UNIVERSITY OF CALIFORNIA
RIVERSIDE

Perceptual Learning: Assessment and Training Across the Mechanical Senses

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Psychology

by

Esteban Sebastian Lelo de Larrea-Mancera

September 2021

Dissertation Committee:
Dr. Aaron R. Seitz, Chairperson
Dr. Khaleel Abdul Razak
Dr. Frederick J. Gallun

The Dissertation of Esteban Sebastian Lelo de Larrea-Mancera is approved by:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Initial Research Evaluation Committee:

Aaron R. Seitz

Megan A. K. Peters

Weiwei Zhang

Qualifying Exam Committee:

Aaron R. Seitz

Weiwei Zhang

Lawrence Rosenblum

Edward Zagha

Tim Labor

Dissertation Committee:

Aaron R. Seitz

Khaleel A. Razak

Frederick J. Gallun

Funding:

UC-MEXUS-CONACYT fellowship #CVU-404659

NIH-NIDCD Grant No. R01 DC 015051.

NIH-NICHD Grant No. R03 HD94234

Note from Author:

To my mother, my sister and the friends who supported me during this period. In honor of my father and my grandmother who left before they had to read this. Many thanks to my research assistants and the staff at the Brain Game Center without whom this work would not have been possible. And many thanks to my advisor Aaron Seitz who helped me struggle through it all and remained patient even in my most passionate of rebellions.

Acknowledgement:

This dissertation includes previously published material in Chapters 2, 3, 3b and 4. Chapter 2 was authored by Larrea-Mancera, E. S. L., Dempsey-Jones, H., Makin, T., & Seitz, A. R. And published in 2019 in the Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 1–6. https://doi.org/10.1109/DEVLRN.2019.8850704. Chapter 3 was authored by Larrea-Mancera, E. S. L., Stavropoulos, T., Hoover, E., Eddins, D., Gallun, F., & Seitz, A. And published in 2020 in the Journal of the Acoustical Society of America, 148(4), 1831–1851. https://doi.org/10.1101/2020.01.08.899088. Chapter 3b was authored by Larrea-Mancera, E. S. L., Stavropoulos, T., Carrillo, A. A., Eddins, D. A., Molis, M. R., Gallun, F. J., & Seitz, A. R. And was uploaded to a pre-print server in 2021. PsyArXiv. https://psyarxiv.com/9u68p/. Lastly, Chapter 4 was authored by Larrea-Mancera, E. S. L. De, Philipp, M. A., Stavropoulos, T., Anna, A., Cheung, S., Koerner, T., Molis, M. R., Gallun, F. J., & Aaron, R. And has been accepted for publication in 2021 by the Journal of Cognitive Enhancement. It is available now as a pre-print in BioRxiv. https://doi.org/https://doi.org/10.1101/2021.01.26.428343.

ABSTRACT OF THE DISSERTATION

Perceptual Learning: Assessment and Training Across the Mechanical Senses

by

Esteban Sebastian Lelo de Larrea-Mancera

Doctor of Philosophy, Graduate Program in Psychology
University of California, Riverside, September 2021
Dr. Aaron R. Seitz, Chairperson

The possibility of directed improvement of perceptual ability is the main force behind this work. In contrast to the majority of perceptual studies that focus on vision, this work centers on the mechanical senses of touch and hearing. In the case of touch perception, aspects of vibratory stimulation that would promote perceptual learning that transfers to untrained features were explored. This line of inquiry has the long-term goal of building useful training for prosthetic limb control. In the case of hearing, a much more nuanced depiction of relevant dimensions of perceptual ability is provided based on research from the fields of psychophysics and auditory neuroscience. We identify and begin to address the need for translation of this scientific laboratory work into a clinical domain. To this end, a number of central auditory processes assessments with potential clinical relevance was validated in a portable automatic rapid testing (PART) platform.

We present robust results across different external noise conditions as well as variations of portable device, headphones, and testing settings (lab vs home). Our results suggest PART may be used to start collecting large enough datasets to create performance norms and ultimately translate these laboratory assessments into clinical practice. Finally, we use PART as the assessment element to evaluate perceptual improvement after an auditory training video-game intervention called *Listen* that incorporates a body of findings in perceptual learning (PL) to promote learning that would generalize to the ability to perceive speech in noise. We present promising preliminary results with a young normal hearing sample and suggest a potential application for people in need. Such application could serve to improve people's lives in meaningful ways through the preservation or improvement of hearing. In conclusion, the work compiled here portrays a somewhat broad picture of conducting PL research across the mechanical senses (touch and audition) where the complexities of assessment and training choices and the different possible scopes of perceptual training are detailed. This dissertation represents a methodological tool for PL research. At the same time it provides examples on the use of PART (auditory assessments) and *Listen* (auditory training), tools developed by the Brain Game Center that have plenty of potential for basic and clinical research beyond the confines of this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: GENERAL INTRODUCTION

