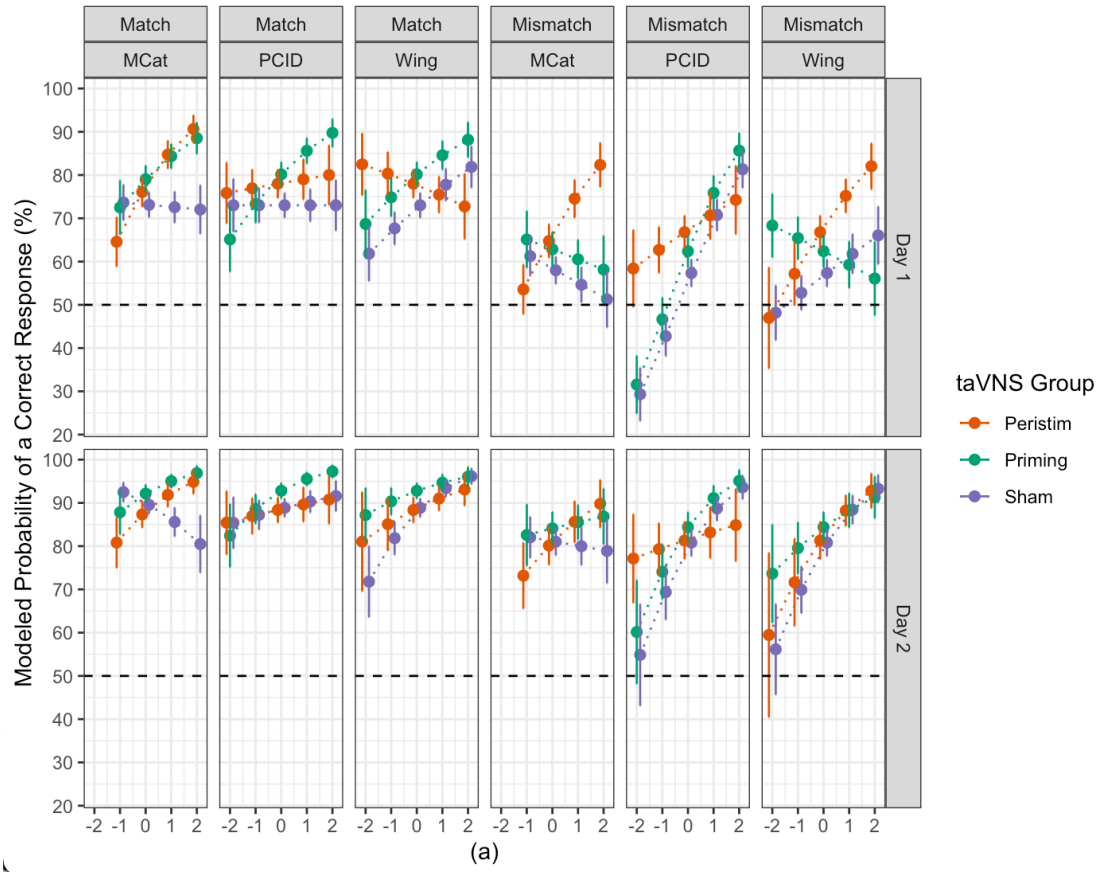
For Match items, for PCID we only see an effect for Priming on day 1 with some separation of effects, Priming greater than Peristim and especially Sham, occurring at higher levels of PCID. Sham improves more day to day at higher levels of PCID, but this essentially moves the Sham learners from a flat line to a somewhat positive PCID trajectory. The advantage of higher levels of PCID for the Priming group is maintained, but mitigated, on day 2 as all learners approached ceiling performance. For MCat, after training on day 1, Sham showed no effect of MCat at about 70-75% across the levels while Peristim and Priming showed a positive effect of MCat, about 65% to 90% across the levels. Day-to-day improvements across levels of MCat were consistent across groups. At Day 2, The positive effect of MCat was maintained for Peristim and Priming, although curiously the Sham group developed a negative trajectory (worse performance at higher MCat levels). For Wing, after training on Day 1 Priming and Sham showed a positive trajectory for Wing while Peristim showed a weakly negative effect. Day-to-day, Sham shows relatively flat improvements across Wing scores while Priming shows larger improvements for those with low Wing scores and Peristim shows larger improvements for those with higher Wing scores. The end result at Day 2 is comparable performance on the high end of Wing for all learners, with some separation at lower levels of Wing: Sham performing the worst, then Peristim, and finally low Wing ability Priming learners on top.

For Mismatch items, For PCID after training on day 1, we see a striking ATI effect for Peristim over Priming and Sham. At low ends of PCID, both Sham and Priming are both hitting

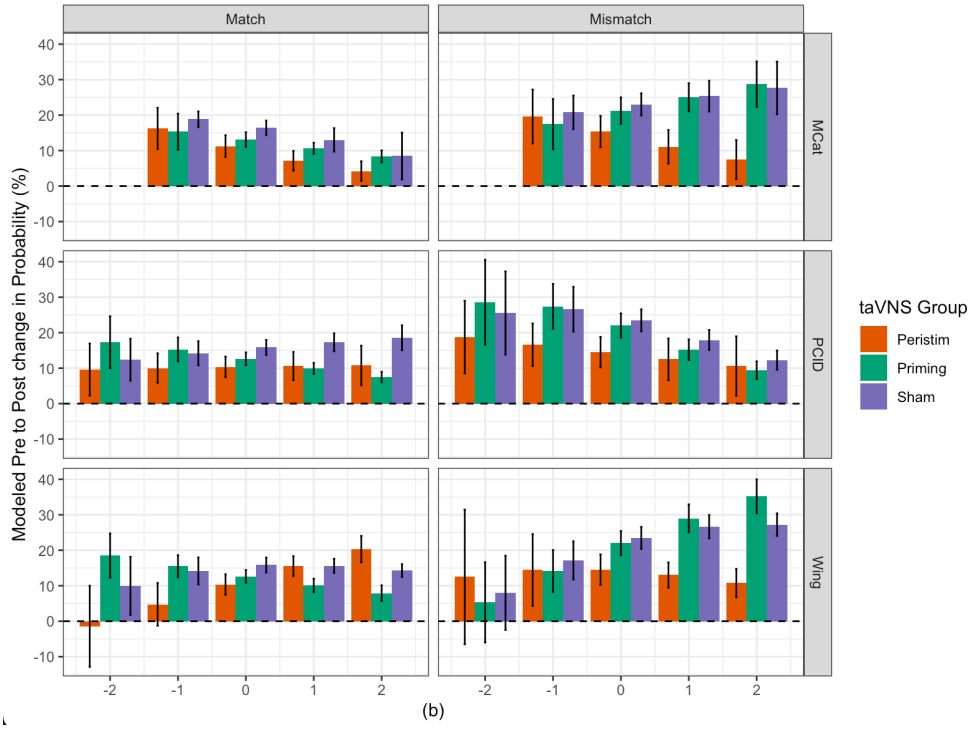
about 30%, below chance performance (indicating they were actively over-endorsing the wrong response option for some items rather than not responding). However, at low ends of PCID for the Peristim group, learners were still at about 60%, above chance for this task. Day-to-day improvements were relatively consistent across groups for PCID levels, and the advantage for Peristim over Sham is still maintained after training on Day 2, but mitigated, and also largely mitigated compared to Priming. Interestingly, at day 2, the influence of PCID is essentially null for the peristim group only, suggesting Peristim taVNS negated the effects of PCID. For MCat, Peristim shows a strong 'rich get richer' ATI effect compared to Priming and Sham after training on day 1. Day-to-day, Priming and Sham improve similarly across MCat levels while Peristim improves the most for lowest levels of MCat. On day 2, this effect is heavily mitigated, but still shows a little separation from Sham on the high end of music experience. For Wing, after training on day 1, both Peristim and Priming are showing ATI effect versus Sham, but in completely different ways. Peristim is performing similar to Sham at low levels of Wing, but at high levels the rich get richer for Peristim. Priming performed similarly to Sham at high levels of Wing, but at low levels Priming performed well above chance compared to Sham. Day-to-day improvements were similar for Priming and Sham with improvement for Peristim being overall lower and more consistent across levels. After training on day 2, Peristim and Sham appear to have identical positive trajectories for Wing, and Priming is similar but still appears to have a small advantage over Sham at low levels of Wing.

Figure 14. Modeled lexical recognition accuracy results split by taVNS group, stimulus condition (match, mismatch), and individual difference (MCat, PCID, Wing): (a) Probability model estimates for logistic MEM of accuracy at training day 1 and day 2. Dotted line represents

chance performance at 50% percent probability. (b) Day 1 to day 2 change in probability estimates for logistic MEM of accuracy to show improvement from training. Dotted line at 0% change day-to-day.


Match
MCat
Match
PCID
Match
Wing
Mismatch
MCat
Mismatch
PCID
Mismatch
Wing
Day 1
Day 2
Modeled Probability of a Correct Response (%)
taVNS Group
Peristim
Priming
Sham

(a)


Match
Mismatch
MCat
PCID
Wing
Modeled Pre to Post change in Probability (%)
taVNS Group
Peristim
Priming
Sham

(b)

The reaction time results for the phonological categorization test are shown in Table 14. A four-way interaction of Stimulus Condition X MCat X Day X taVNS Group survived, as well as a two-way Stimulus Condition X PCID and the simple effect of Wing. Thus, there were only ATI effects for MCat as the other two variables did not interact with taVNS Group. While there was a slope in the expected direction, it's worth noting that the simple effect of Wing was not significant in this sample. It remained in the model because there was a significant contribution of the random slope of Wing by Item; this means that in aggregate across all items there was no consistent significant effect of Wing but that for some items there was a significant effect and others there was not. The effect of PCID was strong for Critical (trained) items and still strong, but weaker for Generalization (untrained) items. For MCat, the direction of effects differed among groups.

Table 14. Linear MEM for phonological categorization test reaction times moderated by IDs.

Fixed effects	**Estimate**	***SE***	***p***	
Intercept (Critical, Day 1, Sham)	6.878	0.030	<.001*	
Training Day (Day 2)	-0.048	0.026	.071^	
taVNS Group (Peristim)	-0.083	0.040	.042*	
taVNS Group (Priming)	-0.046	0.041	.263	
Day 2 X Peristim	-0.010	0.038	.787	
Day 2 X Priming	0.006	0.040	.882	
PCID (Non-Ling. Tone Apt.)	-0.075	0.018	<.001*	
Wing (Music Apt.)	-0.015	0.018	.385	
MCat (Music Exp.)	-0.030	0.026	.255	
MCat X Day 2	0.035	0.025	.159	
MCat X Peristim	0.085	0.040	.038*	
MCat X Priming	0.052	0.043	.224	
MCat X Day 2 X Peristim	-0.050	0.025	.045*	
MCat X Day 2 X Priming	-0.143	0.042	.001*	
Generalization	-0.027	0.030	.381	
Generalization X Day 2	0.075	0.025	.004*	
Generalization X Peristim	0.007	0.025	.772	
Generalization X Priming	-0.017	0.025	.498	
Generalization X PCID	0.027	0.009	.004*	
Generalization X MCat	-0.013	0.016	.419	
Generalization X MCat X Day 2	-0.007	0.016	.680	
Generalization X MCat X Peristim	-0.020	0.025	.421	
Generalization X MCat X Priming	-0.012	0.026	.632	
Generalization X MCat X Day 2 X Peristim	0.061	0.025	.014*	
Generalization X MCat X Day 2 X Priming	0.051	0.026	.051^	
Random effects	**Variance**	***SD***	**Correlation**	
Intercepts | Participant	0.016	0.125		
Day 2 | Participant	0.014	0.118	-.23	
Generalization | Participant	0.003	0.053	-.18	-.11
Intercepts | Item (presented sound)	0.006	0.081		
Day 2 | Item	0.003	0.058	-.37	
Wing | Item	<0.001	0.020	.31	.37
Residual	0.056	0.238		

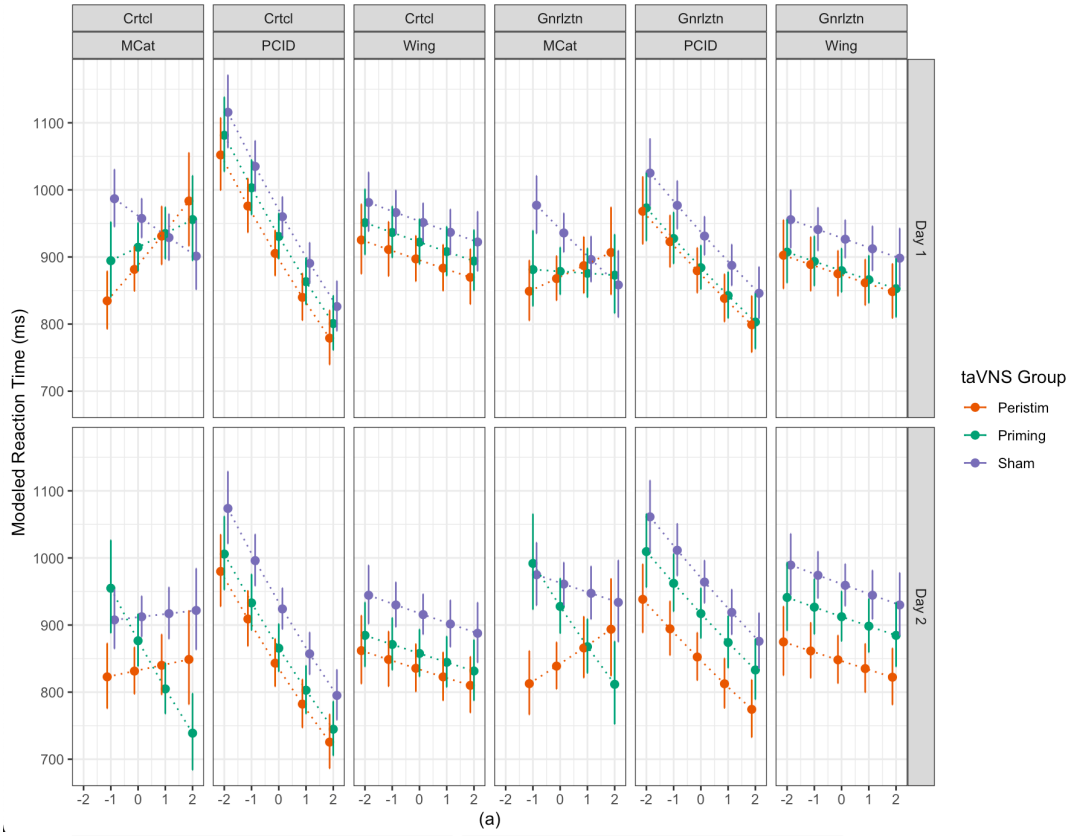
Number of obs.: 9,518; Participants: 72; Items (unique presented sound files): 36

The modeled estimates and standard errors are shown in Figure 15. In Figure 15a, these effects are split by day. Figure 15b reflects the amount of improvement from training day 1 to day 2 for each group by each ID.

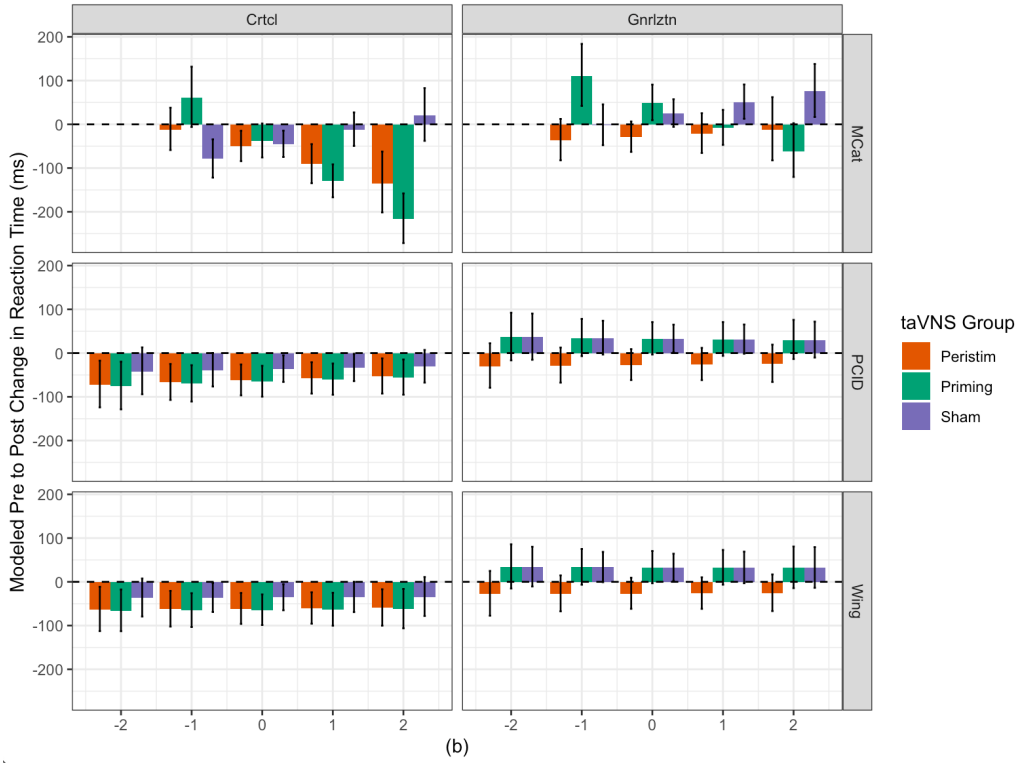
For Critical (trained) items, for MCat the active taVNS show a strong effect of 'the rich get richer' day-to-day as, at the highest level of MCat, Peristim speeds up by about 150 ms and Priming speeds up but about 200 ms. By contrast, the Sham group shows little speed up from day-to-day by MCat. At post, we see that only Priming shows a significant effect of MCat such that more music experience translates to faster RTs, with faster RTs than Sham at higher levels of MCat. Peristim and Sham show flat, parallel lines, so no effect of MCat just overall that Peristim is faster than Sham. Thus, the day-to-day change for Peristim seems to have just corrected some differences that existed at pretest.

For Generalization (untrained) items, day-to-day there appears to be no change for the Peristim group, a small slowdown for Sham at increasing MCat levels, and a slowdown for Priming at low MCat and a speedup at high MCat. Given the existing group differences before any taVNS at pretest, looking at results at post, it appears that the ATI comparison between Priming and Sham on trained items has been very comparably maintained for untrained items, with a (given the large error bars) flat line for Peristim also maintained.

Figure 15. Modeled phonological categorization reaction results split by taVNS group, stimulus condition (critical, generalization), and individual difference (MCat, PCID, Wing): (a) RT model estimates for linear MEM of RTs at pre and post. (b) Pre-to-post change in RT estimates for linear MEM of RTs to show improvement from training. Dotted line at 0% change pre to post.


Crtcl
MCat
Crtcl
PCID
Crtcl
Wing
Gnrlztn
MCat
Gnrlztn
PCID
Gnrlztn
Wing
Day 1
Day 2
Modeled Reaction Time (ms)
taVNS Group
Peristim
Priming
Sham

(a)


Crtcl
Gnrlztn
MCat
PCID
Wing
Modeled Pre to Post Change in Reaction Time (ms)
taVNS Group
Peristim
Priming
Sham

(b)

The reaction time results for the lexical recognition test are shown in Table 15. Three-way interactions of Stimulus Condition X MCat X taVNS Group and PCID X Day X taVNS Group survived, as well as a two-way Stimulus Condition X PCID and the simple effect of Wing. Thus, there were only ATI effects for MCat and PCID as Wing did not interact with taVNS Group. Just like for phonological categorization, a non-significant effect of Wing remained in model because there was a significant random slope of Wing by item that persisted. This again means that in aggregate across all items there was no consistent significant effect of Wing but that for some items there was a significant effect and others there was not.

Table 15. Linear MEM for lexical recognition test reaction times moderated by IDs.