Perceptual Learning (PL) in broad terms refers to the observation that: *repeated practice or training with a particular type of sensory challenge may lead to improved perceptual performance* (see Seitz, 2017*a* for a primer in PL). This might sound obvious at first, but there is much depth to be explored in such an assertion. For example, it might be surprising to some that it applies to anyone, as PL is found throughout the lifespan (Gibson, 1963; Seitz, 2021), and perception is indeed subject to change in response to experience even in older adults (Anderson et al., 2013*ab*; Karawani et al., 2015). This makes PL a very exciting field of study in Psychology and Cognitive Neuroscience as it can reveal knowledge on the human condition and also, it can potentially be applied to our benefit. But how much repetition or training is needed to achieve such benefits? What types of sensory or perceptual challenge is most adequate? And what sort of improvements may we expect? These are questions inherent in the simple definition of PL asserted here, and that any research in PL must address. Additionally, implicit in everything said thus far is the ability to measure perceptual performance, and the use of such assessment to determine whether repeated perceptual challenge promotes perceptual improvement, or in other words, whether PL was observed.

Improved perceptual performance described as PL can be specific to the sensory task and stimuli being practiced, or it can generalize or transfer to other similar perceptual experiences and challenges. Different factors associated with both assessment and training have been described as mediating generalization and transfer of PL (see Fahle, 2005; Petrov, Dosher & Lu, 2005; Seitz, 2017*b*). As hinted above, PL studies can differ widely in their scope, from basic research to practical application. The scope of research has historically dictated the decisions associated with the selection of assessment and training features and the expectations on the training outcomes. Mechanistic studies have used the patterns of specificity and transfer of PL to describe specific neuro-cognitive structures that could have given rise to it. Specific PL has been classically associated with low-level sensory brain areas that match such specificity (Fiorentini & Berardi, 1980; Poggio, Fahle & Edelman, 1992; Schoups et al., 2001). In contrast, generalizable learning has been taken as evidence of plasticity in higher-order brain areas that read out information from multiple specific perceptual features (Dosher & Lu, 1998; Ghose, Yang & Maunsell, 2002). Through the years complex patterns of specificity and transfer of learning have been observed in PL (e.g. Hung & Seitz,2014) moving the debate away from a simple dichotomy to a more complex dynamical and holistic view that includes several processes from different hierarchies of processing (Ahissar & Hochstein, 1997; Ahissar & Hochstein, 2004; Zhang et al., 2011; Watanabe &

Sasaki, 2015), including automatic processes (Seitz & Watanabe, 2003; Seitz & Dinse, 2007; Seitz & Watanabe, 2009; Seitz, Kim & Watanabe, 2009) and involving the whole brain (Maniglia & Seitz, 2018).

Crucial to all the mechanistic studies and PL models mentioned above is the careful and deterministic manipulation of the amounts of training, the types of perceptual challenge (both in assessment and training), and the training outcomes targeted (see Green et al., 2019). However, studies in PL can also have a different scope with focus on application of the learning principles extracted from mechanistic studies for the promotion of some aspect of a person's quality of life through improved perceptual ability. These types of efficacy/effectivity studies (see Green et al., 2019) will typically target generalizable learning and even consider specificity in PL as a sort of curse (see Deveau & Seitz, 2013; Deveau, Lovcik & Seitz, 2014; Deveau & Seitz, 2014). The basic idea for these types of studies is to conduct meaningful assessments that speak to someone's perceptual ability as it manifests in a particular area of their life. This means to conduct training with the aim of improving not only the specific aspects of training such as the stimuli and the task structures used, but with the hopes of a transfer of the task-related improvement to other conditions, thus achieving an efficacious/effective intervention that could manifest in everyday life. This applied side of PL can range from those with special needs like children with specific language impairments (e.g. Merzenich et al., 1996;

Tallal et al., 1996), those needing rehabilitation such as patients with macular degeneration (e.g. Maniglia et al., 2016) to professionals that require extraordinary ability such as radiologists (Seitz, 2017) or professional baseball players (e.g. Deveau, Ozer & Seitz, 2014).

The problem regarding when PL is specific and when does it generalize represents an implicit structural aspect of the current dissertation work. It is explicitly approached in Chapter 2 which investigates whether complexity in the patterns of vibro-tactile stimulation delivered at the fingertips promoted generalization of tactile PL (Larrea-Mancera et al., 2019). While most studies in PL have been conducted in vision, the mechanical senses: touch and audition have been relatively unexplored. Chapter 2 explores different patterns of specificity and transfer of tactile PL in people who were either trained in a perceptual discrimination task with narrow-band simple vibrotactile stimulation delivered at the fingertips, and another trained with broad-band complex stimulation that better resembled stimuli as it is displayed in the real world. In addition to PL being assessed for the trained stimuli, transfer to a set of simple frequencies, one of simple durations, un-trained fingertips, and a novel broad-band set of complex patterns were also tested before and after training. This study aimed both at describing the locus of plasticity based on the patterns of specificity and transfer of PL in each condition, trying to discover where in the processing hierarchy could the mechanism of PL be implemented. And at the

same time, the study was aimed at evaluating whether the bandwidth of vibrotactile stimulation could be an important dimension to consider in efficacy studies aiming at training sensory-motor prosthetic control of amputees who could benefit from improved perceptual discrimination at the interface of the remaining limb and the prosthetic to the extent that PL generalizes to real-world conditions. Even though it was difficult to reach a simple conclusion regarding the effect of stimulation bandwidth used for training due to both groups showing transfer to different perceptual assessments, our results suggest that the pattern of specificity of perceptual learning might be different for particular dimensions, cues, or types of information collected by a given perceptual system. Studies like this contribute to theory building bringing nuance on the relevant factors that modulate PL to be more or less specific to trained attributes, and infuse complexity in the categories or dimensions of learning transfer we think about for this modulation of specificity to occur.