Fixed effects	**Estimate**	***SE***	***p***		
Intercept (Mismatch, Day 1, Sham)	6.901	0.025	<.001		
Training Day (Day 2)	-0.043	0.018	.022		
taVNS Group (Peristim)	-0.043	0.041	.295		
taVNS Group (Priming)	0.057	0.041	.173		
Day 2 X Peristim	-0.041	0.031	.191		
Day 2 X Priming	-0.094	0.031	.003		
PCID (Non-Ling. Tone Apt.)	0.018	0.028	.517		
PCID X Day 2	-0.071	0.020	.001		
PCID X Peristim	-0.057	0.043	.193		
PCID X Priming	0.010	0.040	.796		
PCID X Day 2 X Peristim	0.061	0.033	.065		
PCID X Day 2 X Priming	0.082	0.030	.008		
Wing (Music Apt.)	0.016	0.019	.390		
MCat (Music Exp.)	-0.033	0.027	.232		
MCat X Peristim	0.069	0.042	.109		
MCat X Priming	-0.020	0.046	.658		
Match	-0.116	0.023	<.001		
Match X Peristim	0.029	0.022	.200		
Match X Priming	0.004	0.023	.861		
Match X MCat	0.038	0.015	.010		
Match X MCat X Peristim	-0.031	0.023	.177		
Match X MCat X Priming	-0.074	0.024	.004		
Random effects	**Variance**	***SD***	**Correlation**		
Intercepts | Participant	0.019	0.136			
Day 2 | Participant	0.011	0.104	-.32		
Match | Participant	0.007	0.085	-.19	.06	
Match X Day 2 | Participant	0.006	0.085	.29	-.22	-.61
Intercepts | Item (presented sound)	0.002	0.041			
Match | Item	0.006	0.079	-.49		
MCat | Item	<0.001	0.012	-.70	.24	
Wing | Item	<0.001	0.010	.34	.09	-.84
Residual	0.091	0.301			

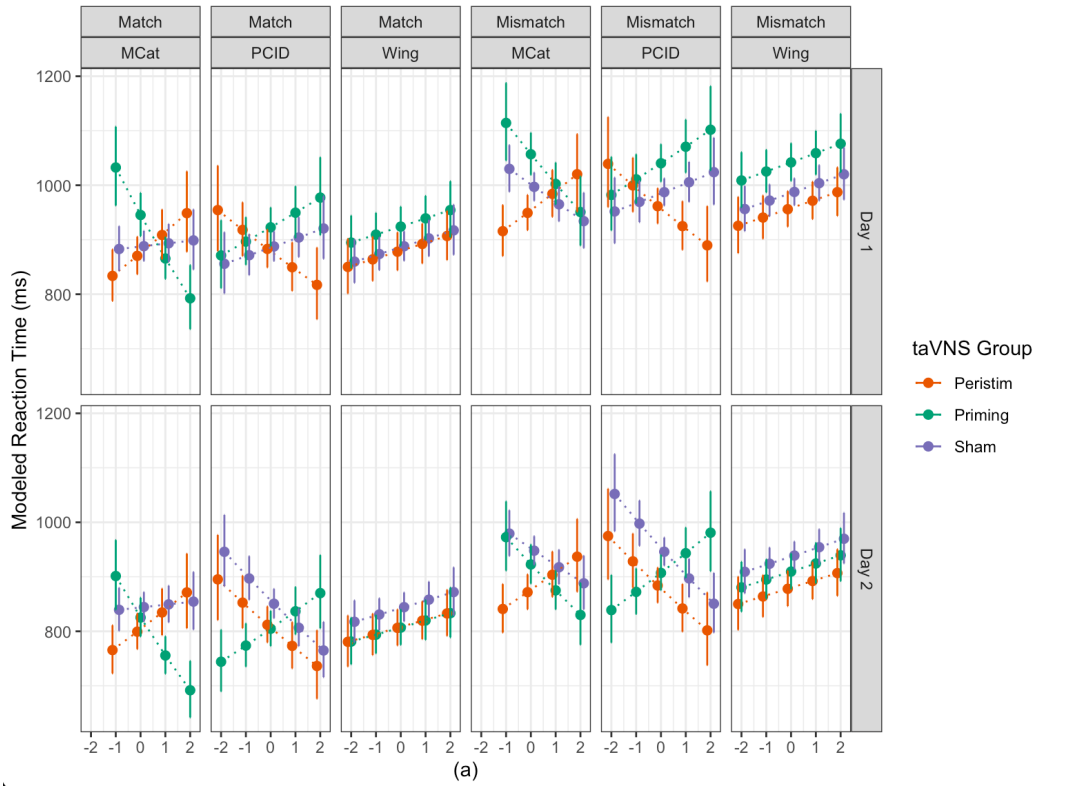
Number of obs.: 22,618; Participants: 72; Items (unique presented sound files): 18

The modeled estimates and standard errors are shown in Figure 16. In Figure 16a, these effects are split by day. Figure 16b reflects the amount of improvement from training day 1 to day 2 for each group by each ID.

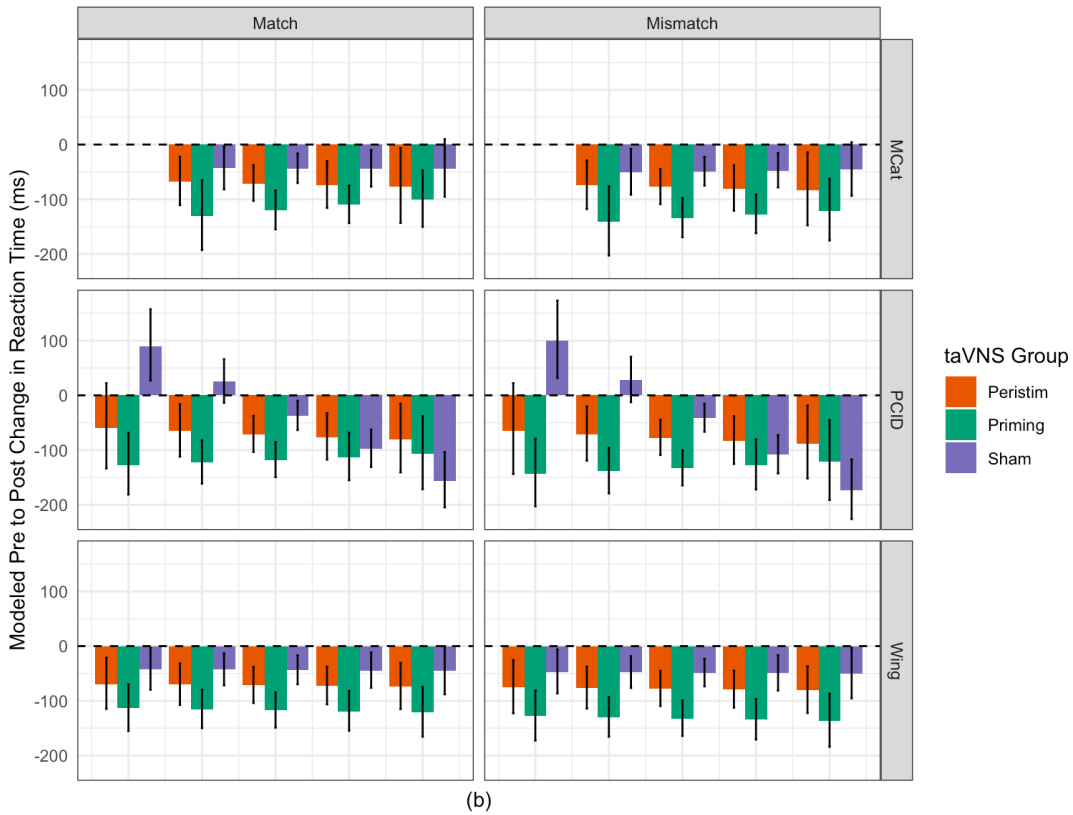
For PCID, the ATI effect did not interact with Stimulation Condition and so the results are consistent across Match and Mismatch items. After training on day 1, there is an effect of 'the rich get richer' for the Peristim taVNS groups vs Sham and Peristim, with the separation of the groups coming out at higher levels of PCID. Day-to-day, the Sham group at lower levels of PCID does not speed up while Priming and, to a lesser extent, Peristim do across the board. After training on day 2, Peristim and Sham have similar slopes such that higher PCID results in faster RTs. For Priming, while the trajectory is counter to the expected direction, the error is such that there is essentially no effect of PCID for the Priming group. Interestingly, at low ends of PCID, Priming is faster than the Sham group.

For MCat on Match items, after training on day 1, there is a significant effect of MCat for the Priming group such that more music experience results in faster RTs, although at lower MCat Priming is slower than the other groups, which have no notably significant effect of MCat. Day-to-day, the magnitude of the speedup for Priming is the greatest, followed by Peristim and Sham which don't show much, if any, speedup. After training on day 2. Priming is faster than Peristim and sham at higher levels of MCat, and Peristim is faster than Priming at the lowest level of MCat. For Mismatch items, after training on day 1, Priming and Sham show a speedup with increased MCat level while Peristim is effectively flat. Day-to-day, the pattern persists that Priming shows the largest speedup followed by Peristim and Sham's more unreliable speedups. At day 2 for Mismatch items, the only group separation is a speedup for Peristim compared to Priming and Sham at low levels of MCat. Effectively Peristim flattened the effect of MCat which resulted in a benefit to participants with low music experience compared to the other groups.

Figure 16. Modeled lexical recognition reaction results results split by taVNS group, stimulus condition (match, mismatch), and individual difference (MCat, PCID, Wing): (a) RT model estimates for linear MEM of RTs at training day 1 and day 2. (b) Day 1 to day 2 change in RT estimates for linear MEM of RTs to show improvement from training. Dotted line at 0% change day-to-day.


Match
Match
Match
Mismatch
Mismatch
Mismatch
MCat
PCID
Wing
MCat
PCID
Wing
Day 1
Day 2
Modeled Reaction Time (ms)
1200
1000
800
-2 -1 0 1 2
taVNS Group
Peristim
Priming
Sham

(a)


Match
Mismatch
MCat
PCID
Wing
Modeled Pre to Post Change in Reaction Time (ms)
100
0
-100
-200
taVNS Group
Peristim
Priming
Sham

(b)

Given the abundance of results, particularly with the different trajectories for the behavioral outcomes, behavioral results are further reframed and condensed into Table 16 for the 'easier' test conditions (trained items for phonological categorization; match items for lexical recognition) and for the 'harder' test conditions (untrained items for phonological categorization; mismatch items for lexical recognition). This table shows whether or not there was a significant effect (via model releveling) for a particular individual difference variable on a particular test, on a particular day, for a particular taVNS group. If significant, the direction of the effect is indicated as positive (higher accuracy; slower RTs), negative (lower accuracy; faster RTs), or "--" (no significant slope; "--^" if it was marginal). All instances of aptitude- (or experience-) by-treatment interactions are surround with '{}' in the table. In context, there were two patterns of ATIs that resulted from this investigation. The first pattern is when taVNS produces an effect of an ID where there wasn't one for Sham; a type of 'the rich get richer' in which, at higher levels of ability, the active taVNS group has either higher accuracy or faster RTs while there is no significant change for the Sham group. The second pattern of ATI observed in this dissertation is when taVNS obviates the negative effects of an ID variable; in other words, when the Sham group shows a significant slope for an ID—lower scores translate to worse outcomes, and vice versa—but active taVNS shows better outcomes for those learners at low ability than the Sham group, and those outcomes are the same across levels of the ID.

Table 16. Summary of the all of the positive, negative, and flat slopes for each individual difference variable in Chapter 6 (PCID, MCat, and Wing) split by task (Phonological Categorization, Lexical Recognition), Stimulus Condition (Critical/Generalization, Match/Mismatch), for each taVNS group.

	PCID			MCat			Wing		
	Peri	**Prim**	**Sham**	**Peri**	**Prim**	**Sham**	**Peri**	**Prim**	**Sham**
Critical/Match Accuracy									
Phon.Cat.: Pre	--^	Pos	Pos	--	--	--	--	--	Pos
Phon.Cat.: Post	Pos	Pos	Pos	--	--	--	--	--	Pos
Lex.Rec.: Day 1	--	{Pos}	{--}	{Pos}	--^	{--}	{--}	{--}^	{Pos}
Lex.Rec.: Day 2	--	{Pos}	{--}	{Pos}	--	{--}	{--}	{--}	{Pos}
Critical/Match Reaction Time									
Phon.Cat.: Pre	Neg	Neg	Neg	--	--	--	--	--	--
Phon.Cat.: Post	Neg	Neg	Neg	--	{Neg}	{--}	--	--	--
Lex.Rec.: Day 1	--	--	--	--	{Neg}	{--}	--	--	--
Lex.Rec.: Day 2	--	--	--	--	{Neg}	{--}	--	--	--
Generalization/Mismatch Accuracy									
Phon.Cat.: Pre	--^	Pos	--	--	--	--	--	--	--
Phon.Cat.: Post	--	Pos	--	--	--^	--	{--}	{--}	{Pos}
Lex.Rec.: Day 1	{--}	Pos	{Pos}	{Pos}	--	{--}	{Pos}	--	{--}
Lex.Rec.: Day 2	{--}	Pos	{Pos}	--	--	--	{--}^	{--}	{Pos}
Generalization/Mismatch Reaction Time									
Phon.Cat.: Pre	Neg	Neg	Neg	--	--	--	--	--	--
Phon.Cat.: Post	Neg	Neg	Neg	--	--	--	--	--	--
Lex.Rec.: Day 1	--	--	--	--	--	--	--	--	--
Lex.Rec.: Day 2	--	--	--	--	--	--	--	--	--

Note. ^ Indicates a marginal slope in the expected direction. {} Surround each part of an ATI effect comparing at least one active taVNS group with sham in the same row for an ID variable. Differences at pretest are not interpretable, and neither are they conclusive at post if they persisted from pretest.

PCID is predictive of accuracy most consistently for the Priming taVNS group, significantly for every task and condition; by contrast the Sham group does not show an effect of PCID for easier lexical recognition trials or harder phonological categorization trials. These results with the context of the plotted results above show an ATI effect in which Priming taVNS produces an effect of PCID: those with higher PCID show larger improvements in accuracy, but only for those that received priming taVNS stimulation (~5-15% improvement for lexical recognition). Peristim taVNS by contrast only shows an effect of PCID for trained items on phonological categorization. However, for Peristim, the more interesting results for PCID are that Peristim does *not* show a significant effect of PCID for mismatch items on lexical recognition, while the Sham group does. Coupled with the graphs, this ATI effect for Peristim suggests that peristimulus stimulation obviated the effect of PCID: there is no longer influence of PCID on the Peristim group because those at the low end of PCID ability have been enhanced by taVNS (~20-30% improvement over Sham at low PCID scores). There are a few taVNS group effects for phonological categorization accuracy that differ from Sham, but they are not interpreted as ATI effects here because they either occur at pretest (before any stimulation where there should not be group differences) or they were differences at posttest that persisted from pretest, so it's unclear whether there would be an ATI effect in the absence of pretest group differences. By contrast, there is no effect of PCID on lexical recognition RTs for any group, and there is an effect of PCID for every group for both conditions of phonological categorization. Thus, PCID results show both types of ATI patterns for accuracy, and no ATIs for RTs.

MCat is not predictive of accuracy or RT at all for the Sham group on either task. Peristim taVNS produces an effect of MCat on accuracy for match items on the lexical

recognition test on both days (~15-20% improvement at the highest level of MCat), and also the mismatch items on day 1 (~30% improvement in accuracy at high MCat). Priming taVNS instead produced an effect of MCat on RTs for match items on lexical recognition across both days (~100-150 ms decrease in RT at the highest level of MCat) and also on the trained phonological categorization items at posttest (~150 ms decrease in RT at high MCat). Thus, MCat results only show ATI patterns for producing effects of MCat for both accuracy and RTs across phonological categorization and lexical recognition.

Wing is not predictive of RT for any group on either task. For accuracy, Wing is predictive of accuracy for the Sham group for easier items on both days and for harder items at posttest and day 2. Both Peristim and Priming taVNS obviate the effects of Wing for match items on lexical recognition on both days (~10-20% improvement at low Wing ability on day 1, and ~10-15% on day 2), and also for harder items on both tasks at day 2 (~5-20% improvement for lexical recognition, ~15-20% for phonological categorization at low Wing scores). Additionally, Peristim taVNS produces an effect of Wing for mismatch items on day 1 (~15% improvement at high Wing scores). Thus, Wing results show ATI patterns only for accuracy results, primarily in obviating the effects of Wing to enhance the outcomes of learners with low music aptitude.

6.2.2 taVNS moderates effects of individual differences on physiological tone learning difficulty

Generalized additive mixed modeling was again used to analyze pupillometry data. Pupillometry data were preprocessed in three steps. (1) Data were downsampled to 50 Hz (one datapoint every 20 ms) as recommended for GAMMs by van Rij et al. (2019), since above that the added detail does not significantly change the results but does significantly increase the time it takes a computer to calculate the model. (2) The 500 ms (phonological categorization) or 750 ms (lexical recognition) baseline period before each trial was subtracted from the trial for each person. (3) Any trials for which more than 33% of the data were missing (due to blinks, saccades, looking offscreen, etc.) were rejected from analysis. GAMMs were implemented with the mgcv package (Wood, 2017) following previous recommendations in applying GAMMs to pupillometry data and language science data (Sóskuthy, 2017; 2021; van Rij et al., 2019), including an autoregressive model and random smooths for participants and items. Model testing was conducted using ordered factors (Wieling, 2018) for the parametric smooths and reference/difference smooths for the random effects structure (Sóskuthy, 2021) to arrive at the model of best fit. Using ordered factors allows direct interpretation of the GAMMs summary table and *p*-values as in Chapter 5, unlike the modeling procedure used in Chapter 4. The basic smooths of time and individual difference become reference smooths (similar to the intercept in a multiple regression or mixed effects model), and all other smooth terms with an ordered factor represent a difference smooth between that reference smooth and when the value of the ordered factor is set to true. For example, the reference smooth s(Time) is for Sham on Day 1. The ordered factor smooth s(Time):IsPrim would be the difference curve between Priming taVNS

and Sham taVNS on Day 1 over the time course of a trial, and a significant *p*-value for this second term would indicate that the difference between Priming and Sham on Day 1 was significant.

Because the IDs of interest are continuous or ordinal and could influence different points in time in the pupil response, the reference smooth in the model becomes a more complicated tensor product interaction. Instead of just s(Time), there is also s(ID), representing a potential nonlinear effect of an individual difference variable on the pupil response overall, and also ti(Time, ID), the tensor production interaction which allows the pupil response to vary by ID in a non-linear way along the time course of the pupil response. Plotting is more necessary than ever in this context to determine the direction of the effects (when the effects of IDs occur and the directions of the effect), and thus plots of model terms are provided after every GAMM. Curves are added together to show cumulative effects on the pupil response to answer research questions. A backward elimination procedure was used like in the previous Chapter, starting with a fully maximal model and using backward elimination to arrive at the model of best fit.

The results of the final GAMMs models are presented below. Again, given the ordered factor model specification, the parametric coefficients represent an overall intercept difference for the pupil response of a given factor, while significant smooth terms indicate significant 'wiggliness' over the course of a given trial and/or ID effect. The results interpreted below are focused with respect to non-linguistic tone aptitude and musicality and any interactive effects with Group.

6.2.2.1 Non-linguistic tone aptitude

For non-linguistic tone aptitude, there was an ATI observed for the Peristim group in the phonological categorization test, shown in Table 17 and Figure 18 with the significant interaction of ti(Time, PCID):IsDay2Peristim. This term indicates that the effect of PCID over the time course of the pupil response is significantly different between Peristim on Day 2 versus Sham on Day 2. Because no tensor product interactions for Generalization items survived, it appears this ATI is consistent across trained and untrained items.

Table 17. GAMM summary table for phonological categorization test X non-linguistic tone aptitude pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Critical, Day 1, Sham)	12.824	6.166	.038*
IsDay2	-27.516	7.192	<.001*
IsDay2Peristim	20.420	13.412	.128
IsDay2Priming	-3.617	11.535	.754
IsGeneralizationDay2Peristim	-11.241	6.711	.094^
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Critical, Day 1, Sham)	17.696	18.455	<.001*
s(PCID) (Critical, Day 1, Sham)	1.010	1.011	.458
ti(Time, PCID) (Critical, Day 1, Sham)	116.659	164.000	<.001*
s(Time):IsDay2	10.323	12.961	<.001*
s(Time):IsDay2Peristim	9.624	12.132	.004*
s(PCID):IsDay2Peristim	3.697	3.886	.145
ti(Time, PCID):IsDay2Peristim	84.063	115.792	<.001*
s(Time):IsDay2Priming	9.565	12.274	.009*
s(X gaze position, Y gaze pos.)	77.759	78.951	<.001*
Random smooth terms	**edf**	**Ref.df**	***p***
s(Time, Participant)	315.948	358.000	<.001*
s(Time, Participant):IsDay2	285.969	360.000	<.001*
s(Time, Item)	133.346	179.000	<.001*
s(Time, Item):IsDay2	127.977	180.000	<.001*
s(Time, Item):IsDay2Peristim	112.952	179.000	<.001*
s(Time, Item):IsDay2Priming	119.052	180.000	<.001*
s(PCID, Item)	13.354	179.000	<.001*
s(PCID, Item):IsDay2Peristim	3.990	179.000	<.001*