A simple observation on PL studies in general and Chapter 2 in particular, is that transfer of PL can only be explored to the extent transfer tests are applied before and after training. Careful selection of such assessments is necessary both to properly address mechanism and to evaluate meaningfully an applied intervention. In Chapter 3 we identify a number of psychophysical assessments of central auditory processing (CAP) that have been documented to have potential clinical utility to understand a wide range of hearing difficulties,

importantly the capacity to understand speech in noisy conditions (Larrea-Mancera et al., 2020). These assessments of auditory function allow us to better understand hearing capacities of individuals in different domains such as spatial hearing, spectral sensitivity or temporal sensitivity. In this chapter, we describe the validity and reliability of several measures of CAP that were implemented in a computer tablet application named PART (for Portable Automatic Rapid Testing), developed by the Brain Game Center under the direction of Dr. Aaron Seitz. This assessment work establishes a number of important hearing dimensions that could be evaluated with potential clinical utility in a PL study that targeted hearing processing ability. The battery tested included tests of different target sounds in competition with distractors including speech, spectro-temporally modulated sound detection, frequency modulation detection, and temporal gap detection. In sum, we found that the battery of assessments we tested on a sample of 150 UCR undergraduates were highly stable across repetitions (test retest reliability), resembled laboratory results of the tests it intends to replicate (construct validity), and further, measurement was stable across the different conditions of equipment, adaptive procedure and environmental noise that were tested.

In order to expand the potential scenarios of application of PART and the hearing assessments we validated in Chapter 3, we replicated the study online (effectively addressing 2020-2021 global pandemic challenges on data collection) by harnessing PART's availability across a number of platforms, and

participant's own devices (Larrea-Mancera et al., 2021a). This replication study represents a direct extension to the work portrayed in Chapter 3 and thus it is labeled Chapter 3b. We found similar distributions of thresholds in a young normal hearing population with no clear effects of headphone, or device variability, although slightly worse thresholds than the original study by half a standard deviation. This difference was found systematically across all assessments and establishes both expectations for deviation from the normative thresholds portrayed in Chapter 3 and suggests caution when interpreting the sensitivity of the measures in remote settings. Further work is currently being conducted to better understand the sources of these small but systematic differences in estimated thresholds collected in remote settings. The studies portrayed in Chapter 3 and 3b lay the foundation for the auditory training study reported in Chapter 4 which used a number of PART hearing assessments to quantify the hypothesized broad-based transfer of training-related effects of an auditory training intervention.

Chapter 4 as mentioned above touches on the problem of specificity of PL from a different perspective than Chapter 2. Instead of aiming to discover the potential mechanisms that may give rise to learning, it aims to promote generalization. The goal of the auditory training (AT) intervention is that it could be applied in clinical or educational settings to improve hearing capacities in a given individual, not only on the trained stimuli, but on other (untrained)

conditions outside of the laboratory, that is in real world conditions. In contrast to the Chapter 2 where specificity of PL is considered to reveal the locus of the mechanism of PL, and is neither good nor bad as it informs learning mechanisms, specificity in Chapter 4 is considered as a “curse”. Chapter 4 reports the efficacy of an AT intervention using *Listen*, an auditory video-game specifically designed to incorporate knowledge from PL and hearing neuroscience to promote generalization of PL to a wide variety of hearing abilities (Larrea-Mancera et al., 2021b). We present promising preliminary results suggesting that our AT transfers to measures of speech intelligibility in noisy conditions using a small sample of young listeners without hearing difficulties.

In sum the present work depicts a somewhat broad picture of conducting PL research across the mechanical senses (touch and audition) where the complexities of assessment and training choices and the different possible scopes of perceptual training are described. Starting with an exploratory mechanistic stroll through the world of touch that affords imaginative leaps into the applications of PL in the realm of robotics and prosthetics, we continue to the world of hearing. Here, a wide body of knowledge of auditory processes is used to generate a solid platform for assessment of relevant hearing capacities and with it, the possibility of nuanced evaluations of different interventions including auditory training (AT). The literary voyage ends with the exploration of the efficacy of a video-game based AT intervention that incorporates PL and hearing

neuroscience principles to promote a wide range of hearing ability and the preliminary finding that it can, potentially for people in need. Therefore, the overarching purpose of this dissertation is to document ways in which research may be conducted to ultimately improve people's lives in meaningful ways through the preservation, improvement or generation of useful PL across the mechanical senses.

REFERENCES

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. https://doi.org/10.1038/387401a0

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. https://doi.org/10.1016/j.tics.2004.08.011

Anderson, S., White-Schwoch, T., Choi, H. J., et al. (2013a). Training changes processing of speech cues in older adults with hearing loss. Front Syst Neurosci, 7, 97.

Anderson, S., White-Schwoch, T., Parbery-Clark, A., et al. (2013b). Reversal of age-related neural timing delays with training. Proc Natl Acad Sci U S A, 110, 4357–4362.

Deveau, J., Lovcik, G., & Seitz, A. R. (2014). Broad-based visual benefits from training with an integrated perceptual-learning video game. *Vision Research*, *99*, 134–140.

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology*, *24*(4), R146–R147.

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, *95*(23), 13988–13993.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology*, *15*(2), 154–160. https://doi.org/10.1016/j.conb.2005.03.010

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. In *Nature* (Vol. 287, Issue 5777, pp. 43–44). https://doi.org/10.1038/287043a0

Ghose, G. M., Yang, T., & Maunsell, J. H. R. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *Journal of Neurophysiology*, *87*(4), 1867–1888.

Gibson, E.J. (1963). Perceptual learning. *Annu. Rev. Psychol*. 14, 29–56.

Green, C.S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., Bingel, U., Chein, J. M., Colzato, L. S., Edwards, J. D., Facoetti, A., Gazzaley, A., Gathercole, S. E., Ghisletta, P., Gori, S., Granic, I., Hillman, C. H., Hommel, B., Jaeggi, S. M., … Witt, C. M. (2019). Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement. Journal of Cognitive Enhancement, 3(1), 2–29.

Hung, S. C., & Seitz, A. R. (2014). Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *Journal of Neuroscience*, *34*(25), 8423–8431.

Karawani, H., Bitan, T., Attias, J., et al. (2015). Auditory perceptual learning in adults with and without age-related hearing loss. Front Psychol, 6, 2066.

Larrea-Mancera E. S. L., Dempsey-Jones, H., Makin, T. and Seitz, A. R. (2019) Does Training on Broad Band Tactile Stimulation Promote the Generalization of Learning? Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Oslo, Norway, 2019, pp. 1-6.

Larrea-Mancera, E. S. L., Stavropoulos, T., Hoover, E. C., Eddins, D. A., Gallun, F. J., & Seitz, A. R. (2020). Portable Automated Rapid Testing (PART) for auditory research: Validation in a normal hearing population. BioRxiv, 2020.01.08.899088.

Larrea-Mancera, E. S. L. De, Stavropoulos, T., Carrillo, A. A., Eddins, D. A., Molis, M. R., Gallun, F. J., & Seitz, A. R. (2021a). Portable Automated Rapid Testing (PART) of auditory processing abilities in young normally-hearing listeners : A remotely administered replication with participant-owned devices . *PsyArXiv*.

Larrea-Mancera, E. S. L., De, Philipp, M. A., Stavropoulos, T., Anna, A., Cheung, S., Koerner, T., Molis, M. R., Gallun, F. J., & Aaron, R. (2021b). Training with an auditory perceptual learning game transfers to speech in competition. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2021.01.26.428343

Maniglia, M., Pavan, A., Sato, G., Contemori, G., Montemurro, S., Battaglini, L., & Casco, C. (2016). Perceptual learning leads to long lasting visual improvement in patients with central vision loss. *Restorative Neurology and Neuroscience*, *34*(5), 697–720. https://doi.org/10.3233/RNN-150575

Maniglia, M., & Seitz, A. R. (2018). Towards a whole brain model of Perceptual Learning. Current Opinion in Behavioral Sciences, 20, 47–55.

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal Processing Deficits of Language-Learning Impaired Children Ameliorated by Training. Science, 271(5245), 77–81.

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, *112*(4), 715–743. https://doi.org/10.1037/0033-295X.112.4.715

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast Perceptual Learning in visual hyperacuity. *Science*, *256*(5059), 1018–1021.

Schoups, A., Vogels, R., Qian, N., & ORBAN, G. . (2001). Practising orientation identi®cation improves orientation coding in V1 neurons. *Nature*, *412*(August), 549–553.

Seitz, A. R., & Watanabe, T. (2003). Is subliminal learning really passive. *Nature*, *422*(6927), 36. https://doi.org/10.1038/422036a

Seitz, A. R., & Dinse, H. R. (2007). A common framework for perceptual learning. *Current Opinion in Neurobiology*, *17*(2), 148–153. https://doi.org/10.1016/j.conb.2007.02.004

Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research*, *49*(21), 2604–2610. https://doi.org/10.1016/j.visres.2009.08.003

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards Evoke Learning of Unconsciously Processed Visual Stimuli in Adult Humans. *Neuron*, *61*(5), 700–707. https://doi.org/10.1016/j.neuron.2009.01.016

Seitz, A. R. (2017a). Perceptual learning. Current Biology, 27(13), R631–R636. https://doi.org/10.1016/j.cub.2017.05.053

Seitz, A. R. (2017b). Generalizable Learning: Practice Makes Perfect — But at What? *Current Biology*, *27*(6), R225–R227. https://doi.org/10.1016/j.cub.2017.01.064

Seitz, A. R. (2021). Perceptual Learning: Changes across the Lifespan. *Current Biology*, *31*(2), R69–R72. https://doi.org/10.1016/j.cub.2020.11.024

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M., & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. Science, 271(5245), 81–84. https://doi.org/10.1126/science.271.5245.81

Watanabe, Takeo; Sasaki, Y. (2015). Perceptual learning: Toward a comprehensive theory. *Annu Rev Psychol*, *3*(66), 197–221.