Number of obs.: 637,867; Participants: 72; Items: 36

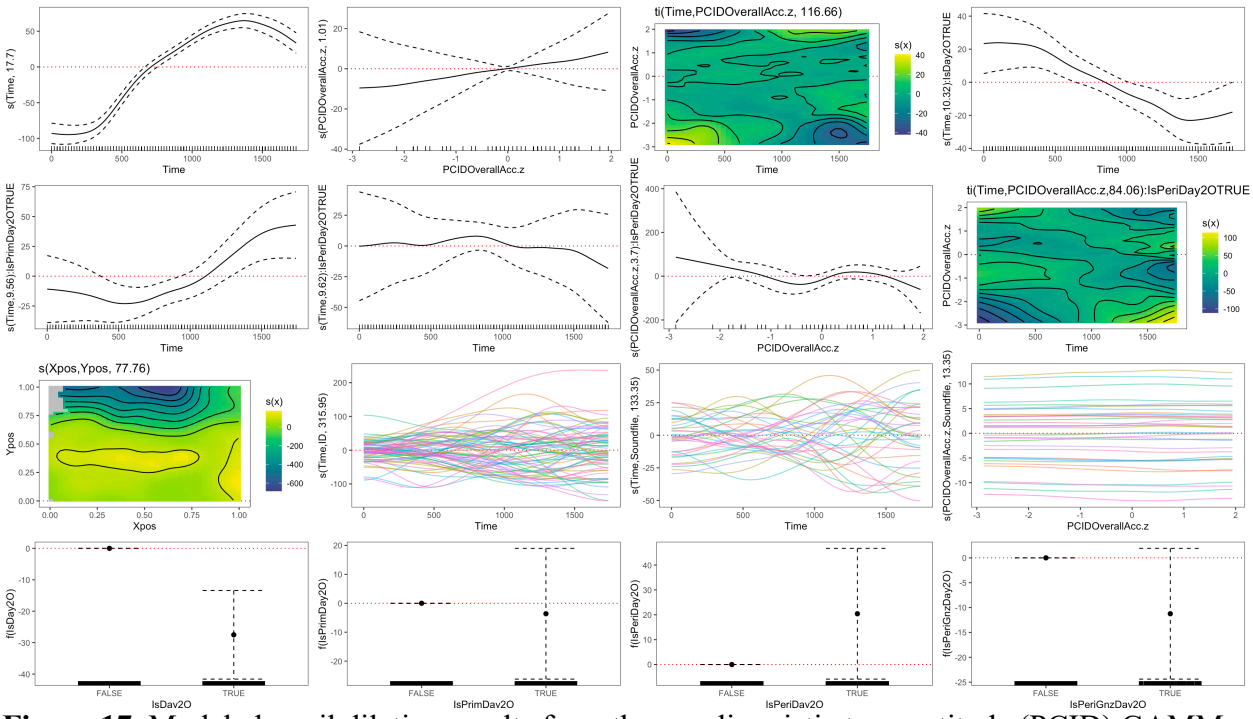


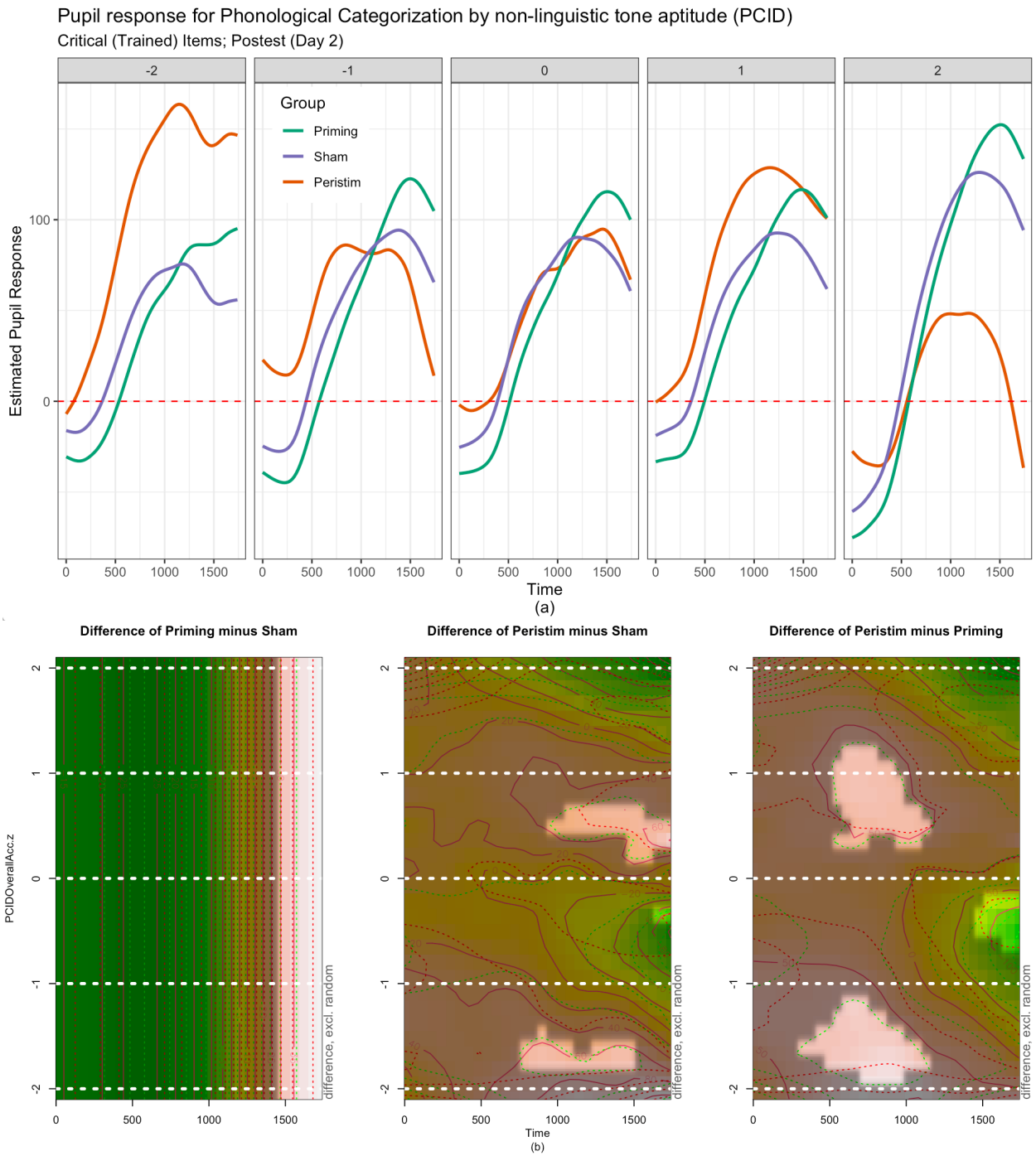
Figure 17. Modeled pupil dilation results from the non-linguistic tone aptitude (PCID) GAMM for phonological categorization. Predicted pupillometry fixed smooth effects and interactions for accurate responses. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point. Bottom row shows predicted pupillometry parametric effects. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. Predicted pupillometry fixed smooth effects for GAMM of phonological categorization pupil response for accurate responses at pre and post. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point.

Due to the difficulty in interpreting the tensor product interactions just from difference plots, the main ATI effect of interest above is summed and plotted below in Figure 18. It can be seen in Figure 18a, which looks at the curves for groups at posttest split by specific levels of PCID, that Priming and Sham vary rather little from PCID, at most showing a slight early decrease and later increase in the pupil response. For Peristim, we see a stronger effect with a high pupil response at low PCID and a lower pupil response at high PCID. Note here that the pupillometry effects here parallel those in Chapter 5: Priming taVNS shows a smaller pupil

response for much of the first 1,000 ms compared to Sham. In Chapter 5, Sham and Peristim showed the same pupil response, but here, now that we split the response by PCID, we see that averaging across the effect of PCID is what lent the curve to looking no different from Sham. As one final parallel, we see a larger response at the end of the curve for Priming, which we now know from Chapter 5 is an artifact of Tone 4 only.

Figure 18b shows summed response difference heatmap plots using defaults of the itsadug package v.2.4 (van Rij et al., 2020). Highlighted areas in these plots represent areas of the strongest differences between two conditions. Colors heading to yellow-to-green represent smaller pupil dilation while colors heading to red-to-white represent a larger pupil response. On the left plot we see that the effect of PCID on the y-axis was constant between Priming and Sham, thus the difference between groups was consistent and showed up as a higher response for Priming at the end of the pupil response. In the middle and right plots we see that Peristim had a larger pupil response around the middle of the pupil response than Sham and even more than Priming at differing levels of PCID, most strongly at about 1.5 standard deviations below the mean of PCID for the sample and about 0.5-1.0 standard deviations above the mean of PCID. There was also a small decrease in pupil size for Peristim compared to both groups at the end of the trial at about 0.5 standard deviations below the mean of PCID.

Figure 18. Modeled phonological categorization pupillometry results split by taVNS group and level of non-linguistic tone aptitude (PCID, z-scored) for trained items at posttest: (a) Estimated pupillometry model estimates from GAMM of PCID at posttest faceted by z-scored levels of PCID. (b) taVNS group change in pupillometry estimates from GAMM of PCID at posttest. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).



There was also an ATI observed for the lexical recognition test, shown in Table 18 and Figure 19 with multiple significant ti(Time, PCID) interactions for Priming compared to Sham,

Peristim compared to Sham, as well as Priming and Peristim each at Day 2 compared to Sham at Day 2. Because no effect of Stimulus Condition survived, it appears these ATI effects are consistent across match and mismatch items.

Table 18. GAMM summary table for lexical recognition test X non-linguistic tone aptitude pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Match/Mismatch, Day 1, Sham)	22.20	13.02	.0882
IsPeristim	-29.74	22.04	.177
IsPriming	50.16	22.30	.025
IsDay2	25.93	14.32	.070
IsDay2Peristim	-27.36	24.63	.267
IsDay2Priming	-41.72	27.26	.126
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Match/Mismatch, Day 1, Sham)	17.966	18.374	<.001
s(PCID) (Match/Mismatch, Day 1, Sham)	1.018	1.019	.019
ti(Time, PCID) (Match/Mismatch, Day 1, Sham)	229.016	256.952	<.001
s(Time):IsPeristim	6.946	8.258	.782
s(PCID):IsPeristim	1.035	1.036	.253
ti(Time, PCID):IsPeristim	182.143	206.888	<.001
s(Time):IsPriming	13.024	14.505	<.001
s(PCID):IsPriming	1.237	1.257	.005
ti(Time, PCID):IsPriming	205.681	229.783	<.001
s(Time):IsDay2	13.811	15.465	<.001
s(PCID):IsDay2	1.023	1.024	.759
ti(Time, PCID):IsDay2	225.480	255.418	<.001
s(Time):IsDay2Peristim	9.914	11.529	.304
s(PCID):IsDay2Peristim	1.032	1.033	.831
ti(Time, PCID):IsDay2Peristim	158.057	185.940	<.001
s(Time):IsDay2Priming	13.394	14.857	<.001
s(PCID):IsDay2Priming	5.622	5.696	.013
ti(Time, PCID):IsDay2Priming	199.942	223.548	<.001
s(X gaze position, Y gaze pos.)	78.659	78.996	<.001
Random smooth terms	**edf**	**Ref.df**	***p***
s(Time, Participant)	247.132	354.000	<.001
s(Time, Participant):IsDay2	245.736	354.000	<.001
s(Time, Item)	7.493	89.000	.065
s(Time, Item):IsPeristim	0.288	89.000	<.001

s(Time, Item):IsPriming	0.107	89.000	<.001
s(Time, Item):IsDay2	6.559	89.000	.016
s(Time, Item):IsDay2Peristim	0.219	89.000	<.001
s(Time, Item):IsDay2Priming	0.101	89.000	<.001
s(PCID, Item)	.141	89.000	<.001*
s(PCID, Item):IsPeristim	0.108	89.000	<.001*
s(PCID, Item):IsPriming	0.080	89.000	<.001*
s(PCID, Item):IsDay2	0.104	89.000	<.001*
s(PCID, Item):IsDay2Peristim	0.094	89.000	<.001*
s(PCID, Item):IsDay2Priming	0.084	89.000	<.001*

Number of obs.: 2,637,160; Participants: 72; Items: 18

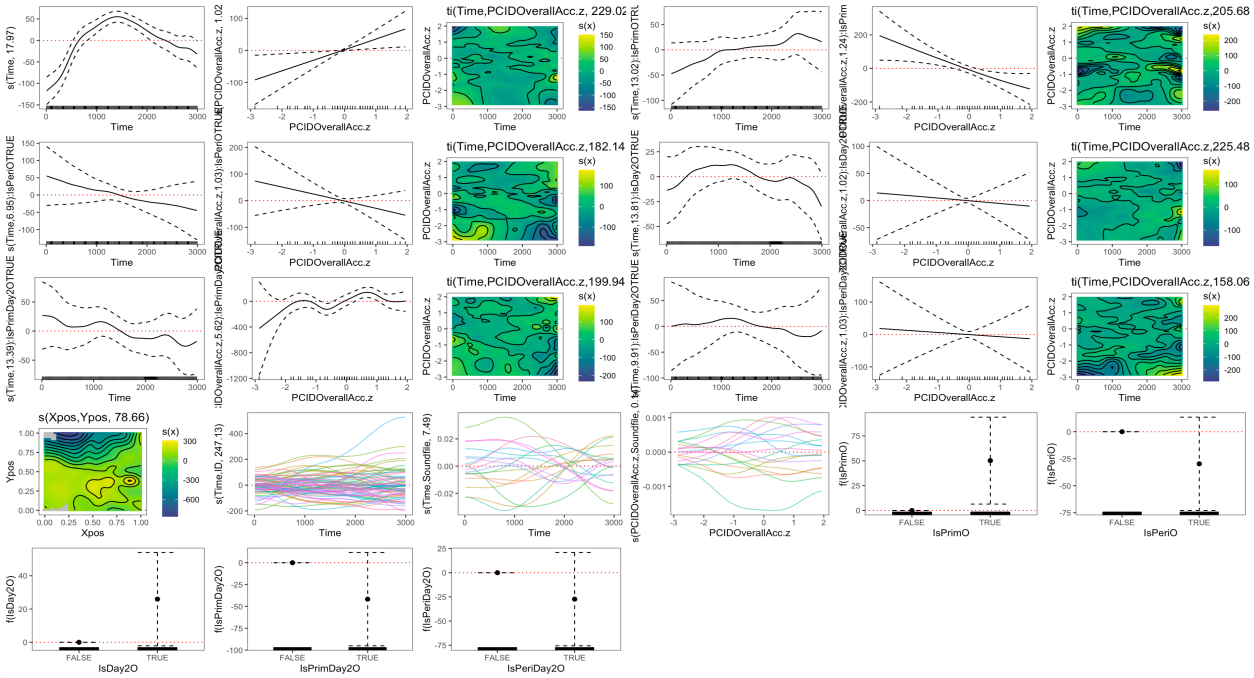


Figure 19. Modeled pupil dilation results from the non-linguistic tone aptitude (PCID) GAMM for lexical recognition. Predicted pupillometry fixed smooth effects and interactions for accurate responses. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. Predicted pupillometry fixed smooth effects for GAMM of lexical recognition pupil response for accurate responses at day 1 and 2. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point.

Due to the difficulty in interpreting the tensor product interactions just from difference plots, the ATI effects of interest above are summed and plotted below in Figure 20 for day 1 and Figure 21 for day 2. It can be seen in Figure 20a that there is quite a bit more wiggliness in these curves versus phonological categorization. Firstly, on day 1 it appears Sham's pupil response gradually increases with PCID. Peristim's response appears broadly similar to Sham's, and Figure 20b reveals that the main significant difference between the groups is between PCID at 0 and +1 standard deviation around average PCID: Peristim here shows a smaller pupil response between about 1,500 and 3,000 ms. For Priming, there are two trends worth noting compared to Sham: at below-average PCID, Priming shows a much larger pupil response than Sham, and between 1 and 2 SDs of above-average PCID, Priming shows (at a smaller magnitude) a smaller pupil response than Sham. Comparing Peristim and Priming, Peristim has a smaller pupil response than Priming on the later part of the pupil response ranging from about -2 SD of PCID up through +1 SD of PCID.

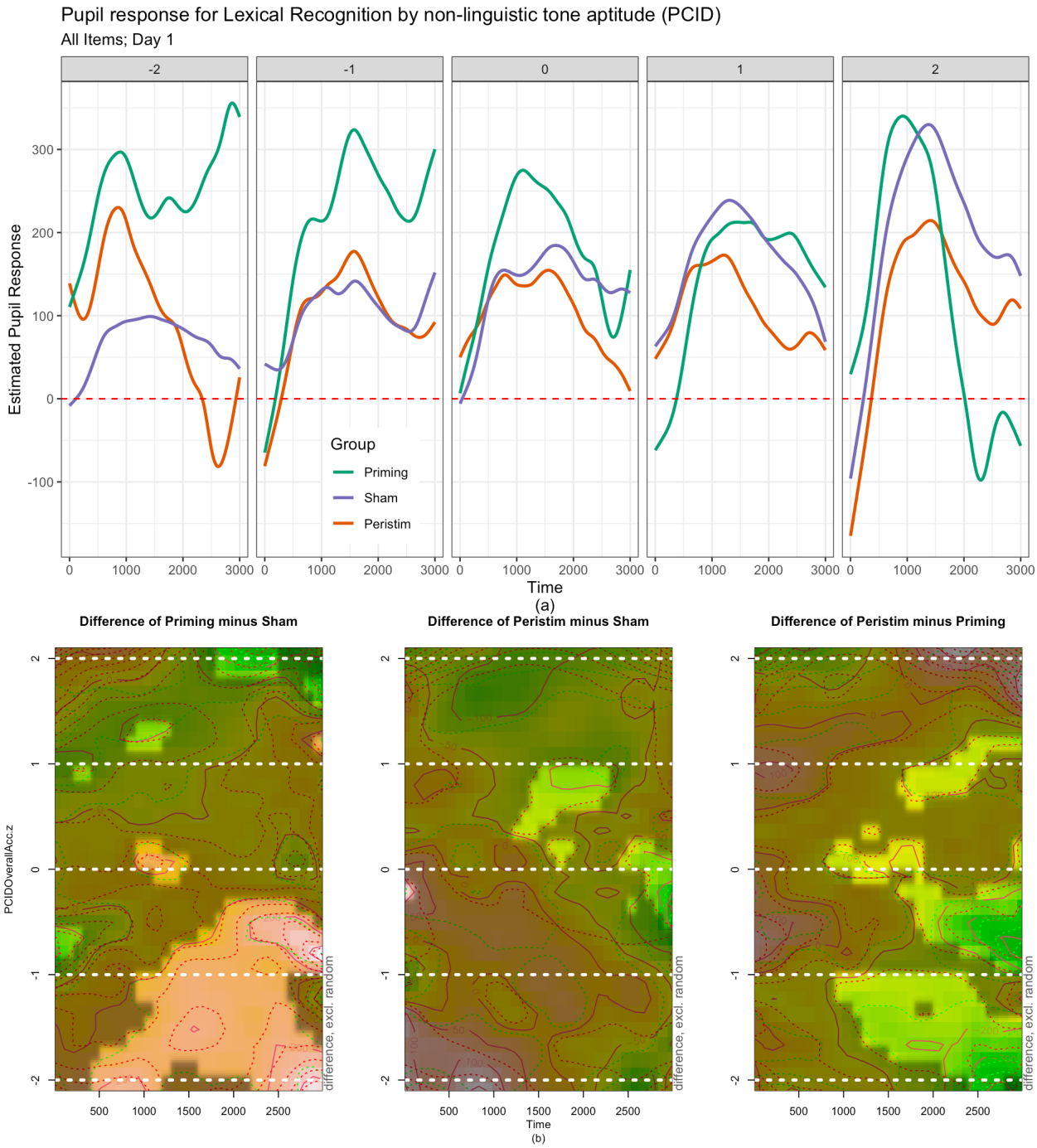


Figure 20. Modeled lexical recognition pupillometry results split by taVNS group and level of non-linguistic tone aptitude (PCID, z-scored) for all items after training on day 1: (a) Estimated pupillometry model estimates from GAMM of PCID on day 1 faceted by z-scored levels of PCID. (b) taVNS group change in pupillometry estimates from GAMM of PCID at day 1. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).

Looking at the end of the second day of training in Figure 21, we see the ATI effects have been largely mitigated, but the patterns are still largely the same. Figure 21b on the left shows that the differences between Priming and Sham are smaller, but the pattern is largely the same, with large swaths of descriptively higher pupil responses for Priming (only pockets of significance around -.5 and +.5 SD around PCID), and a decrement in pupil response compared to Sham at the highest values of PCID, although now not significant. For Peristim minus Sham, we see a very similar effect as on day 1, a smaller pupil response at the end of the trial between 0 and +1 SDs above the PCID mean. The differences between Peristim and Priming remain but are also mitigated such that the largest differences are for a smaller pupil response for Peristim between 0 and +1 SDs of PCID.

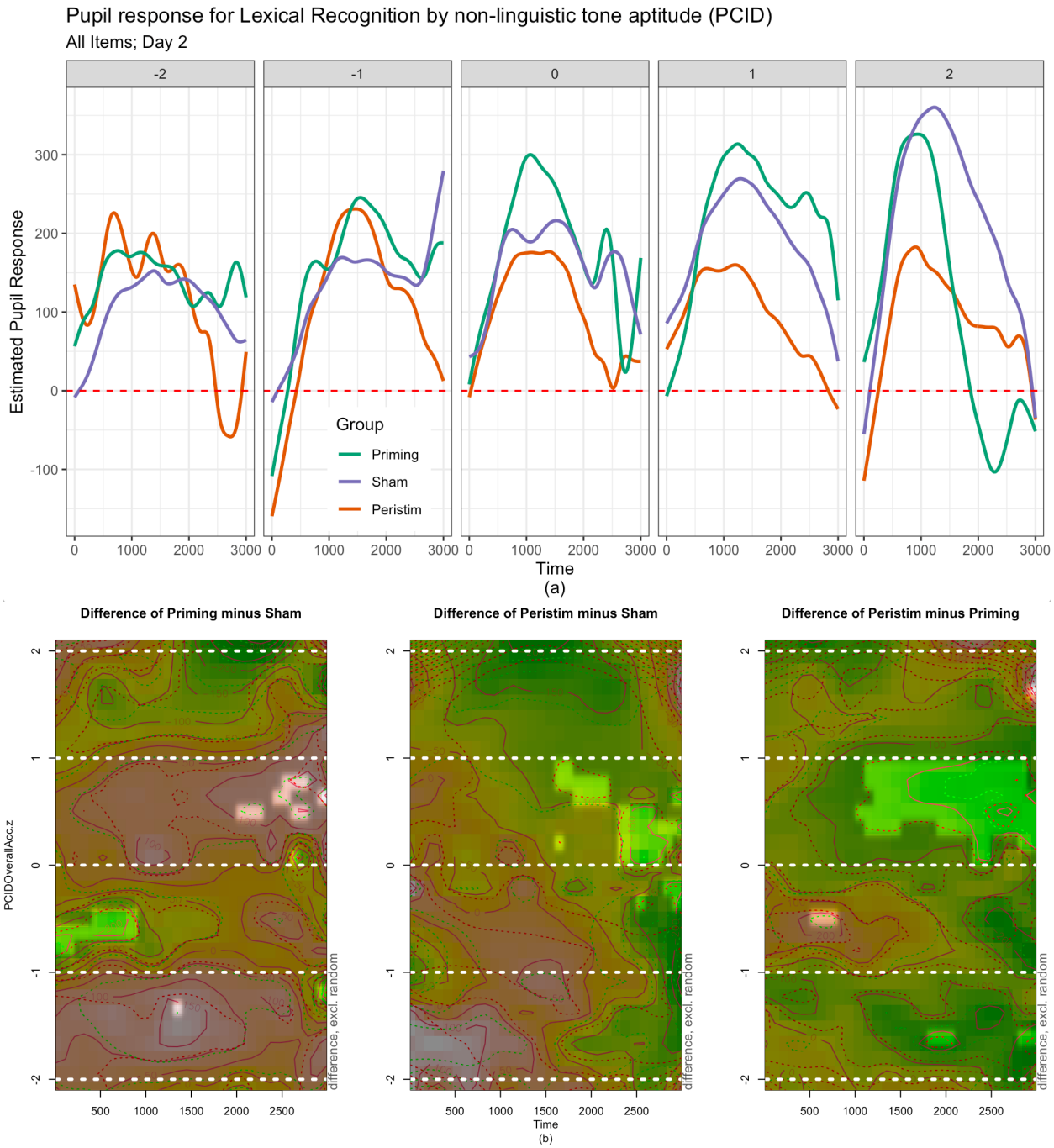


Figure 21. Modeled lexical recognition pupillometry results split by taVNS group and level of non-linguistic tone aptitude (PCID, z-scored) for all items after training on day 2: (a) Estimated pupillometry model estimates from GAMM of PCID on day 2 faceted by z-scored levels of PCID. (b) taVNS group change in pupillometry estimates from GAMM of PCID at day 2. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).