Zhang, J.-Y., Wang, R., Klein, S., Levi, D., & Yu, C. (2011). Perceptual learning transfers to untrained retinal locations after double training: A piggyback effect. *Journal of Vision*, *11*(11), 1026–1026. https://doi.org/10.1167/11.11.1026

CHAPTER TWO: Does Training on Broad Band Tactile Stimulation Promote the Generalization of Learning?

This chapter presents the first example of assessment and training on the mechanical sense of touch. It was presented in the Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (2019) in Oslo, Norway and published in the Meeting Proceedings. The published version of the manuscript can be found online here: https://ieeexplore.ieee.org/document/8850704/

ABSTRACT

Given the clear role of sensory feedback in successful motor control, there is a growing interest in integrating substitutionary tactile feedback into robotic limb devices. To enhance the utility of such feedback, here we investigate how to best improve the limited generalization of tactile learning across body parts and stimulus properties. Specifically, we sought to understand how perceptual learning with different types of tactile stimuli may give rise to different patterns of learning generalization. To address this, we utilized vibro-tactile effectors to present patterns of stimulation in a match-to-sample paradigm. One group of participants trained on narrow-band stimulation consisting of simple sinusoidal

vibrations, and the other on broad-band stimulation generated from music. We hypothesized that training on broad-band tactile stimulation would promote greater generalization of learning outcomes. We found training with broad-band stimuli generalized to underlying stimulus features of frequency discrimination but showed weaker generalization to un-trained digits. This study provides a first step towards devising perceptual learning paradigms that will generalize broadly to the untrained perceptual contexts.

INTRODUCTION

Movements are produced by a combination of motor outflow and sensory inflow. Movements for object manipulation, such as grasping a cup or playing a musical instrument, require rapid integration of motor control and sensory feedback. It is the assimilation of these two processes that leads to the intuitive execution of movements (Wolpert, Ghahramani & Jordan, 1995). Patients experiencing sensory loss from the body demonstrate how even simple tasks, like lifting a cup, are devastated by the absence of somatosensory feedback (Richardson et al., 2016). As such, it is being increasingly recognised that the functionality of artificial limbs is severely restricted by the absence of this essential source of action information (Bensmaia & Miller, 2014). Artificial tactile feedback is realised through the delivery of direct somatosensory stimulation through targeted reinnervation (Kuiken et al., 2016), direct (Tan et al., 2013) and transcutaneous electrical nerve stimulation (Horch, Meek, Taylor & Hutchinson, 2011), as well as cutaneous stimulation – most commonly using vibro-tactors (Schofield, Evans, Carey & Hebert, 2014). A key challenge for successful tactile integration across these approaches is the ability of the perceptual systems to successfully interpret the artificial stimulation.

The field of Perceptual Learning provides a window into the ways that sensory experiences shape current perceptions of the world. Perceptual learning

studies typically train participants on simple stimuli that resemble basic dimensions of neural coding (e.g. pure frequencies, durations, and intensities). The extent to which learning is specific to these basic stimulus-features has been taken as evidence of low-level sensory learning (Fiorentini & Berardi, 1980; Merzenich et al., 1988; Karni & Sagi, 1991; Poggio, Fahle & Edelman, 1992; Sagi, 2011). For example, tactile perceptual learning studies have shown that the primary somatosensory cortex is selectively tuned to simple frequencies of mechanical sinusoids delivered to the fingertips (Mountcastle, Steinmetz & Romo, 1990; Recanzone et al., 1992; Hernández, Salinas, García & Romo, 1997; Harris, Harris & Diamond, 2001). Other studies have found that tactile PL can generalize from trained to un-trained digits (Hernández, Salinas, García & Romo, 1997) and that generalization of learning may reflect topography of their representation in the somatosensory cortex with greater learning generalization to overlapping representations (Harris, Harris & Diamond, 2001; Harrar, Spence & Makin, 2014; Dempsey-Jones et al., 2015), however see also (Sathian & Zangaladze, 1997; Spengler et al., 1997) for complete generalization across fingers.

However, recent theories suggest that perceptual learning is best understood through a model where multiple components, including low-level sensory representations, as well as higher order read-out weights, decision rules, and attention are combined together to generate the observed changes in

performance (Seitz, 2017). This model suggests that to achieve generalization of learning one should find stimuli that would activate a broader range of neural and cognitive processes during learning. For example, Kowalsky, Depireux & Shamma (1996) found that auditory cortical responses for broad-band “complex” stimulation were more robust than for narrow-band stimuli “pure” stimulation. Moreover, event related potential (ERP) recordings in humans find that broad-band frequencies are better perceived and are easier to recall than narrow-band frequencies (Ahlo et al., 1996; Tervaniemi, Schröger, Saher & Näätänen, 2000; Tervaniemi, 2003). Additionally, a number of studies suggest musical structures facilitate neural encoding (Brattico, Näätänen & Tervaniemi, 2002; Larrea-Mancera, Rodríguez-Agudelo & Solís-Vivanco, 2017) and generalization of learning (Schellenberg, 2004; Kraus & Chandrasekaran, 2010).