6.2.2.2 Musicality: Self-rated Musicianship

For music experience, there was no ATI observed in the phonological categorization test, shown in Table 19 and Figure 22. There was simply an effect of MCat that applied equally across all groups.

Table 19. GAMM summary table for phonological categorization test X music experience pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Crit/Gnz, Day 1, Sham)	12.529	6.328	.048*
IsDay2	-28.786	7.551	<.001*
IsDay2Peristim	20.558	12.060	.088^
IsDay2Priming	-3.709	12.211	.761
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Crit/Gnz, Day 1, Sham)	17.702	18.466	<.001*
s(MCat) (Crit/Gnz, Day 1, Sham)	1.006	1.006	.615
ti(Time, MCat) (Crit/Gnz, Day 1, Sham)	16.921	24.450	.005*
s(Time):IsDay2	10.449	13.106	<.001*
s(Time):IsDay2Peristim	8.896	11.625	.007*
s(Time):IsDay2Priming	9.863	12.662	.004*
s(X gaze position, Y gaze pos.)	78.143	78.976	<.001*
Random smooth terms	**edf**	**Ref.df**	***P***
s(Time, Participant)	327.616	360.000	<.001*
s(Time, Participant):IsDay2	306.462	359.000	<.001*
s(Time, Item)	133.399	180.000	<.001*
s(Time, Item):IsDay2	127.324	179.000	<.001*
s(Time, Item):IsDay2Peristim	119.700	180.000	<.001*
s(Time, Item):IsDay2Priming	119.223	180.000	<.001*
s(MCat, Item)	13.438	143.000	<.001*

Number of obs.: 637,867; Participants: 72; Items: 36

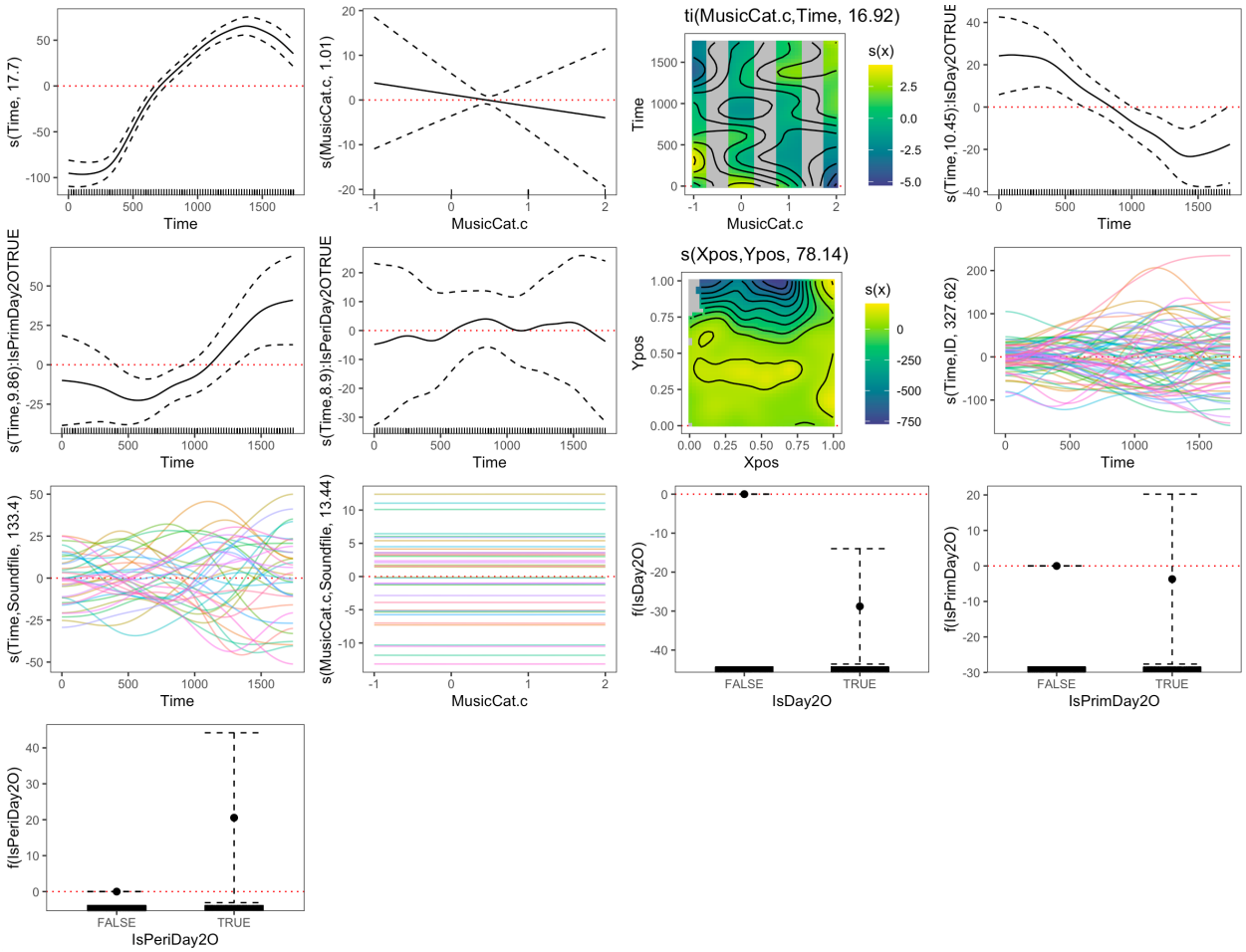
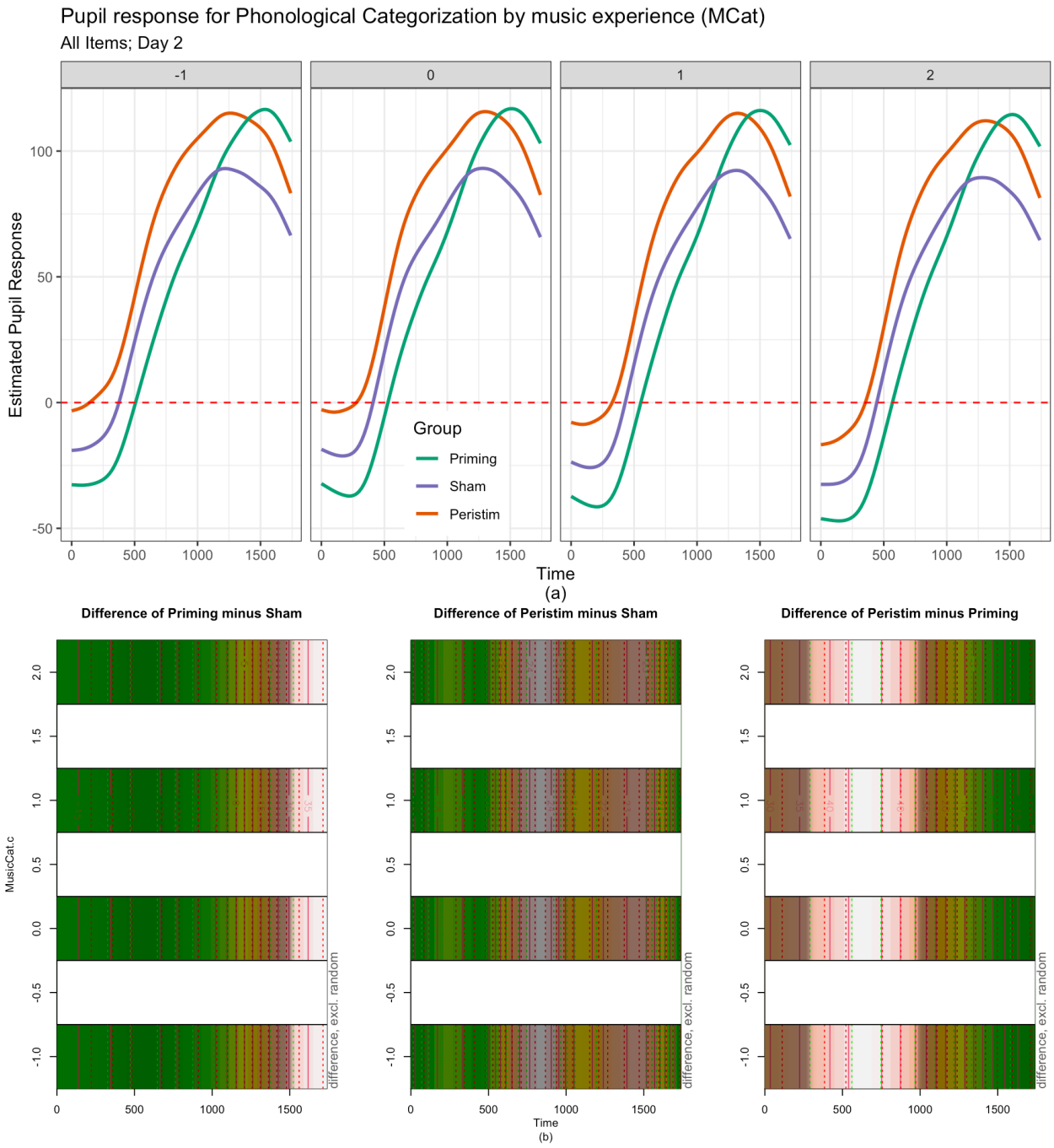


Figure 22. Modeled pupil dilation results from the music experience (MCat) GAMM for phonological categorization. Predicted pupillometry fixed smooth effects and interactions for accurate responses. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. Predicted pupillometry fixed smooth effects for GAMM of phonological categorization pupil response for accurate responses at pre and post. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point.

In Figure 23, the effect of MCat is shown. It can be seen in Figure 23b that the interaction applies equally across all groups and in Figure 23a appears to have a minor effect on the pupil response. At most, it can be seen that, as music experience increases, pupil diameter decreases in the first 500 ms of the pupil response.

Figure 23. Modeled phonological categorization pupillometry results split by taVNS group and level of music experience (MCat, centered at 0 for music-loving nonmusicians) for trained items at posttest: (a) Estimated pupillometry model estimates from GAMM of MCat at posttest faceted by the four self-rated levels of MCat. (b) taVNS group change in pupillometry estimates from GAMM of MCat at posttest. Highlighted areas on the heatmaps represent estimated significant differences between groups, and each slice represents one of the four facets of the pupil response represented in part (a).



There was an ATI observed for the lexical recognition test, shown in Table 20 and Figure 24 with multiple significant ti(Time, PCID) interactions for Priming compared to Sham, Peristim

compared to Sham, as well as Priming and Peristim each at Day 2 compared to Sham at Day 2. Because no effect of Stimulus Condition survived, it appears these ATI effects are consistent across match and mismatch items.

Table 20. GAMM summary table for lexical recognition test X music experience pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Match/Mismatch, Day 1, Sham)	26.392	13.935	.058^
IsPeristim	-37.121	22.979	.106
IsPriming	34.198	22.190	.123
IsDay2	24.608	15.652	.116
IsDay2Peristim	-26.811	26.496	.312
IsDay2Priming	3.765	26.472	.887
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Match/Mismatch, Day 1, Sham)	17.966	18.374	<.001*
s(MCat) (Match/Mismatch, Day 1, Sham)	1.005	1.005	.408
ti(Time, MCat) (Match/Mis.,, Day 1, Sham)	41.628	46.868	<.001*
s(Time):IsPeristim	10.502	13.001	.026*
s(MCat):IsPeristim	1.020	1.020	.300
ti(Time, MCat):IsPeristim	38.367	44.674	<.001*
s(Time):IsPriming	12.940	15.149	.001*
s(MCat):IsPriming	1.004	1.004	.518
ti(Time, MCat):IsPriming	29.775	36.982	<.001*
s(Time):IsDay2	13.144	15.451	<.001*
s(MCat):IsDay2	1.034	1.036	.858
ti(Time, MCat):IsDay2	38.566	44.734	<.001*
s(Time):IsDay2Peristim	12.084	14.474	<.001*
s(MCat):IsDay2Peristim	1.002	1.002	.872
ti(Time, MCat):IsDay2Peristim	36.190	43.125	<.001*
s(Time):IsDay2Priming	11.626	14.042	<.001*
s(MCat):IsDay2Priming	1.862	1.881	.102
ti(Time, MCat):IsDay2Priming	34.720	41.707	<.001*
s(X gaze position, Y gaze pos.)	78.629	78.996	<.001*
Random smooth terms	**edf**	**Ref.df**	***p***
s(Time, Participant)	322.765	354.000	<.001*
s(Time, Participant):IsDay2	314.785	354.000	<.001*
s(Time, Item)	9.300	89.000	.049^
s(Time, Item):IsPeristim	0.270	89.000	<.001*
s(Time, Item):IsDay2	4.339	89.000	.056^
s(Time, Item):IsDay2Peristim	0.237	89.000	<.001*
s(MCat, Item)	0.207	71.000	.103
s(MCat, Item):IsPeristim	.111	70.000	.006*
s(MCat, Item):IsDay2	0.096	70.000	<.001*

Number of obs.: 2,637,160; Participants: 72; Items: 18

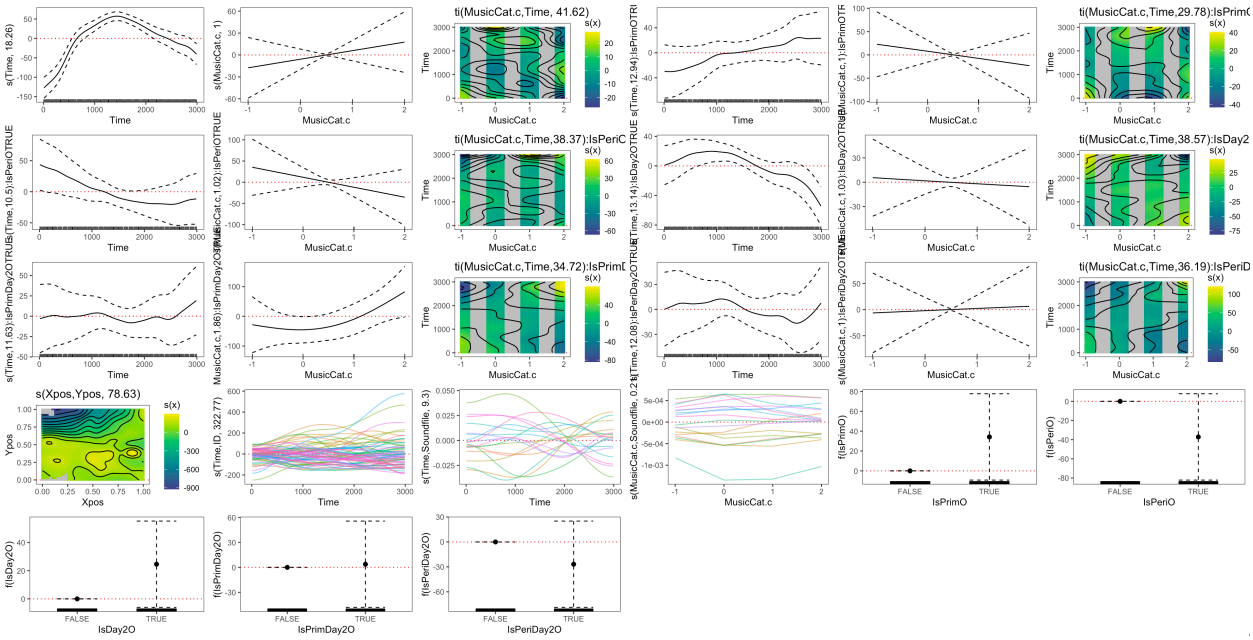
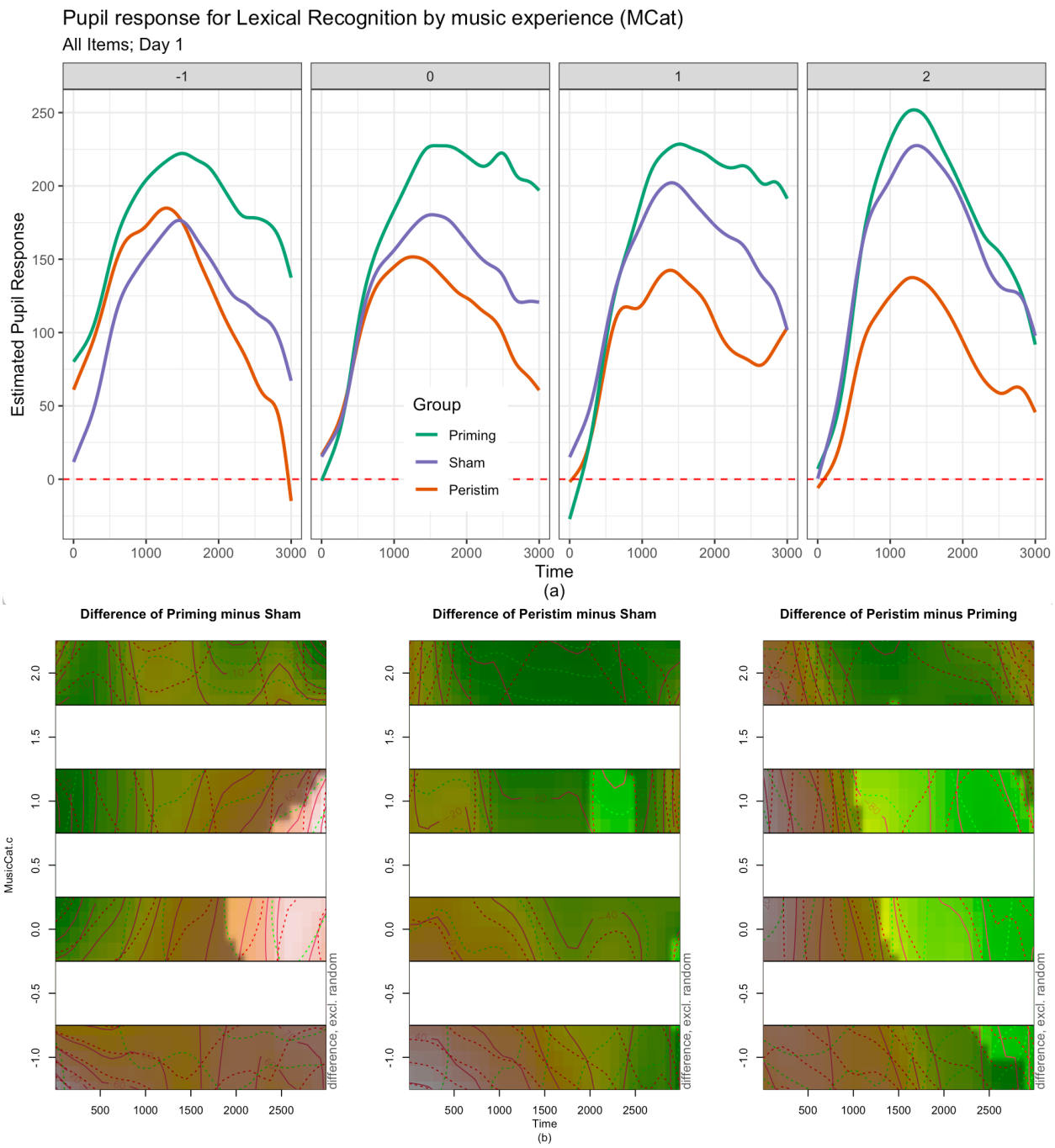


Figure 24. Modeled pupil dilation results from the music experience (MCat) GAMM for lexical recognition. Predicted pupillometry fixed smooth effects and interactions for accurate responses. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. Predicted pupillometry fixed smooth effects for GAMM of lexical recognition pupil response for accurate responses at day 1 and 2.

The tensor product interactions are again complex and laid out more comprehensibly in Figure 25 for lexical recognition on Day 1 and Figure 26 for lexical recognition on Day 2. In Figure 25a, it can be seen that across self-rated musicianship category, the results are relatively consistent. In general, it appears there is a trend for Sham that pupil response raises with more music experience, a trend for Priming to have a high and relatively consistent response across MCat levels, and a trend for Peristim to have a somewhat reduced peak pupil response with increasing experience. The relative groups differences are maintained throughout: Peristim with the smallest pupil response, then Sham in the middle and Priming on top. Looking at Figure 25b, the largest differences between groups show that the ATI effects are driven by a larger pupil

response for Priming compared to Sham from about 2,000 ms to the end of the trial for music-loving non-musicians and amateur musicians. Priming showed an even larger increase over Peristim starting as early as 1,000 ms for those two categories as well as for non-musicians. Peristim's only major difference from Sham was a smaller pupil response from about 2,000 to 2,500 ms for amateur musicians.

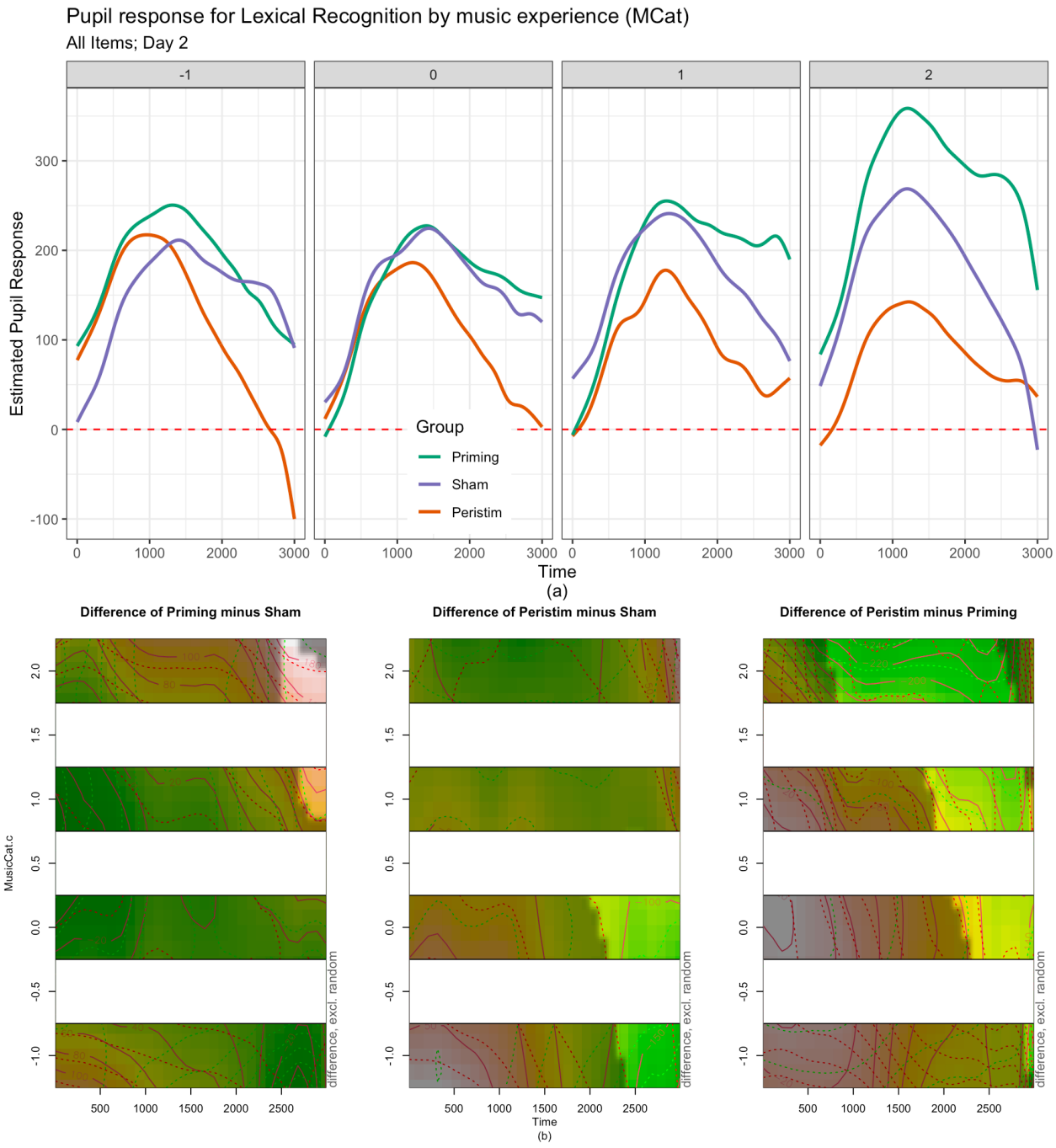
Figure 25. Modeled lexical recognition pupillometry results split by taVNS group and level of music experience (MCat, centered at 0 for music-loving nonmusicians) for all items after training on day 1: (a) Estimated pupillometry model estimates from GAMM of MCat on day 1 faceted by self-rated levels of MCat. (b) taVNS group change in pupillometry estimates from GAMM of MCat at day 1. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).



For lexical recognition after the second training day, the effects show similar patterns, but for different levels of music experience. In Figure 26a, it can be seen that across self-rated

musicianship category, the results are relatively more consistent than for day 1. Priming and Sham now pattern more similarly for both categories of non-musicians and now instead show separation for the musician categories. Both Priming and Sham show a slight upward trajectory for music experience still, while Peristim shows a small decline in the peak response with more experience. Priming and Sham show a larger increase compared to Peristim with increasing music experience. Looking at Figure 26b, the largest differences between groups show that the ATI effects are driven by a larger pupil response for Priming compared to Sham from about 2,500 ms to the end of the trial for amateur musicians and serious amateur musicians. Priming showed an even larger increase over Peristim starting as early as 1,000 ms for serious amateur musicians and from about 2,000 ms for amateur musicians and music-loving non-musicians. Peristim's difference from Sham showed a larger effect size than day 1, but instead of amateur musicians, it showed a smaller pupil response from about 2,000 till the end of the trial for non-musicians and music-loving non-musicians.

Figure 26. Modeled lexical recognition pupillometry results split by taVNS group and level of music experience (MCat, centered at 0 for music-loving nonmusicians) for all items after training on day 2: (a) Estimated pupillometry model estimates from GAMM of MCat on day 2 faceted by self-rated levels of MCat. (b) taVNS group change in pupillometry estimates from GAMM of MCat at day 2. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).


Pupil response for Lexical Recognition by music experience (MCat)
All Items; Day 2
-1
0
1
2
Estimated Pupil Response
300
200
100
0
-100
0
1000
2000
3000
Group
Priming
Sham
Peristim
Time
(a)
Difference of Priming minus Sham
Difference of Peristim minus Sham
Difference of Peristim minus Priming
MusicCat.c
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
difference, excl. random
500
1000
1500
2000
2500
Time
(b)

6.2.2.3 Musicality: Music aptitude

For music aptitude, just like music experience, there was no ATI observed in the phonological categorization test, shown in Table 21 and Figure 27. There was simply an effect of Wing that applied equally across all groups.

Table 21. GAMM summary table for phonological categorization test X music aptitude pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Crit/Gnz, Day 1, Sham)	9.299	6.494	.152
IsDay2	-29.139	7.547	<.001*
IsDay2Peristim	21.892	12.172	.072^
IsDay2Priming	-2.394	12.164	.844
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Crit/Gnz, Day 1, Sham)	17.600	18.39	<.001*
s(Wing) (Crit/Gnz, Day 1, Sham)	2.176	2.24	.219
ti(Time, Wing) (Crit/Gnz, Day 1, Sham)	134.763	183.79	<.001*
s(Time):IsDay2	10.380	13.02	<.001*
s(Time):IsDay2Peristim	9.611	12.35	<.001*
s(Time):IsDay2Priming	9.291	12.00	.018*
s(X gaze position, Y gaze pos.)	77.877	78.96	<.001*
Random smooth terms	**edf**	**Ref.df**	***p***
s(Time, Participant)	310.685	358.00	<.001*
s(Time, Participant):IsDay2	306.376	358.00	<.001*
s(Time, Item)	133.405	179.00	<.001*
s(Time, Item):IsDay2	127.185	179.00	<.001*
s(Time, Item):IsDay2Peristim	120.083	180.00	<.001*
s(Time, Item):IsDay2Priming	119.146	179.00	<.001*
s(Wing, Item)	13.419	179.00	<.001*

Number of obs.: 637,867; Participants: 72; Items: 36

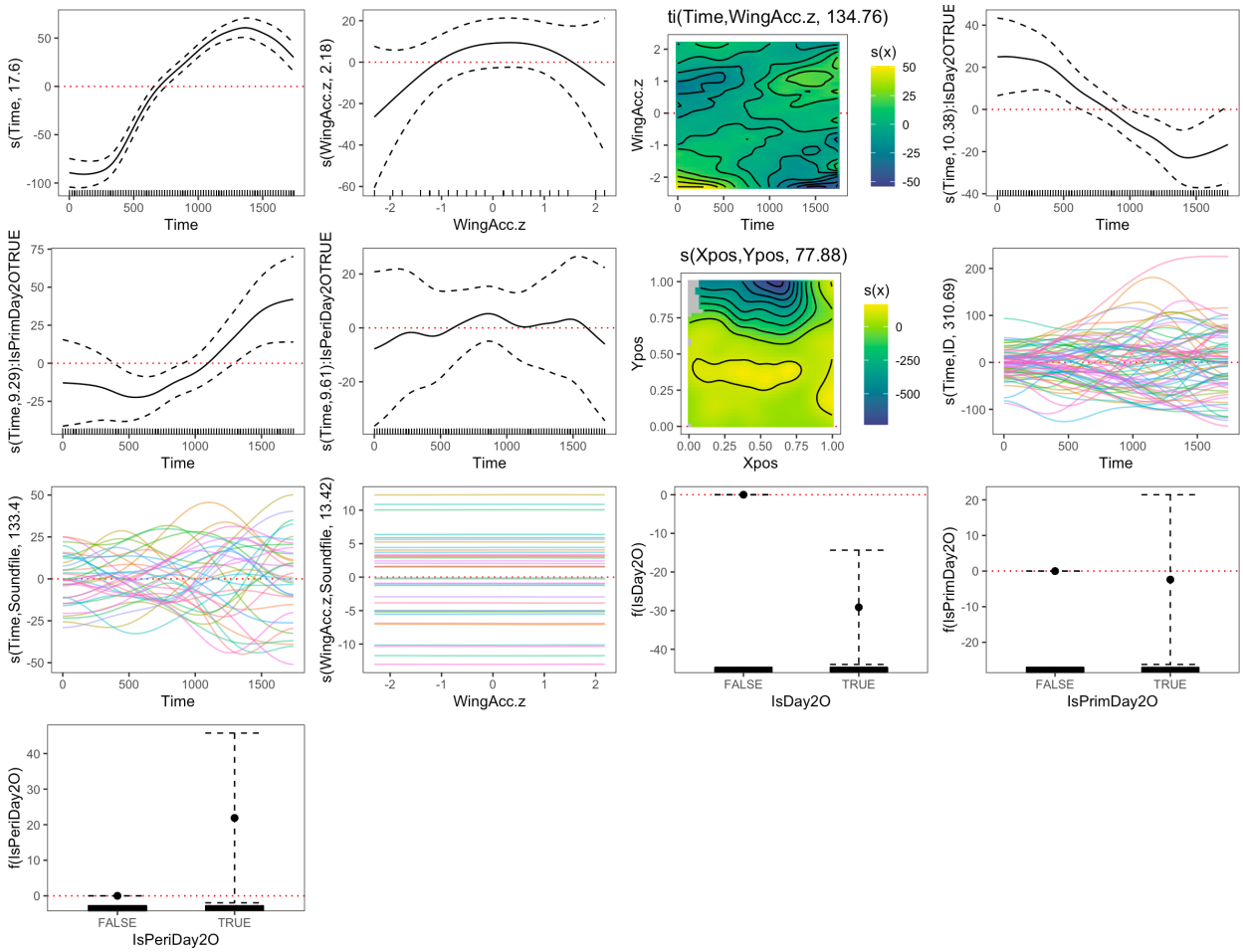


Figure 27. Modeled pupil dilation results from the music aptitude (Wing) GAMM for phonological categorization. Predicted pupillometry fixed smooth effects and interactions for accurate responses. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. Predicted pupillometry fixed smooth effects for GAMM of phonological categorization pupil response for accurate responses at pre and post. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point.

Looking at Figure 28a, we can see that the overall effect of Wing across groups results in a larger pupil response for those with higher music aptitude, although at the very highest levels it drops back down a bit. Looking at Figure 28b, keeping in mind these differences are averaging across not only the other IDs but also any effects of tone observed previously, we overall no

difference between Peristim and Sham, a larger response at the tail for the Priming group vs Sham, and Peristim with a larger response than priming from about 250 to 1,000 ms.

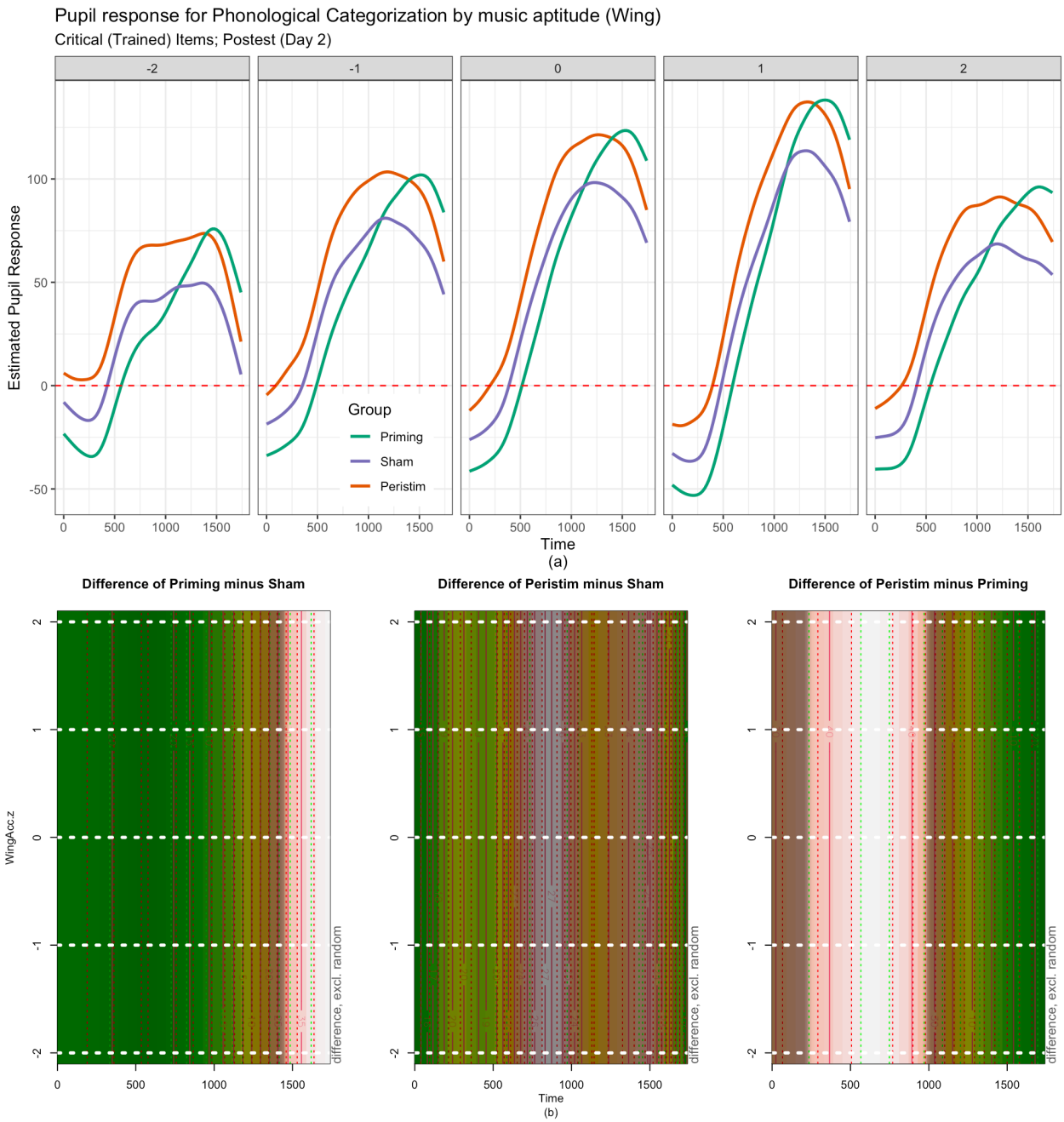


Figure 28. Modeled phonological categorization pupillometry results split by taVNS group and level of music aptitude (Wing, z-scored) for trained items at posttest: (a) Estimated pupillometry model estimates from GAMM of Wing at posttest faceted by z-scored levels of Wing. (b) taVNS group change in pupillometry estimates from GAMM of Wing at posttest. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).

There was an ATI observed for the lexical recognition test, shown in Table 22 and Figure 29 with multiple significant ti(Time, PCID) interactions for Priming compared to Sham, Peristim compared to Sham, as well as Priming and Peristim each at Day 2 compared to Sham at Day 2.

Table 22. GAMM summary table for lexical recognition test X music aptitude pupillometry analysis.

Parametric coefficients	**Estimate**	***SE***	***p***
Intercept (Mismatch, Day 1, Sham)	28.714	11.840	.015*
IsPeristim	-46.027	28.251	.103
IsPriming	49.346	23.659	.037*
IsDay2	24.975	12.940	.054^
IsDay2Peristim	-17.010	26.092	.515
IsDay2Priming	-35.456	27.642	.200
IsDay2PeristimMatch	0.017	0.053	.749
Fixed smooth terms	**edf**	**Ref.df**	***p***
s(Time) (Mismatch, Day 1, Sham)	18.096	18.514	<.001*
s(Wing) (Mismatch, Day 1, Sham)	1.282	1.292	.033*
ti(Time, Wing) (Mismatch,, Day 1, Sham)	224.996	259.072	<.001*
s(Time):IsPeristim	6.176	7.185	.880
s(Wing):IsPeristim	2.730	2.755	.399
ti(Time, Wing):IsPeristim	127.296	147.054	<.001*
s(Time):IsPriming	1.036	1.044	.018*
s(Wing):IsPriming	4.770	4.860	.011*
ti(Time, Wing):IsPriming	146.218	167.679	<.001*
s(Time):IsDay2	11.454	13.740	<.001*
s(Wing):IsDay2	1.022	1.023	.450
ti(Time, Wing):IsDay2	224.860	260.741	<.001*
s(Time):IsDay2Peristim	1.439	1.532	.361
s(Wing):IsDay2Peristim	1.437	1.459	.798
ti(Time, Wing):IsDay2Peristim	120.284	141.353	<.001*
s(Time):IsDay2Priming	1.144	1.169	.116
s(Wing):IsDay2Priming	5.797	5.869	<.001*
ti(Time, Wing):IsDay2Priming	175.793	192.105	<.001*
s(Time):IsDay2PeristimMatch	1.837	2.294	.052^
s(X gaze position, Y gaze pos.)	78.639	78.996	<.001*
Random smooth terms	**edf**	**Ref.df**	***p***
s(Time, Participant)	266.225	360.00	<.001*
s(Time, Participant):IsDay2	249.376	355.000	<.001*
s(Time, Item)	9.409	89.000	.156

s(Time, Item):IsPeristim	0.170	89.000	<.001*
s(Time, Item):IsPriming	0.113	89.000	<.001*
s(Time, Item):IsDay2	4.111	89.000	.016*
s(Time, Item):IsDay2Peristim	0.142	89.000	<.001*
s(Time, Item):IsDay2Priming	0.117	89.000	<.001*
s(Time, Item):IsDay2PeristimMatch	0.502	89.000	<.001*
s(Wing, Item)	0.105	89.000	<.001*
s(Wing, Item):IsPeristim	0.085	89.000	<.001*
s(Wing, Item):IsDay2	0.077	89.000	<.001*

Number of obs.: 2,637,160; Participants: 72; Items: 18

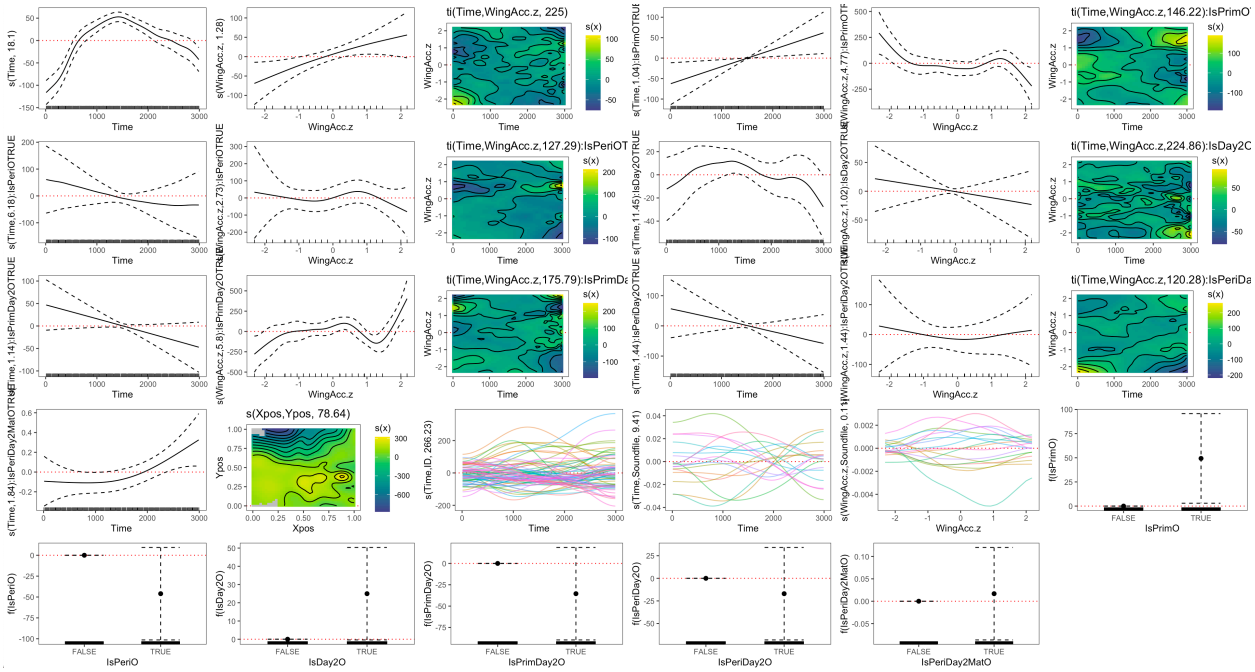


Figure 29. Modeled pupil dilation results from the music aptitude (Wing) GAMM for lexical recognition. Predicted pupillometry fixed smooth effects and interactions for accurate responses. Where the confidence interval is different from zero, there is a significant difference for that effect at that time point. Where the confidence interval for TRUE is different from zero, there is an overall intercept difference for that effect. Predicted pupillometry fixed smooth effects for GAMM of lexical recognition pupil response for accurate responses at day 1 and 2.