Here, we sought to address the extent to which the generalization of tactile perceptual is mediated by the complexity of the training stimuli. Participants discriminated either sequences of narrow-band tones or broad-band “tactile music” (described below), presented on vibro-tactile effectors. We investigated how resultant learning generalized to untrained stimuli. We hypothesized that broad-band stimulation training would produce generalization to underlying stimulus dimensions such as frequency and duration discrimination. To our knowledge, no studies in the tactile domain have examined how broad-

band stimuli, which may be considered to be more ecological, might yield different patterns of generalization from training.

METHODS

Participants

We recruited 46 undergraduate students from the University of California, Riverside (13 male, mean age=20.3, SD=2.18), who were paid $10-15 an hour based on performance. They were randomly assigned to either of 2 training groups and completed 10 sessions over a period of 2 weeks. 4 participants were excluded for poor performance during training. All participants signed an informed consent, as approved by the UCR Human Subject Review Board, reported normal hearing and vision, and no history of psychiatric or neurological disorders.

Materials

All experiments were controlled using a Mac Mini (Apple, Inc., Cupertino, CA) running OSX 10.5.6. Tactile stimulation was delivered using vibro-tactile electromagnetic solenoid-type stimulators and a Dancer Design vibro-tactile amplifier tactamp 4.2 (Dancer Design, 2017). Stimulation patterns were

generated in Matlab (Mathworks Inc., Natick MA), with the use of Psychophysics Toolbox (Brainard, 1997).

Stimuli

Training Sessions

Participants trained on a match-to-sample task discrimination task using one of two types of stimulation (see fig. 2.2).

Narrow-Band Group

This group experienced stimulus sequences made up of 8 frequencies (16, 32, 64, 128, 256, 512, 1024 and 2048 Hz) with 0.25 seconds of duration each, presented in pseudorandom order with no repeats for a total sequence duration of 2 sec.

Broad-band Group

This group experienced vibratory ‘music’ patterns composed by a British music studio for the specific use with vibro-tactile stimulators. These sequences were made up of spectrally broad-band sounds laid out in time with musical rhythm.

Test Sessions

We evaluated generalization of tactile discrimination, by testing discrimination of 4 (untrained) types of vibro-tactile stimuli:

Frequency Test

Narrow-band vibration stimuli of 0.5s duration that adaptively varied above a baseline of 128Hz.

Duration Test

Narrow-band vibration stimuli of 128Hz frequency, adaptively varying in duration above a 0.5s baseline.

Didgeridoo Test

An untrained broad-band sequence of vibrations made by an Australian traditional instrument with stimulus differences manipulated in the same manner as used in trained (see below for details). The didgeridoo was chosen for the purpose of including a type of broad-band stimulation as different as possible from the utilized stimulation for training, but which was also musical. Since the didgeridoo has an unusual spectra with instances of “missing fundamental” frequencies, it was a suitable candidate for our test of broad-band perception with a novel stimuli set.

Untrained-Fingers Test

Trained broad-band stimuli (described above) was used to test untrained fingers (homologous to the trained fingers).

Procedure

In each session, participants laid their hands on a piece of foam (to dampen spread of vibrations) and placed the middle finger of one hand on one stimulator and the index finger of the other hand on the other stimulator. Headphones playing white noise were worn to prevent auditory feedback of the tactile stimulation. In each trial, a sequence of two mechanical vibro-tactile stimuli were delivered to the fingertips: a first stimulus, a 1s inter-stimulus interval (ISI), then a second stimulus. The task was to report (within 5 seconds) if the second stimulus matched the first (yes or no) via foot-pedals (‘left’ or ‘right’). Visual response feedback was provided (see fig. 2.1). On Day 1, a practice was given of 15 trials of each task first with the left-index and right-middle fingers) and again with the right-index and left-middle fingers.

To adjust difficulty, the difference between stimuli was modified adaptively via a 2-down 1-up staged staircase (Levitt, 1971). That is, after 2 correct responses, the difference between stimuli was reduced (making the judgment harder). After 1 incorrect response, the difference would increase (making the judgment easier). The size of the difference adjustment (step sizes) decreased progressively: 20% change for the first two reversals, 15% for the third, 10% for the fourth, and 5% from the fifth and on. In the *Frequency Test* and *Duration Test*, adaptive adjustments were made to the frequency or duration differences between the first and second stimuli, respectively.