The ATI effects of interest are summed and plotted below in Figure 30 for day 1 and Figure 31 for day 2. It can be seen in Figure 30a that the pupil response for Sham shows a gradual increase in pupil response with increasing music aptitude, Priming taVNS varies

nonlinearly across Wing, and Peristim appears to vary the least across Wing. In Figure 30b we can see that the Priming group shows a greater response than Sham at low and high values of Wing, Peristim shows a reduced pupil response compared to Sham at average to just below average levels of Wing, and Priming shows a greater response than Peristim at high levels of Wing and to a lesser degree just below average levels of Wing.

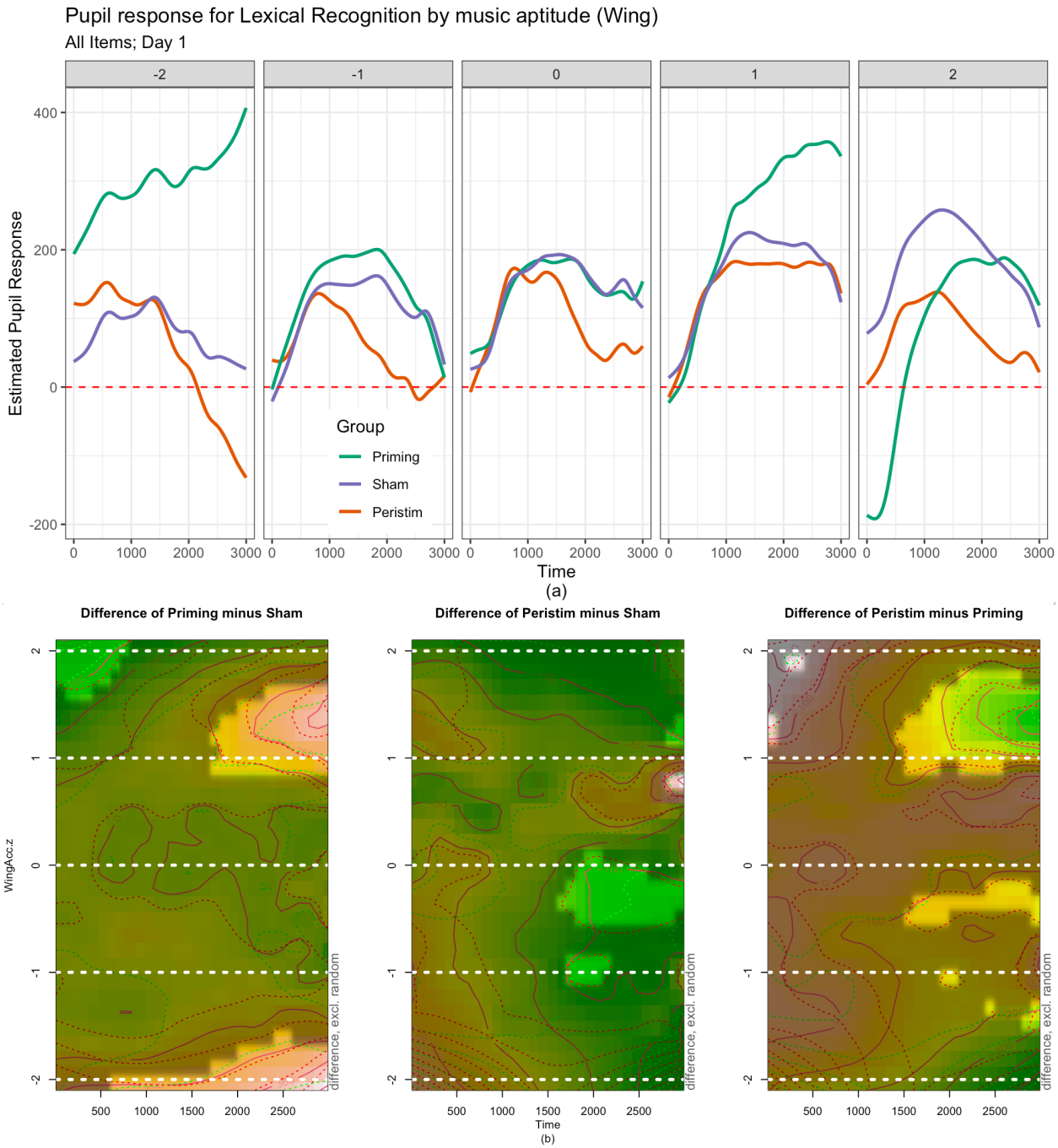


Figure 30. Modeled lexical recognition pupillometry results split by taVNS group and level of music aptitude (Wing, z-scored) for all items after training on day 1: (a) Estimated pupillometry model estimates from GAMM of Wing on day 1 faceted by z-scored levels of Wing. (b) taVNS group change in pupillometry estimates from GAMM of Wing at day 1. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).

In Figure 31a, after the second day of training we see that Sham follows a similar trajectory of peak increases with increasing music aptitude. Priming now follows a similar trajectory and indeed in Figure 31b we can see there are no longer notable differences between Priming and Sham. In Figure 31b, we can also see that group differences between Peristim and Sham in the same location at average and just below average Wing there is a smaller pupil response for Peristim vs Sham, and this is also true for Peristim compared to Priming.

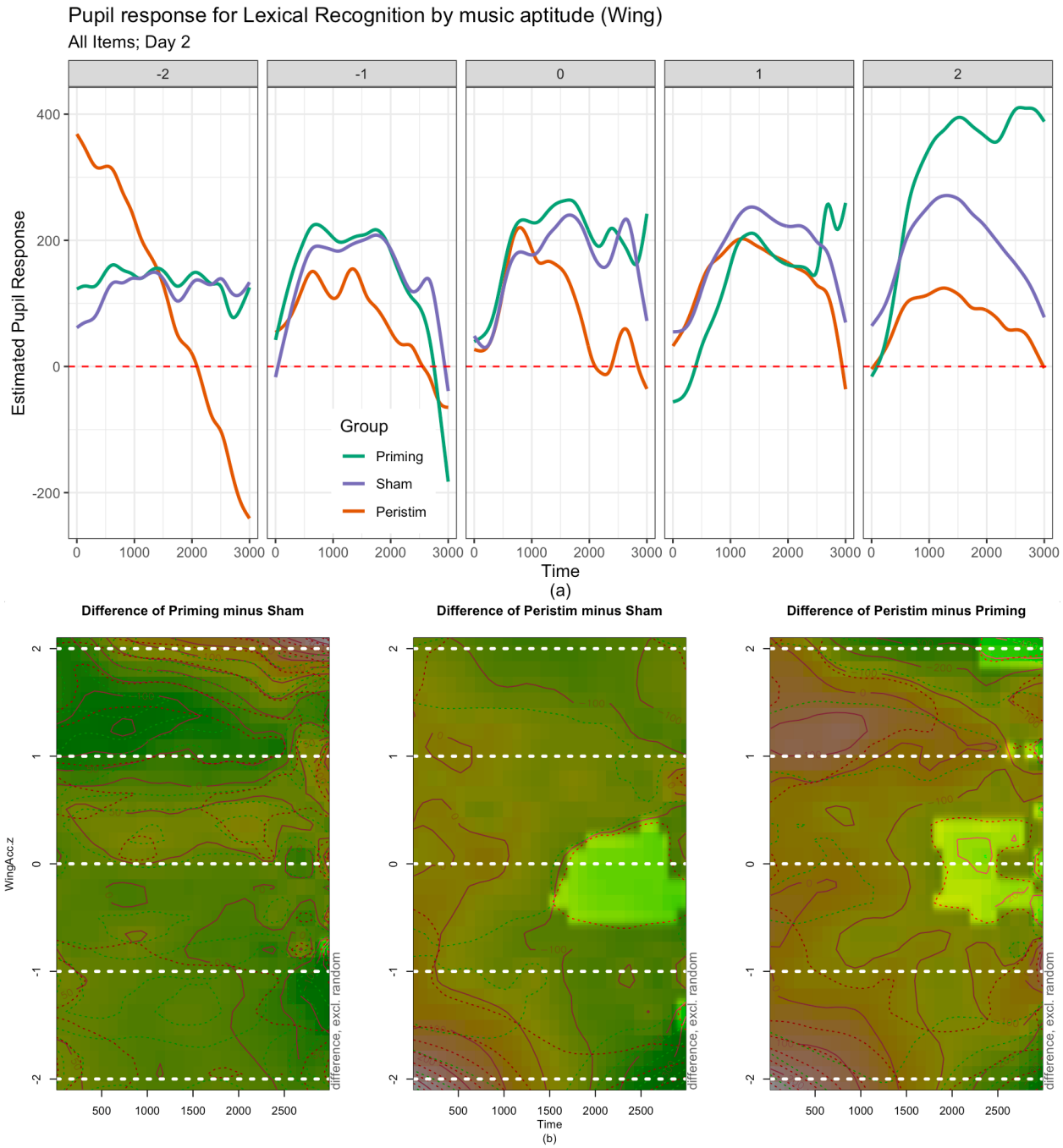


Figure 31. Modeled lexical recognition pupillometry results split by taVNS group and level of music aptitude (Wing, z-scored) for all items after training on day 2: (a) Estimated pupillometry model estimates from GAMM of Wing on day 2 faceted by z-scored levels of Wing. (b) taVNS group change in pupillometry estimates from GAMM of Wing at day 2. Highlighted areas on the heatmaps represent estimated significant differences between groups, and white dotted lines reflect the slices of the pupil response represented in part (a).

6.3 Discussion

Two types of taVNS interventions were observed to elicit performance and cognitive effort enhancements over a sham control differentially across a number of learner individual differences (IDs) in a double-blind study of learning novel tone categories and words with lexical tone. Not only does the effectiveness of taVNS vary with the difficulty of the tonal information to be learned as found in Chapter 5, but here we see the degree to which the effectiveness of taVNS varies is based in part on IDs of the learner. Specifically, the IDs under study were non-linguistic tone aptitude (measured by the PCID task), music experience (measured by a self-rated score, MCat), and music aptitude (measured by the Wing), three IDs shown in prior literature to be highly predictive of lexical tone learning outcomes.

First, in looking at the bigger picture, it is easy to see that none of the IDs had a negative impact on outcomes for any group. All significant slopes were positive for accuracy and negative for RT, meaning higher ID ability meant better accuracy or faster RTs when significant. Next, if we consider the existing literature on these IDs for Mandarin tone learning (e.g., Bowles et al., 2016; Pelzl, 2019), they are reproduced in the sham control group results here: musicality (music experience and music aptitude) is only significant for accuracy outcomes, not reaction time, and non-linguistic tone aptitude is significant for both accuracy and reaction time outcomes. While music aptitude hasn't been predictive of RT results previously and isn't technically predictive here, we do see an interesting surviving nonsignificant effect wherein the effect of music aptitude is collectively null but does significantly vary by item. This is a novel finding that showed up for both phonological and lexical RT results and may be indicative of further research

needed to tease apart specific types of items and conditions for which music aptitude may be predictive of tone-related RT.

To answer the research question on whether taVNS-facilitated behavioral outcomes are moderated by individual differences, there were multiple significant aptitude- and experience-by-treatment interactions. The ATIs observed for accuracy were varied across both phonological categorization and lexical recognition, across peristim and priming taVNS, and across non-linguistic tone aptitude, music experience, and music aptitude. Peristim taVNS showed ATIs relating to non-linguistic tone aptitude, music experience, and music aptitude for accuracy, and no ATIs related to RT. Priming taVNS showed ATIs relating to non-linguistic tone aptitude and music aptitude for accuracy, and related to music experience for RT. This last finding is particularly interesting given that music experience is not typically associated previously with better RTs. Priming taVNS shows this advantage only for those with more music experience for the easier conditions of trained tone items and match trials for lexical recognition. Dittinger et al. (2016), upon which the present phonological categorization and lexical recognition tasks were based, does show an effect of musicianship for accuracy on both equivalent tasks, but also an RT advantage for phonological categorization for musicians. It's worth noting that the musicians in Dittinger et al. (2016) were professional musicians with an average of 17 years of instrument practice, while the highest level of music experience in the present corpus was a 'serious amateur musician'. Thus, after training, priming induced an RT effect and peristim induced an accuracy effect in amateur musicians that parallels related effects found for professional level musicians.

The findings are diverse, but a few patterns do emerge. For one, accuracy outcomes are more impacted than RT outcomes by ATIs for active taVNS vs sham. Secondly, it appears the

majority of ATI effects were observed for lexical recognition over phonological categorization. This difference may be due to a variety of reasons, not the least of which being a lack of active taVNS stimulation during or right before the phonological test while the lexical recognition test had stimulation during or right before. It could also be a result of more test trials for lexical recognition or that there is no pretest for lexical recognition like there is for phonological recognition to effectively cancel out the interpreting of any preexisting group differences. Yet another explanation could also be the added complexity in the lexical recognition test being an additional source of variation: at issue for the phonological categorization test is simply fuzziness in the phonological domain while the lexical recognition test reflects fuzziness for both the phonological domain and the phonolexical mapping of new phonology on preexisting semantic representations. It is possible this form-to-meaning mapping is more strongly benefitted by taVNS and, indeed, in Chapter 5 we saw high levels of performance for phonological categorization on T1 and T2.

Both peristim and priming taVNS showed ATI effects on accuracy, but only priming showed ATI effects on RT. Especially interesting in light of the results in Chapter 4 and 5 is the fact that there are multiple accuracy improvements for priming taVNS revealed here where they were not previously, particularly at high non-linguistic tone aptitude on day 1 for match items, and even mismatch items on both days for those with low music aptitude. It's clear that ATI analyses allow for a much more nuanced investigation of the effects of a particular treatment.

All three individual differences, non-linguistic tone aptitude, music experience, and music aptitude, revealed ATI effects, both by the pattern of producing a relationship that was not present for sham to enhance high aptitude/experience learners and also by obviating ID effects,

enhancing the performance of low-aptitude learners and putting them on the same level as high-aptitude learners. Only music experience did not show the latter pattern as it was not related to sham performance in this sample. Further, the ATI effects on accuracy for all three IDs were roughly in the same magnitude, each showing effects that varied from about a 5% to a 30% improvement in probability of an accurate response over sham. Previous research into ATI effects with Mandarin tone learning has primary looked at contrasting high versus low variability training, one exemplar being Perrachione et al. (2011), which found similar ATI patterns as observed here, a larger boost for low-aptitude learners when paired with low-variability training and a larger boost for high-aptitude learners when paired with high-variability training. The results in the present dissertation don't readily fall on so clean an interpretation given the entirely low-variability design and the presence of both types of effects, but future research may benefit from an investigation of taVNS to supplement different types of instructional interventions, such as low-aptitude learners in a high-variability training.

To answer the research question on whether IDs moderate the effects of taVNS on the deployment of cognitive effort, there were also multiple significant aptitude- and experience- by-treatment interactions. The ATIs were again varied across phonological categorization and lexical recognition. Here again, similar to behavioral outcomes, lexical recognition showed more evidence of ATIs than phonological recognition. Case in point, the only ATI observed for cognitive effort on the phonological categorization test was for non-linguistic tone aptitude for the peristim group, whereas for lexical recognition there were effects for all tested IDs. The ATI effect for phonological categorization shows that, while the effort for the sham and priming groups seems to gradually increase with higher non-linguistic tone aptitude, the peristim group

has a somewhat reversed effect: cognitive effort starts out high for the peristim group at low aptitude and generally declines (if a little nonlinearly) toward lower effort at higher aptitude. For non-linguistic tone aptitude in lexical recognition on day 1, the peristim group shows consistent effort across levels of aptitude while the priming group shows relatively consistent effort but higher overall than peristim, and especially sham at lower levels of aptitude; by contrast, the effort for the sham group gradually increases as aptitude increases. After the second day priming and sham are both gradually rising while peristim effort gradually declines with increasing aptitude, and overall group differences are muted compared to day 1. Music experience shows very similar effect across days, the one exception being, instead of a flat peristim effect on day 1, on both days effort for the peristim group declines with increasing experience. For music aptitude and lexical recognition, the ATI terms in the model were significant, but the interpretability of the results are less clear: the effects of priming are inconsistent, and the effect of peristim appears flat across levels of music aptitude while sham shows an increase in effort on day 1, but all groups are largely showing flat trajectories on day 2.

In sum, in addition to the ATI effects for behavioral outcomes there were also ATI effects observed in the degree of cognitive effort deployed across tasks, depending on the type of taVNS applied. The sham group overall showed a trend of increased effort as the aptitude or experience increased, which is counterintuitive given that better aptitude and experience corresponded to similar or improved performance, but may reflect more cognitive resources being able to be brought to bear on the task at hand. It may thus be the case that, while music experience may have led to increased, more robust—or effortful—pitch processing or auditory memory (e.g., Patel, 2011), this did not lead to better tone learning accuracy due to the short duration of the

training and the low number of words to learn. Given the results for a similar lexical recognition test used in Dittinger et al. (2016), it may also be the case that sham participants were still actively trying to learn during the lexical recognition test despite the lack of feedback.

When ATI effects occurred in the physiological data, the pattern was the following: over the course of an individual difference variable peristim taVNS resulted in a reduction of effort compared to sham and, by and large, priming taVNS resulted in an increase in effort compared to sham. This result for peristim, in light of comparable if not better accuracy than sham, indicates that peristim taVNS led to better encoding of words during training, such that less effort was needed in the process of recognizing previously learned words. This result for priming taVNS, also in light of comparable if not better accuracy than sham and comparable if not better RT than sham, suggests that priming taVNS may have led to a more optimal state of arousal for exerting mental effort on the tasks (Aston-Jones & Cohen, 2005), which resulted in better accuracy and/or reaction time.

The pupillometry results do augment the behavioral findings for lexical recognition in a few ways. The priming group producing an effect of non-linguistic tone aptitude for match items in lexical recognition parallels some evidence for a decline in effort in the early and later parts of the trials for those with higher aptitude, even if priming engenders a large pupil response vs sham at lower levels of aptitude (and indeed, for the other IDs), and the reduction in effort at higher levels is reduced to a trend on day 2. The peristim group producing an accuracy effect of music experience in lexical recognition match items parallels a decrease in cognitive effort for peristim at higher levels of experience while the sham group increases in effort, and the trajectory of these behavioral and physiological ATIs was maintained on day 2. The priming

group produces an RT effect of music experience on both days of lexical recognition match items, and in general priming shows a larger pupil response compared to sham, particularly for music-loving non-musicians and amateur musicians on day 1, and this effect shifts up to amateur and serious amateur musicians on day 2.

Unfortunately, there is no clear corollary between the behavioral ATI effects and the physiological ATI effects. For one, the phonological categorization task did not have any conclusive behavioral ATI effects for non-linguistic tone aptitude. Non-linguistic tone aptitude was the only ID for which there was a physiological ATI effect for phonological categorization. In light of the differential effects of tone difficulty observed in Chapter 4, it is possible that this Chapter essentially averaged across disparate effects resulting in few clear interpretations. Inconclusive ATI effects for behavioral outcomes on phonological categorization are also due to group differences at pretest before any taVNS was applied, suggesting future research should attempt to balance on pretest scores in addition to the IDs before group assignment. In addition, neither set of pupillometry analyses included models in which all three IDs were included simultaneously due to the already complex GAMMs, but ideally the IDs would be controlling for each other and potentially clarifying effects as they did for the behavioral results. Interestingly, the clearest pupillometry results came out for music experience, the least continuous measure of the three IDs under investigation. This may lead to new a methodological consideration for GAMMs: maybe there is such a thing as too many knots in the parameterization of tensor product interaction; that is, maybe the potential complexity of continuous nonlinear interactions should be *a priori* constrained in certain scenarios where we don't think the results should be overly complex to reduce spurious pockets of significant differences at a granularity that isn't

interpretable practically. This would also be one way to reduce the complexity of a GAMMs model with many estimated tensor product interactions. The complexity of the statistical modelling for both behavioral and physiological data in this Chapter is quite high, so future research that seeks to test ATI performance more explicitly across different tones at the phonological or even lexical level may benefit from an *a priori* power analysis based on the present exploratory analyses to maximize the power to detect effects in such a complicated paradigm.

Even after only one day of training for naïve learners of tone, taVNS effected multiple aptitude- and experience- by-treatment interactions on lexical recognition outcomes, most of which persisted into the second day of training. What is more, after two days of training, taVNS also effected an ATI on a phonological categorization posttest for both taVNS groups, even though participants were not stimulated during or directly before the posttest. The observed ATI effects manifested in a combination of either (1) buttressing those with low aptitude to perform as if they had high aptitude or (2) giving an advantage to high aptitude learners where there wasn't one previously. As we've seen in each Chapter, peristim and priming taVNS result in disparate learning benefits. Once again, despite peristim having less total stimulation duration, we see more effects of improved accuracy outcomes relating to peristim (10 ATIs for peristim, 6 for priming) while priming shows more benefits to RT than peristim (3 ATIs for priming, 0 for peristim).

6.4 Conclusions

The current Chapter's results further expand the scope of those in Chapter 4 and Chapter 5 to include nuance with the varying effects of learner individual differences. After a two-day training, active taVNS showed a combination of effects allowing learners to either capitalize on their high aptitude and experience or to level the playing field for learners with low aptitude to enhance Mandarin tone learning outcomes at the phonological and lexical levels. Both peristim and priming active taVNS administrations produce differential types of impact on learning outcomes and cognitive effort deployed at test. Both administrations also produce these ATI effects for both phonological and lexical accuracy, reaction time, and effort, the one exception being priming on phonological categorization effort.

While there were many results of two types of ATIs, few clear patterns emerged. It is possible results are so varied because some are spurious. Given the complexity of the analyses and the relatively low sample sizes for the active taVNS groups ($n < 20$ each), a larger sample is warranted in replicating, extending, and clarifying these results. Additionally, the varied results may also reflect an even more complex story. Given the degree to which tone difficulty impacted phonological categorization outcomes in Chapter 5, it's possible that being able to split ATIs by tone would further elucidate results. Doing so would be more analytically complex, but it does seem clear from the pupillometry results here that some of the effects observed for tones in Chapter 5 have likely been averaged over, resulting in an estimate of effort that is not reflective of easy and difficult items equally.