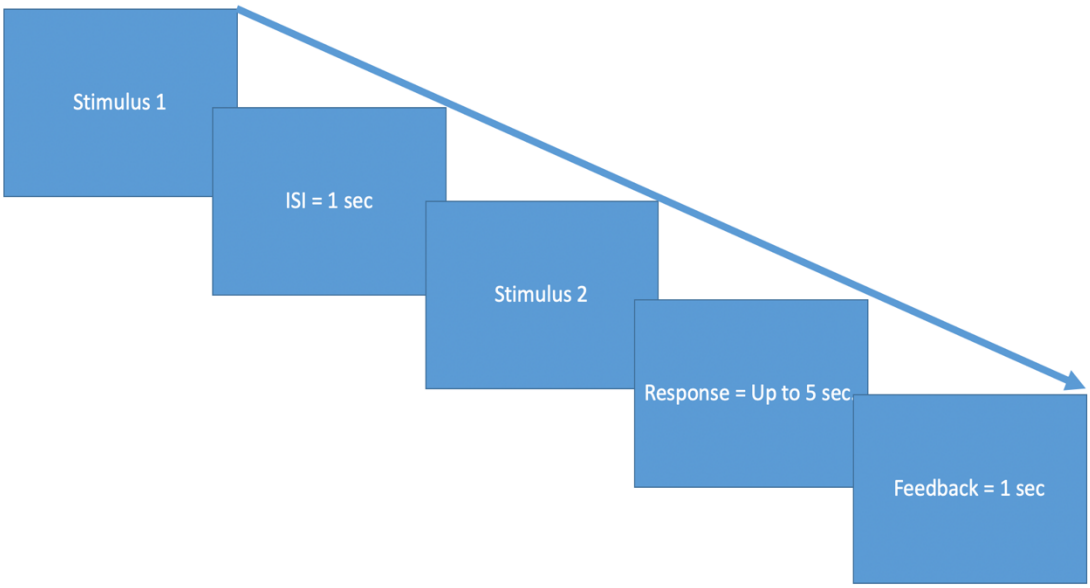

Figure 2. 1 Diagram of a single trial. Stimulus 1 (sample) was followed by an inter-stimulus-interval (ISI) of 1 second. This was followed by the presentation of stimulus 2 (match or non-match). Participants had 5 seconds to make a response. Immediately following the participant's response feedback was presented for 1 second.

For the *Broad-band Training; Narrow-band Training; Untrained-Fingers Test; and Didgeridoo Test,* we used a re-sampling procedure to determine the difference between first and second stimuli. For any one trial, the stimuli were divided into 8 equal segments. Each section was pseudo-randomly allocated to either be stretched or compressed (by changing the sampling rate of each segment). Half of the segments were stretched and half compressed, keeping the total duration of the stimulus unchanged. The magnitude of this re-sampling procedure was the adaptive parameter for these tasks. That is, the extent of

stretch/ compression for each segment was larger in easy trials (bigger difference between stimuli one and two), and smaller in harder trials.

Test sessions consisted of 60 trials for each of the four tasks (order randomized) and were conducted on the second and tenth day. Each testing session lasted around 10 minutes.

Training sessions consisted of 200 trials each and were conducted on days 3-9. These were divided into 5 blocks. Feedback (correct/ incorrect) was presented after each trial, and a score on a ten-point scale was presented after each block based on staircase. This score was used to assign monetary bonus. In each block, participants with scores of 9 received an extra $0.5, and scores of 10 received an extra $1. Participants were always trained on a middle and index finger concurrently (whether the index/ middle was left/ right was pseudo-randomly assigned). Each training session took about 40 minutes to finish.

Data Analysis

The threshold used for analyses was the median of the last 6 reversals of each test, or the last 24 reversals of training. Threshold values beyond +/-2 SD were considered outliers and were removed from group level analyses. We first looked at the training data to evaluate learning effects over days (A). To do so, we conducted two mixed-model ANOVAs (one for thresholds and reaction times, separately). We used the within-subject factor Session (Pre vs Post), and the

between-subject factor Group (Narrow vs Broad-Band). We next looked at test performance, to evaluate generalization of learning to different stimulus features or different (untrained fingers) (B). This was achieved by conducting four (one per test type) 2 X 2 mixed-model ANOVAs also with within-subject factor: Time (Pre vs Post); and between-subject factor: Group (Narrow-band vs Broad-band). Significant main effects and interactions were followed-up with post-hoc tests, namely related-samples t-tests with Bonferroni corrections for multiple comparisons.

RESULTS

Training Results

To address learning on the training tasks, we contrasted performance of the first vs last training days. Thresholds improved on both tasks (Fig. 2.2; main effects of Session; $F_{(1,40)}=23.06$, $p<.001$, $\eta^2=.345$) as did reaction times ($F_{(1,40)}=25.73$, $p<.001$, $\eta^2=.38$). We found no group differences (thresholds, $F_{(1,40)}=0.16$, $p=.689$, $\eta^2=.005$; reaction times, $F_{(1,40)}=2.608$, $p=.114$, $\eta^2=.061$), nor interactions with group that were statistically significant. Post hoc tests showed significant improvements in threshold (broad-band, $t_{(20)}=2.77$, $p=.024$, Cohen's $d=.61$; narrow-band , $t_{(20)}=3.93$, $p=.001$, Cohen's $d=.86$) and reaction times (broad-band,

$t_{(18)}$=2.77, p=.024, Cohen's d=.606); narrow-band, $t_{(20)}$=4.54, p<.001, Cohen's d=.991) for both groups.