Nonetheless, the spirit of Ingvalson et al. (2011) has been confirmed in this Chapter; it's clear from these results that, over the long term, a greater understanding of learner individual

differences in Mandarin tone learning will enable neurostimulation interventions to be tailored to achieve the greatest benefits for each individual learner. While the paradigm here was a two-day training intervention for *ab initio* learners and showed a proof-of-concept, further research must look at longer term interventions, delayed posttest outcomes, and more ecologically valid learning targets such as real words with all four Mandarin tones and multisyllabic words. Given the range of potential ATI effects for neurostimulation and IDs observed in this Chapter, it's possible that early incorporation of taVNS for learners with both high and low non-linguistic tone aptitude and musicality may be beneficial, but another potential use case may be later incorporation of taVNS. Indeed, of the many Mandarin tone training studies that show learners plateauing in performance below criterion (e.g., Bowles et al., 2016; Chandrasekaran et al., 2010; Li & DeKeyser, 2017; Wong & Perrachione, 2007), it would be insightful to replicate such a paradigm measuring non-linguistic tone aptitude and musicality, and then apply taVNS to those underperforming learners. Likewise, applying taVNS early in such a design may help avoid the plateau effect entirely.

Non-linguistic aptitude and musicality have a strong relationship to Mandarin tone learning and processing, and this Chapter has shown that the influence of both is fungible by way of neurostimulation. While the results of this Chapter are exploratory, modifying the effects of these individual differences in a safe, non-invasive, low-cost way to achieve up to a 30% gain in accuracy after only a day or two of training lend further credence to the potential for a very practical applicability for these taVNS interventions being integrated into real-world learning environments in the future.

Chapter 7: General Discussion

7.1 Summary of Findings

Mandarin lexical tone learning has repeatedly been identified as a difficult linguistic feature for non-native speakers of tonal languages like English, even for native English learners of Mandarin at high proficiencies (e.g., Pelzl et al., 2019b). Sound perception training has been shown to help native English speakers perceive lexical tone differences, but acquiring lexical tone as a feature still remains difficult, even after as many as 18 training sessions (Bowles et al., 2016; Chandrasekaran et al., 2010; Li & DeKeyser, 2017; 2019; Liu & Chandrasekaran, 2013; Wang et al., 1999; 2003; Wong et al., 2011; Wong & Perrachione, 2007). This dissertation investigated transcutaneous auricular vagus nerve stimulation (taVNS), a type of non-invasive peripheral nerve stimulation delivering electrical current to the ear canal, and its potential impact as tool to enhance Mandarin tone learning.

The logic for taVNS comes from incipient research showing a connection between positive effects on learning and memory (Jacobs et al., 2015) and plasticity in the primary auditory cortex for pure tone learning (Borland et al., 2018; Engineer et al., 2011; Kilgard, 2012). Priming administrations of taVNS are thought to operate via inducing tonic shifts in arousal prior to a task and thus cortical excitability to optimize subsequent learning (Groves et al., 2005). Peristimulus administrations of taVNS are thought to operate via inducing phasic changes in task-related attention and consolidation of specific pieces of information. Both types of administration were investigated here.

Participants in three groups, peristim taVNS, priming taVNS, and a sham taVNS control participated in a double-blind two-day Mandarin phonological and lexical tone training study. Behavioral data including accuracy and RT were collected, as was physiological data in the form of pupillometry due to its ties both to cognitive effort and the most well-studied VNS mechanism of action, the production of norepinephrine. Active taVNS groups received stimulation before or during multiple training and testing tasks across the two days.

Initially, primary learning outcomes were assessed for lexical tone learning: accuracy and RT on lexical recognition and recall tests. Peristim taVNS resulted in better accuracy outcomes while priming taVNS resulted primarily in better RT outcomes. Cognitive effort was assessed during a passive word learning task on both days. From day-to-day, we observed a reduction in effort for peristim taVNS and a small increase in effort for priming taVNS compared to sham to reach similar or improved behavioral outcomes compared to sham. The results are suggestive of different underlying mechanisms: peristim enhances the early encoding of individual items and results in less effort long term while priming enhances the overall state of arousal, optimizing the resources that the learner can bring to bear on the task and speeding up lexical retrieval.

Next, learning outcomes were assessed for phonological tone learning: accuracy, RT, and cognitive effort on a phonological categorization pretest and posttest. Peristim again resulted in better accuracy outcomes on trained and untrained (speaker and syllable) items while priming resulted in more robust RT outcomes for trained items. Interestingly, peristim showed only a slightly smaller magnitude improvement than priming for trained items, but peristim was the only group to show RT improvements on untrained items. These effects were moderated by tone difficulty with stronger effects of taVNS observed for more difficult tones. Analyses of cognitive

effort confirmed a difficulty effect at pretest, with T4 more difficult with a larger pupil response than T2, in turn more difficult than T1. Interestingly, opposite to the lexical learning task, here priming taVNS resulted largely in less effort over the course of the posttest items, and peristim showed similar effort to sham.

Finally, learning outcomes for both phonological and lexical tone learning were assessed with respect to individual differences of non-linguistic tone aptitude, music experience, and music aptitude, three variables with a history in the tone learning literature. The extent to which the effects of these IDs were moderated by taVNS interventions was assessed with accuracy, RT, and cognitive effort on the phonological categorization pretest and posttest and the lexical recognition test on both days. Results revealed a complex picture of aptitude- and experience-by-treatment interactions, in which peristim and priming taVNS each resulted in a combination of effects to reflect taVNS-related improvements for either low aptitude learners to make their outcomes comparable to high aptitude learners or for high aptitude learners to improve their outcomes where they were previously no different from low aptitude learners. There were also a number of interactions with cognitive effort, largely showing peristim taVNS to result in a reduction of effort with increasing aptitude or experience for both phonological and lexical tests, but also some showing priming taVNS to produce a larger overall response compared to sham for the lexical recognition test. Interestingly, pupillometry results for the lexical recognition test patterned with the passive word learning task: peristim showed reduced effort while priming showed increased effort compared to sham. By contrast, in general the trend for phonological categorization was the reverse: less effort for priming and more for peristim compared to sham.

Combining behavioral results, there are some notable connections across the chapters. The peristim accuracy advantage observed in Chapter 4 for the mismatch condition in lexical recognition appears to have been driven largely by peristim taVNS improving those with low non-linguistic tone aptitude. The small accuracy advantage observed in Chapter 4 for priming was apparently driven by those with high non-linguistic tone aptitude and low music aptitude. The priming taVNS advantage for RT found in Chapters 4 and 5 for both phonological and lexical outcomes was driven by those with more music experience receiving priming taVNS, allowing them to recognize tones and tone words more quickly. Better accuracy for peristim on T4 in Chapter 5 was likely driven by peristim obviating the effect of music aptitude, bolstering the accuracy of those with low music aptitude. Participants that receive priming taVNS largely show more effort than sham for lexical tasks but less effort for trained materials for a phonological task, while participants that receive peristim taVNS show even stronger evidence for more effort than sham for a phonological task but less effort for lexical tasks.

In sum, this dissertation has revealed: (1) priming and peristim administrations of taVNS facilitated vocabulary learning of words with Mandarin tone, (2) priming and peristim administrations of taVNS facilitated learning of new phonological tone categories, and (3) the effects of individual differences were substantially and differentially impacted by priming and peristim administrations taVNS. These conclusions resulted from analyses of behavioral and physiological data and examine accuracy, reaction time, and cognitive effort across a two-day training. Simply put, the evidence herein supports taVNS as a practical treatment intervention for enhancing language learning and reveals a number of considerations for its use and implementation.

7.2 Fuzzy Phonolexical Representations are Mitigated by taVNS

Perhaps the most straightforward underlying driver of tone word learning difficulty observed in this dissertation is that *ab initio* Mandarin tone word learning for native speakers of English resulted in early and underspecified phonolexical representations. The fuzzy lexical representation hypothesis (Gor et al., 2021) posits that L2 learners often misinterpret the surface form of a word with a different word given ambiguity in the phonological mapping of word's mental representation. The training design of the current corpus consisted of nine tone pseudowords, each of which formed minimal pairs by consonant, vowel, and tone with other pseudowords within the nine. This tightly controlled but artificial design may have maximized the fuzziness of any representations formed for these *ab initio* learners given their still low familiarity with these new items (see Cook & Gor, 2015; Gor & Cook, 2020). Indeed, in line with the OM (Bordag et al., 2021; forthcoming), it appears likely that participants in this study, regardless of group, did not reach their optima for phonological or phonolexical encoding after the two-day training as evidenced by the combination of (1) effects of learning observed day-to-day, and (2) less than ceiling performance on tests of phonological categorization and lexical recognition and the end of those two days for only nine monosyllabic pseudowords. It is likely that with additional days of training many participants could have continued to improve toward their optima, allowing the fuzziness of the new representations to further resolve. Although minimal pairs as a source of fuzziness may be reversed to become a source of enhanced sensitivity to L2 contrasts at higher proficiencies (Bundgaard-Nielsen et al., 2011; Llompart &

Reinisch, 2020; Wiener et al., 2019), Pelzl et al. (2021a; 2021b) found evidence of fuzzy tone word representations in even advanced learners of Mandarin. Pelzl et al. (2021a) observed fuzziness to arrive due to both encoding and processing deficits, suggesting successful encoding could still result in subpar processing, and these findings have interesting resonance with the results of this dissertation.

Peristim taVNS is posited to be related to learning and consolidation of specific information, and time and again across this dissertation the results suggest peristim taVNS has facilitated encoding of information, resulting in better accuracy and a reduction of cognitive effort. The best piece of evidence for this is the phonological categorization task, before and during which participants are not stimulated, yet still the peristim participants showed the greatest evidence of better encoding the information in the training: not only did peristim participants show the best accuracy improvements across trained and untrained items, they also showed the only improvements in reaction time for untrained items, suggesting more robust learning of some tonal categories compared to the other groups. Peristim taVNS, despite being implemented in a combination of training and testing tasks in the present corpus, may be showing large effects in efficacy at test precisely because it has been implemented during the training, or encoding, tasks, thus producing measurable recognition improvements at test.

Priming taVNS is posited to be related more broadly to tonic shifts in arousal and attention with the idea of placing someone in an optimal state for learning. However, despite priming participants receiving more total stimulation than the peristim group, they did not show as conclusive a benefit on accuracy, and usually not to the same magnitude when they did. However, priming taVNS showed the most consistent and robust reaction time speedups for

recognition of trained items, but notably not untrained items. These findings suggest that the optimal use for priming taVNS may not be for learning new information, but for accessing *already learned* information.

Thus, a logical follow-up to the present dissertation would an extension of the present design with four groups: (1) a sham taVNS control, (2) peristim taVNS throughout training and testing tasks, (3) priming taVNS before training and testing tasks, and (4) an active taVNS group in which peristim is applied only during learning tasks and priming is only applied during recognition tests. In this way, we could attempt to improve encoding for new information and mitigate processing deficits for recognition of encoded information.

Additionally, in light of the OM and given that form encoding needs to reach its optimum for a lexical entry to function properly (Bordag et al., 2021), an additional component of a revised design could include phonological tone categorization training to reach some criterion (rather than a fixed amount of training for each person) before introducing lexical learning. It may clarify the results of peristim's role in encoding (by perhaps leading participants more quickly to their optima for the individual items), and it could allow an evaluation of group differences in the lexical recognition test when the phonological categorization component of the phonolexical mapping is held relatively more constant (i.e., at a similar place on the ontogenetic curve for phonological tone word form) than in the present design. Given the present design's use of English words with similar lexical characteristics in a controlled pseudoword training, the semantic ontogenetic curves are already held constant at their optima in this design as it is solely based on L1 semantic representations in one-to-one translation (Bordag et al., 2021).

The aptitude- and experience- by-treatment interaction effects observed in Chapter 6 underline the need to incorporate relevant individual differences in future investigations of taVNS interventions as well. A recent study investigated individual differences that affect L2 phonological encoding to mitigate fuzzy phonolexical representations (Daidone & Darcy, 2021). The authors found encoding to be enhanced by L2 vocabulary size at higher proficiencies and phonological short-term memory (PSTM) at lower proficiencies. Although they admit the impact of individual differences may depend on the contrast(s) under examination, PSTM is a compelling ID to include in future investigations of fuzzy lexical representations for early learners of Mandarin tone. The authors only found PSTM to be predictive for a phonological contrast along a dimension not used in the native language (similar to lexical tone here), and that those with higher PSTM are able to hold the relevant phonetic details in memory longer, assisting in transfer to long-term representations. While individual differences are well-established factors that have been shown to mitigate the fuzziness of Mandarin tone word learning and research has pushed forward in the direction of aptitude-by-instructional-treatment interaction studies, this dissertation has opened pandora's box by showing preliminary evidence for aptitude-by-neurostimulation-treatment studies. Should the effects in this dissertation persist upon further study, it is possible the picture could become even more complex: if taVNS interacts with aptitude in a low-variability training design like this one, would it do so in the same way and to the same degree for a high-variability design?

7.3 Practical and Pedagogical Implications for taVNS Interventions

The results of this dissertation reveal a number of implications for incorporating taVNS interventions to enhance Mandarin tone learning. Firstly, the results herein show that taVNS can provide an accuracy improvement for difficult tone learning across a two-day training, as much as a 30% improvement in accuracy over a sham control for both phonological tone contrasts and tone word learning after only a two-day training. The findings of this dissertation highlight the potential for the future utility of practical Mandarin tone word training for native speakers of non-tonal languages being facilitated by taVNS. Because Mandarin learners have been observed to so often plateau in learning lexical tone over multiple training sessions (e.g., Bowles et al., 2016; Chandrasekaran et al., 2010; Li & DeKeyser, 2017; Wong & Perrachione, 2007), this dissertation shows taVNS to be a promising intervention for overcoming lackluster ultimate attainment. A longer training study over more days than two is needed in order to examine whether taVNS is helping learners get closer to ceiling performance but minimally taVNS may be practically useful for enhancing the encoding of Mandarin tone word phonolexical representations generally allowing for more efficient lexical retrieval when encountered post-training.

Considering the results of this dissertation in total, it's clear that further research extending the findings here may help formalize specific recommendations with respect to when during learning, what aspects of learning, and what type of learner will most benefit from a taVNS intervention. At least for Mandarin tone learning, it appears that the hardest aspects of an already difficult new contrast may benefit the most from a taVNS intervention, in this case harder to identify tones (T4 > T2 > T1), tone word recognition in difficult conditions (mismatch

> match), and tone word recall from memory. While investigations of interactions with individual differences show a combination of taVNS effects that benefit either low ability learning or high ability learners, the investigation here was novel and exploratory. Minimally it's this benefit for low ability learners that may ultimately show utility for taVNS in overcoming plateau performance below criterion after repeated Mandarin tone training sessions. In the end, future investigations are needed to better hone in on when one group of learners may benefit versus the other.

While implementing a tVNS study with a consumer-grade device is relatively straightforward, determining the stimulation parameters (e.g., priming vs. peristim, shape of the stimulation waveform, duration of stimulation) that lead to optimal learning benefit is still largely unknown. While the corpus studied here used a research-grade stimulator for precise control of the stimulating waveform and required some training, the earbuds are commercially available as is a handheld stimulation device that is easy to use with more limited waveform options, which was not used here (costing under $500 total).

Finally, an important consideration of the results in this dissertation is that taVNS was an effective intervention even though stimulation was delivered 0.2 mA below a user's perceptual threshold in this corpus. Results from a recent systematic review of tVNS safety (Redgrave et al., 2018) additionally confirm that the most common side effects of tVNS are skin irritation at the electrode site (18.2%), headache (3.6%), and common cold symptoms (1.7%). The minimal chance of discomfort and negative side effects coupled with the generally unobtrusive nature in which the stimulation can be applied, through earbuds, speak to the relatively easy adaptability

of supplementing practical language learning situations with the adoption of taVNS for noticeable gains in performance.

7.4 Limitations and Future Work

There are a number of limitations and questions raised in this dissertation that need to be addressed in future research. For one, the n-sizes were somewhat limited for the active taVNS groups, and in particular conclusions for the complex modeling of ATI effects in Chapter 6 should be treated with caution. Additionally, given the complex nature of the corpus design, directionality can't be conclusively established at this stage with respect to whether better phonological learning led to better lexical learning or if there was a bidirectional benefit. In Table 1 (section 3.2.3), the order of events can be seen to be phonological pretesting followed by phonological training, then word learning, then lexical testing on day 1 with phonological training followed by word learning, ending in lexical testing and phonological posttesting on day 2. It is outside the scope of this dissertation to follow the chaining of training and testing events more closely, but one could imagine, especially in the context of the OM (Bordag et al., 2021; forthcoming), that each phonological tone training phase helps to reduce the fuzziness in the phonological domain of that form and in turn reduces the fuzziness at the lexical training phase and test, which also in part rely on that phonological domain knowledge. Likewise, lexical training phases are also in part training participants in the phonological domain (especially given the visual contour aid during training) in addition to the phonolexical mapping of the word.

The scope of the training in the corpus needs to be extended in a number of ways to increase generalizability. Participants only had to learn nine monosyllabic pseudowords with T1, T2, and T4 over a two-day training. Most Mandarin words are disyllabic, and it has been noted that learning disyllabic tone words may be harder than monosyllabic words (Chang & Bowles, 2015). One reason may be that Mandarin tones undergo sandhi in context to make them additionally difficult. T3 in particular can manifest very similarly to the T2 contour, which is likely a major reason why Pelzl's (2019a) review identifies T2 as the hardest tone for learners but T4 was observed to be the most difficult here: T3, while typically quite easy to identify in isolated syllables, was not present in this corpus design to create a more difficult contrast with T2. The training should also be extended to increase the difficulty by adding a larger number of learning targets and generalizability outcome measures and include delayed posttests so the durability of these taVNS effects can be better established.

Patterns of cognitive effort for active taVNS groups were largely opposite and reversed around sham between phonological categorization and lexical recognition. It will be important to tease these differences apart in future research, particularly whether these reflect mechanistic differences between the priming and peristim groups' effects on phonological and lexical learning or whether they are artifacts of this particular design. For example, perhaps the reduction of effort observed in lexical recognition for peristim is an artifact of active stimulation during the lexical recognition test and removing stimulation for that task might instead show a match in effort with the sham group while still showing behavioral benefits (cf. the phonological categorization results). As another example, note the learning benefits found in Jacobs et al. (2015) for taVNS employed continuous stimulation (like for priming) during a consolidation

phase, a period of rest between the training and testing. Given the present design wherein the priming taVNS group received three periods of priming on each training day, it may be that some of the positive behavioral outcomes produced here were not due to priming *per se*, but rather to consolidation or a combination thereof since the latter two periods of 'priming' on each training day were both before *and after* training material.

It will also be important to establish whether taVNS enhances only early phonolexical outcomes or if they may persist into even more complex contexts like multiword units or sentences. Given there is research in animal models to suggest VNS interventions may uniquely benefit auditory training (e.g., Borland et al., 2018; Engineer et al., 2011; Kilgard, 2012), this hasn't been established in humans and a comparison of taVNS efficacy with a difficult visual language learning paradigm is warranted. Given the fuzzy lexical encoding problem found with even advanced learners of Mandarin (Pelzl et al., 2021a), taVNS may also be a useful intervention at later stages of learning. It would also be worthwhile to investigate *ab initio* learning with a different language that varies by non-tonal, but still difficult, phoneme contrasts to see if tone word learning is uniquely benefitted by taVNS.

Indeed, taVNS affects cognition generally, so taVNS may have broad implications for learning in general. Pairing it with other difficult learning targets may be warranted. Indeed, the study of programming language learning may be a logical transition given some inherent similarities to natural language (e.g., Fedorenko et al., 2019; Pandža, 2016) as well as recent research showing a role for natural language aptitude on programming outcomes (Prat et al., 2020).

7.5 Conclusion

Transcutaneous auricular vagus nerve stimulation (taVNS) had positive effects on phonolexical tone word learning for native English speakers naïve to lexical tone as measured by phonological categorization, lexical recognition, and lexical recall tests. Advantages were found even after one training session and persisted into a second training day. Two different administrations of taVNS were studied, peristimulus (peristim) stimulation and priming stimulation, and were found to effect different types of benefits for learning. Despite the peristim administration having less total stimulation duration per day than priming, it results in as good or better accuracy than priming, showing that amount of stimulation is less important than where and how in the process of learning and testing it is administered.

This body of work also contributes more generally to research on Mandarin tone learning. It is the first to investigate cognitive effort for Mandarin lexical tone learning, phonological tone learning, and separately by tone, and it showed that tone word difficulty can be interpreted with respect to cognitive effort. The results were largely complementary with behavioral findings, and also revealed effort for difficult items associated with phonological tone categorization to heavily decline after only a two-day training, although the hardest tone was still moderately more effortful to process. Finally, results from sham participants (i.e., without active stimulation) suggest that tone word learning effort increases at increasing levels of non-linguistic tone aptitude and musicality, possibly indicating increased readiness to respond to tests of phonological and lexical tone.

Building on the findings for the sham taVNS control, this dissertation has provided evidence from a double-blind study that (1) peristim taVNS can enhance encoding of Mandarin tone words when time-locked to the targets of learning, helping generalize to new phonological exemplars, (2) priming taVNS can enhance phonological and lexical retrieval of newly learned information, and (3) learners are differentially advantaged by peristim or priming taVNS depending on both the difficulty of the information and their non-linguistic tone aptitude and musicality.

The promise of tVNS lies not only in these results, but in the fact that tVNS can be induced safely with consumer-grade equipment straight out of the box. Given the recency of research applying neurostimulation to second language learning and the diversity of neurophysiological mechanisms targeted by different techniques, future research should not only focus on developing effective protocols but also weigh the relative efficacy of these methods for different aspects of language learning while considering their practical limitations. When paired with traditional language learning approaches, neurostimulation may provide an indispensable boost for adult language learning to overcome inherent difficulties in second language learning.

References

Antoniou, M., & Wong, P. C. (2016). Varying irrelevant phonetic features hinders learning of the feature being trained. *The Journal of the Acoustical Society of America*, *139*(1), 271-278.

Aston-Jones, G., & Cohen, J.D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Ben-Menachem, E., Hamberger, A., Hedner, T., Hammond, E. J., Uthman, B. M., Slater, J., ... & Wilder, B. J. (1995). Effects of vagus nerve stimulation on amino acids and other metabolites in the CSF of patients with partial seizures. *Epilepsy research*, *20*(3), 221-227.

Bent, T., Bradlow, A.R., & Wright, B.A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 97–103.

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PloS One*, *8*(4), e60676.

Bordag, D., Gor, K., & Opitz, A. (2021). Ontogenesis Model of the L2 Lexical Representation. *Bilingualism: Language and Cognition,* 1-17. doi:10.1017/S1366728921000250

Bordag, D., Gor, K., & Opitz, A. (forthcoming). Refining Key Concepts of the Ontogenesis Model of the L2 Lexical Representation. *Bilingualism: Language and Cognition*.

Borland, M.S., Engineer, C.T., Vrana, W.A., Moreno, N.A., Engineer, N.D., Vanneste, S., … & Kilgard, M.P. (2018). The interval between VNS-tone pairings determines the extent of cortical map plasticity. *Neuroscience*, *369*, 76-86.

Borland, M.S., Vrana, W.A., Moreno, N.A., Fogarty, E.A., Buell, E.P., Sharma, P., ... & Kilgard, M.P. (2016). Cortical map plasticity as a function of vagus nerve stimulation intensity. *Brain Stimulation, 9*(1), 117-123.

Bowles, A.R., Chang, C.B., & Karuzis, V.P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*(4), 774-808.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911.

Capone, F., Assenza, G., Di Pino, G., Musumeci, G., Ranieri, F., Florio, L., ... & Di Lazzaro, V. (2015). The effect of transcutaneous vagus nerve stimulation on cortical excitability. *Journal of Neural Transmission*, *122*(5), 679-685.

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

Chang, C. B., and Bowles, A. R. (2015). Context Effects on Second-Language Learning of Tonal Contrasts. *Journal of the Acoustic Society of America*, *138*(6), 3703–3716. doi:10.1121/1.4937612

Colflesh, G., Karuzis, V., & O'Rourke, P. (2016). Effects of Working Memory Training on L2 Proficiency and Working Memory Capacity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 289-294

Colzato, L., & Beste, C. (2020). A literature review on the neurophysiological underpinnings and cognitive effects of transcutaneous vagus nerve stimulation: challenges and future directions. *Journal of Neurophysiology*, *123*(5), 1739-1755.

Cook, S. (2012). *Phonological Form in L2 Lexical Access: Friend or Foe?* Unpublished Doctoral dissertation, University of Maryland.

Cook, S. V., and Gor, K. (2015). Lexical access in L2: Representational deficit or processing constraint? *The Mental Lexicon*, *10*, 247–270. doi: 10.1075/ml.10.2.04coo

Cook, S. V., Pandža, N. B., Lancaster, A. K., and Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*:1345. doi: 10.3389/fpsyg.2016.01345

Cooper, A., & Wang, Y. (2013). Effects of tone training on Cantonese tone-word learning. *The Journal of the Acoustical Society of America*, *134*(2), EL133-EL139.

Daidone, D., & Darcy, I. (2021). Vocabulary size is a key factor in predicting second language lexical encoding accuracy. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.688356

Darcy, I., Daidone, D., and Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*, 372–420. doi: 10.1075/ml.8.3.06dar

DaSilva, A.F., Truong, D.Q., DosSantos, M.F., Toback, R.L., Datta, A., & Bikson, M. (2015). State-of-art neuroanatomical target analysis of high-definition and conventional tDCS montages used for migraine and pain control. *Frontiers in Neuroanatomy*, *9*: 89.

Delogu, F., Lampis, G., & Belardinelli, M. O. (2006). Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers? *Cognitive Processing*, *7*(3), 203-207.

Delogu, F., Lampis, G., & Belardinelli, M. O. (2010). From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception. *European Journal of Cognitive Psychology*, *22*(1), 46-61.

Dietrich, S., Smith, J., Scherzinger, C., Hofmann-Preiß, K., Freitag, T., Eisenkolb, A., & Ringler, R. (2008). A novel transcutaneous vagus nerve stimulation leads to brainstem and cerebral activations measured by functional MRI/Funktionelle Magnetresonanztomographie zeigt Aktivierungen des Hirnstamms und weiterer zerebraler Strukturen unter transkutaner Vagusnervstimulation. *Biomedical Engineering/Biomedizinische Technik*, *53*(3), 104-111.

Dittinger, E., Barbaroux, M., D'Imperio, M., Jäncke, L., Elmer, S., & Besson, M. (2016). Professional music training and novel word learning: from faster semantic encoding to longer-lasting word representations. *Journal of Cognitive Neuroscience*, *28*(10), 1584-1602.

Dittinger, E., Chobert, J., Ziegler, J. C., & Besson, M. (2017). Fast brain plasticity during word learning in musically-trained children. *Frontiers in Human Neuroscience*, *11*, 233.

Doughty, C.J., & Long, M.H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology*, *7*(3), 50-80.

Eckstein, M.K., Guerra-Carrillo, B., Singley, A.T.M., & Bunge, S.A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69-91.

Engineer, N.D., Riley, J.R., Seale, J.D., Vrana, W.A., Shetake, J.A., Sudanagunta, S.P., … & Kilgard, M.P. (2011). Reversing pathological neural activity using targeted plasticity. *Nature*, *470*(7332), 1-15.

Fedorenko, E., Ivanova, A., Dhamala, R., & Bers, M. U. (2019). The language of programming: A cognitive perspective. *Trends in Cognitive Sciences*, *23*(7), 525–528. https://doi-org.proxy-um.researchport.umd.edu/10.1016/j.tics.2019.04.010

Finocchiaro, C., Maimone, M., Brighina, F., Piccoli, T., Giglia, G., & Fierro, B. (2006). A case study of primary progressive aphasia: improvement on verbs after rTMS treatment. *Neurocase*, *12*(6), 317-321.

Follesa, P., Biggio, F., Gorini, G., Caria, S., Talani, G., Dazzi, L., ... & Biggio, G. (2007). Vagus nerve stimulation increases norepinephrine concentration and the gene expression of BDNF and bFGF in the rat brain. *Brain Research*, *1179*, 28-34.

Frangos, E., Ellrich, J., & Komisaruk, B. R. (2015). Non-invasive access to the vagus nerve central projections via electrical stimulation of the external ear: fMRI evidence in humans. *Brain Stimulation*, *8*(3), 624–636. https://doi.org/10.1016/j.brs.2014.11.018

George, M.S., & Aston-Jones, G. (2010). Noninvasive techniques for probing neurocircuitry and treating illness: vagus nerve stimulation (VNS), transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS). *Neuropsychopharmacology*, *35*(1), 301.

Gilzenrat, M.S., Nieuwenhuis, S., Jepma, M., & Cohen, J.D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252–269.

Gor, K., & Cook, S. V. (2020). A mare in a pub? Nonnative facilitation in phonological priming. *Second Language Research*, *36*(1), 123-140.

Gor, K., Cook, S., Bordag, D., Chrabaszcz, A., & Opitz, A. (2021). Fuzzy lexical representations in adult second language speakers. *Frontiers in Psychology*, *12*, 732030. https://doi.org/10.3389/fpsyg.2021.732030

Gottfried, T. L. (2007). Music and language learning: Effect of musical training on learning L2 speech contrasts. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 221-237). Amsterdam: John Benjamins.

Gottfried, T. L., & Ouyang, G. Y. H. (2005). Production of Mandarin tone contrasts by musicians and non-musicians. *The Journal of the Acoustical Society of America*, *118*(3), 2025.

Gottfried, T. L., & Ouyang, G. Y.-H. (2006). Training musicians and nonmusicians to discriminate Mandarin tones. *Journal of the Acoustical Society of America, 120,* 3167.

Gottfried, T. L., Staby, A. M., & Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *Journal of the Acoustical Society of America, 115*, 2545.

Groves, D.A., Bowman, E.M., & Brown, V.J. (2005). Recordings from the rat locus coeruleus during acute vagal nerve stimulation in the anaesthetised rat. *Neuroscience Letters*, *379*(3), 174-179.

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269-279.

Hayes-Harb, R., and Masuda, K. (2008). Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research*, *24*, 5–33. doi: 10.1177/0267658307082980

Ingvalson, E. M., Barr, A. M., & Wong, P. C. (2013). Poorer phonetic perceivers show greater benefit in phonetic-phonological speech learning. *Journal of Speech, Language, and Hearing Research*, *56*(3), 1045–1050. https://doi.org/10.1044/1092-4388(2012/12-0024)

Ingvalson, E.M., Ettlinger, M., & Wong, P.C. (2014). Bilingual speech perception and learning: A review of recent trends. *International Journal of Bilingualism, 18*(1), 35-47.

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of phonetics*, *39*(4), 571-584.

Jacobs, H.I., Riphagen, J.M., Razat, C.M., Wiese, S., & Sack, A.T. (2015). Transcutaneous vagus nerve stimulation boosts associative memory in older individuals. *Neurobiology of Aging*, *36*(5), 1860-1867.

Kaan, E., Wayland, R., Bao, M., & Barkley, C. M. (2007). Effects of native language and training on lexical tone perception: An event-related potential study. *Brain Research*, *1148*, 113-122.

Kilgard, M.P. (2012). Harnessing plasticity to understand learning and treat disease. *Trends in Neurosciences*, *35*(12), 715-722.

Kirkham, J., Lu, S., Wayland, R., & Kaan, E. (2011). Comparison of vocalists and instrumentalists on lexical tone perception and production tasks. In: *Proceedings of the 17th International Congress of Phonetic Sciences*, pp. 1098–1101.

Klooster, D.C., de Louw, A.J., Aldenkamp, A.P., Besseling, R.M.H., Mestrom, R.M.C., Carrette, S., ... & Boon, P. (2016). Technical aspects of neurostimulation: Focus on equipment, electric field modeling, and stimulation protocols. *Neuroscience & Biobehavioral Reviews*, *65*, 113-141.

Kraus, T., Kiess, O., Hösl, K., Terekhin, P., Kornhuber, J., & Forster, C. (2013). CNS BOLD fMRI effects of sham-controlled transcutaneous electrical nerve stimulation in the left outer auditory canal–a pilot study. *Brain Stimulation*, *6*(5), 798-804.

Kuchinsky, S.E., Ahlstrom, J.B., Vaden Jr, K.I., Cute, S.L., Humes, L.E., Dubno, J.R., & Eckert, M.A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*(1), 23-34.

Kuchinsky S. E., & Vaden K. I. (2020). Aging, Hearing Loss, and Listening Effort: Imaging Studies of the Aging Listener. In: Helfer K.S., Bartlett E.L., Popper A.N., Fay R.R. (Eds.) *Aging and Hearing. Springer Handbook of Auditory Research, vol 72*. Springer, Cham. https://doi.org/10.1007/978-3-030-49367-7_10

Kuipers, J.R., & Thierry, G. (2011). N400 amplitude reduction correlates with an increase in pupil size. *Frontiers in Human Neuroscience*, *5*, 61.

Lee, C. Y., & Hung, T. H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *The Journal of the Acoustical Society of America*, *124*(5), 3235-3248.

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*(4), 593-620.

Li, M., & Dekeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 mandarin tonal word production. *The Modern Language Journal*, *103*(3), 607-628.

Liu, C., & Chandrasekaran, B. (2013). Effects of phonological training on tone perception for English listeners. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, No. 1, p. 060057). Acoustical Society of America.

Llanos, F., McHaney, J. R., Schuerman, W. L., Han, G. Y., Leonard, M. K., & Chandrasekaran, B. (2020). Non-invasive peripheral nerve stimulation selectively enhances speech category learning in adults. *npj Science of Learning*, *5*(1), 1-11.

Loewenfeld, I.E. (1999). Otto Lowenstein: Neurologic and ophthalmologic testing methods during his lifetime. *Documenta Ophthalmologica*, *98*(1), 3-20.

Llompart, M., & Reinisch, E. (2020). The phonological form of lexical items modulates the encoding of challenging second-language sound contrasts. *Journal of Experimental Psychology: Learning and Memory*, *46*(8), 1590–1610. doi:10.1037/xlm0000832

Lu, S., Wayland, R., & Kaan, E. (2015). Effects of production training and perception training on lexical tone perception–A behavioral and ERP study. *Brain Research*, *1624*, 28-44.

Maddox, W. T., Chandrasekaran, B., Smayda, K., & Yi, H.-G. (2013). Dual systems of speech category learning across the lifespan. *Psychology and Aging, 28*(4), 1042–1056. https://doi.org/10.1037/a0034969

Manta, S., Dong, J., Debonnel, G., & Blier, P. (2009). Enhancement of the function of rat serotonin and norepinephrine neurons by sustained vagus nerve stimulation. *Journal of Psychiatry & Neuroscience, 34*(4), 272-280.

Marshall, L., Mölle, M., Hallschmid, M., & Born, J. (2004). Transcranial direct current stimulation during sleep improves declarative memory. *Journal of Neuroscience*, *24*(44), 9985-9992.

Meinzer, M., Jähnigen, S., Copland, D.A., Darkow, R., Grittner, U., Avirame, K., … & Flöel, A. (2014). Transcranial direct current stimulation over multiple days improves learning and maintenance of a novel vocabulary. *Cortex*, *50*, 137-147.

Miniussi, C., Cappa, S.F., Cohen, L.G., Flöel, A., Fregni, F., Nitsche, M.A., … & Walsh, V. (2008). Efficacy of repetitive transcranial magnetic stimulation/transcranial direct current stimulation in cognitive neurorehabilitation. *Brain Stimulation*, *1*, 326-336.

Mottaghy, F.M., Hungs, M., Brügmann, M., Sparing, R., Boroojerdi, B., Foltys, H., ... & Töpper, R. (1999). Facilitation of picture naming after repetitive transcranial magnetic stimulation. *Neurology*, *53*(8), 1806-1806.

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, *35*(4), 418-440.

Naeser, M. A., Martin, P. I., Ho, M., Treglia, E., Kaplan, E., Bashir, S., & Pascual-Leone, A. (2012). Transcranial magnetic stimulation and aphasia rehabilitation. *Archives of Physical Medicine and Rehabilitation*, *93*(1), S26-S34.

Ohlenforst, B., Zekveld, A.A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., … & Kramer, S.E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. https://doi.org/10.1016/j.heares.2017.05.012

Ohn, S.H., Park, C.I., Yoo, W.K., Ko, M.H., Choi, K.P., Kim, G.M., ... & Kim, Y.H. (2008). Time-dependent effect of transcranial direct current stimulation on the enhancement of working memory. *Neuroreport*, *19*(1), 43-47.

Ollen, J.E. (2006). A criterion-related validity test of selected indicators of musical sophistication using expert ratings. Unpublished doctoral dissertation, Ohio State University, Ohio.

Öztürk, L., Büning, P. E., Frangos. E., de Lartigue, G., & Veldhuizen, M. G. (2020). tVNS increases liking of orally sampled low-fat foods: A pilot study. *Frontiers in Human Neuroscience*, *14*(600995), 1-10. doi: 10.3389/fnhum.2020.600995

Pandža, N. B. (2016). Computer programming as a second language. In D. Nicholson (Ed.), *AISC: Vol. 501. Advances in Human Factors in Cybersecurity* (pp. 439-445). Switzerland: Springer. DOI: 10. 1007/978-3-319-41932-9_36

Pandža, N. B., Phillips, I., Karuzis, V. P., O'Rourke, P., & Kuchinsky, S. E. (2020). Neurostimulation and pupillometry: New directions for learning and research in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 56-77. https://doi.org/10.1017/S0267190520000069

Pascual-Leone, A., Walsh, V., & Rothwell, J. (2000). Transcranial magnetic stimulation in cognitive neuroscience–virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology*, *10*(2), 232-237.

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology*, *2*, 1–14. doi:10.3389/fpsyg.2011.00142

Pelzl, E. (2019a). What makes second language perception of Mandarin tones hard?: A non-technical review of evidence from psycholinguistic research. *Chinese as a Second Language*, *54*(1), 51-78.

Pelzl., E., Lau, E.F., Guo, T., & DeKeyser, R. (2019b). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, *41*(1), 59-86.

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2021a). Advanced second language learners of Mandarin show persistent deficits for lexical tone encoding in picture-to-word form matching. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.689423

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2021b). Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings

from behavioral and event-related potentials experiments. *Studies in Second Language Acquisition*, *43*(2), 268-296.

Perrachione, T.K., Lee, J., Ha, L.Y., & Wong, P.C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461-472.R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., ... & Naylor, G. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*, 5S-27S.

Prat, C. S., Madhyastha, T. M., Mottarella, M. J., & Kuo, C. H. (2020). Relating natural language aptitude to individual differences in learning programming languages. *Scientific Reports*, *10*(1), 1-10.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Redgrave, J., Day, D., Leung, H., Laud, P. J., Ali, A., Lindert, R., & Majid, A. (2018). Safety and tolerability of transcutaneous vagus nerve stimulation in humans; a systematic review. *Brain Stimulation*, *11*(6), 1225-1238.

Reis, J., Robertson, E.M., Krakauer, J.W., Rothwell, J., Marshall, L., Gerloff, C., … & Cohen, L. G. (2008). Consensus: Can transcranial direct current stimulation and transcranial

magnetic stimulation enhance motor learning and memory formation? *Brain Stimulation*, *1*, 363-369.

Sakai, K.L., Noguchi, Y., Takeuchi, T., & Watanabe, E. (2002). Selective priming of syntactic processing by event-related transcranial magnetic stimulation of Broca's area. *Neuron*, *35*(6), 1177-1182.

Samuels, E.R., & Szabadi, E. (2008). Functional neuroanatomy of the noradrenergic locus coeruleus: its roles in the regulation of arousal and autonomic function part I: principles of functional organisation. *Current Neuropharmacology*, *6*(3), 235-253.

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*, *40*(3), 529-549.

Sebastián-Gallés, N., & Díaz, B. (2012). First and second language speech perception: Graded learning. *Language Learning*, *62*, 131-147.

Shao, J., & Zhang, C. (2020). Dichotic perception of lexical tones in Cantonese-speaking congenital amusics. *Frontiers in Psychology*, *11*, 1411.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological science*, *17*(8), 675-681.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.

Van Leusden, J. W., Sellaro, R., & Colzato, L. S. (2015). Transcutaneous Vagal Nerve Stimulation (tVNS): a new neuromodulation tool in healthy humans? *Frontiers in Psychology*, *6*, 102.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R.H., & Wood, S.N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1-23.

van Rij, J., Wieling, M., Baayen, R., van Rijn, H. (2020). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.4.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005-2015.

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., & Castellanos, F.X. (2015). *Inscapes*: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222-232.

Voeten (2019). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 1.3.

Vonck, K., Raedt, R., Naulaerts, J., De Vogelaere, F., Thiery, E., Van Roost, D., … & Boon, P. (2014). Vagus nerve stimulation… 25 years later! What do we know about the effects on cognition? *Neuroscience & Biobehavioral Reviews*, *45*, 63-71.

Walker, B. R., Easton, A., & Gale, K. (1999). Regulation of limbic motor seizures by GABA and glutamate transmission in nucleus tractus solitarius. *Epilepsia*, *40*(8), 1051-1057.

Walsh, V., & Pascual-Leone, A. (2003). *Transcranial magnetic stimulation: A neurochronometrics of mind*. Cambridge, MA: MIT Press.

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain and Language*, *78*(3), 332-348.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, *113*(2), 1033-1043.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649-3658.

Wayland, R., Herrera, E., & Kaan, E. (2010). Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, *38*(4), 654-662.

Wiener, S., Lee, C. Y., and Tao, L. (2019). Statistical regularities affect the perception of second language speech: Evidence from adult classroom learners of Mandarin Chinese. *Language Learning*, *69*(3), 527–558. doi:10.1111/lang.12342

Wing, H. D. (1968). *Tests of musical ability and appreciation: An investigation into the measurement, distribution, and development of musical capacity* (2nd ed.). London: Cambridge University Press.

Wong, H. (1953). Outline of the Mandarin phonemic system. *Word*, *9*(3), 268-276. DOI: 10.1080/00437956.1953.11659474

Wong, F. C. K., Chandrasekaran, B., Garibaldi, K., & Wong, P. C. M. (2011). White matter anisotropy in the ventral language pathway predicts sound-to-word learning success. *Journal of Neuroscience*, *31*, 8780–8785.

Wong, P.C., & Perrachione, T.K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*(4), 565–585.

Wood, S.N. (2017). Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

Yakunina, N., Kim, S. S., & Nam, E. C. (2017). Optimization of transcutaneous vagus nerve stimulation using functional MRI. *Neuromodulation: Technology at the neural interface*, *20*(3), 290-300.

You, D. S., Kim, D.-Y., Chun, M. H., Jung, S.E., & Park, S.J. (2011). Cathodal transcranial direct current stimulation of the right Wernicke's area improves comprehension in subacute stroke patients. *Brain & Language*, *199*, 1-5.

Zekveld, A.A., Koelewijn, T., & Kramer, S.E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, *22*, 1-25.

Zhang, Y., Liu, J., Li, H., Yan, Z., Liu, X., Cao, J., Park, J., Wilson, G., Liu, B., & Kong, J. (2019). Transcutaneous auricular vagus nerve stimulation at 1 Hz modulates locus coeruleus activity and resting state functional connectivity in patients with migraine: an fMRI study. *NeuroImage: Clinical*, *24*, 101971. https://doi.org/10.1016/j.nicl.2019.101971