Generalization of Learning

To understand generalization, we compared changes across test sessions (see fig. 2.3).

Frequency Test

Data from this task addresses the hypothesis that training generalized to component frequencies of the training tasks. Impressively, we found a significant interaction between Time and Group ($F_{(1,35)}$=4.73, p=.036, η^2=.119), suggesting greater generalization to frequency discrimination from broad-band vs narrow-band training. These results suggest an advantage of training with broad-band compared to narrow band stimuli to frequencies discrimination.

Duration Test

Data from this task addresses the hypothesis that training generalized to duration discrimination, another component of both training tasks. Again, while the magnitude of the broad-band effect was greater, the interaction failed to reach significance ($F_{(1,38)}$=2.76, p=.105, η^2=.055), however, a significant main effect of Time ($F_{(1,38)}$=9.41, p=.004, η^2=.188) is suggestive of learning in both groups. This shows that broad-band training transferred as much, or possibly more, to discriminating basic tone durations.

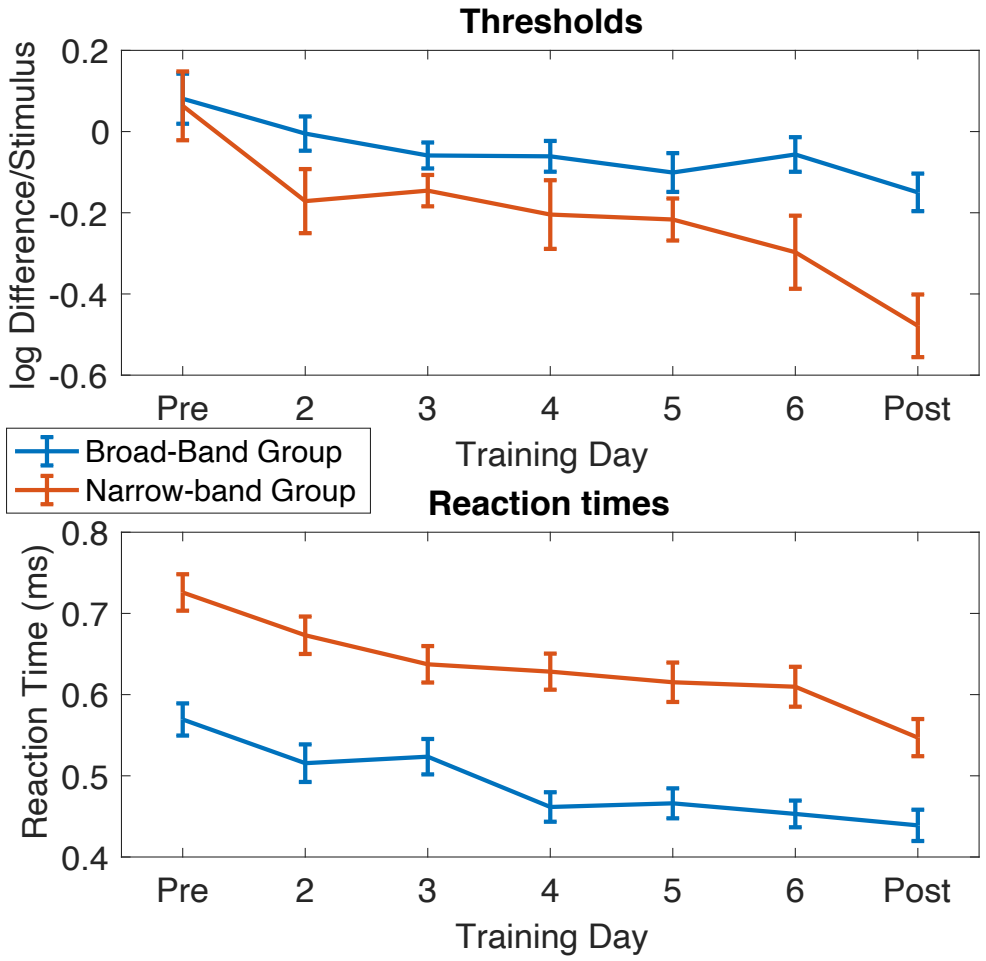

Figure 2. 2: Training Progression. Performance for each training session for both groups. Top, thresholds for each session (lower numbers mean improved performance). Bottom, reaction-times for the last 100 trials of each training session (higher numbers indicate worse performance). Error bars show within subjects standard error.

Didjeridoo Test

Data from this task addresses the hypothesis that training generalized to a novel broad-band vibro-tactile stimulus. Here we failed to find an interaction between training groups ($F_{(1,39)}$=1.60, p=.213, η^2=.038), nor a significant effect of Time ($F_{(1,39)}$=1.56, p=.219, η^2=.037). This suggests that neither training led to generalization of learning to a new set of broad-band vibrations.

Untrained-Fingers Test

Data from this task addresses the hypothesis that training generalized to un-trained digits. Here, we found a significant interaction ($F_{(1,35)}$=7.21, p=.011, η^2=.098), this time favoring the narrow-band group. This suggests that broad-band training generalizes to untrained digits to a lesser extent than the narrow-band training.

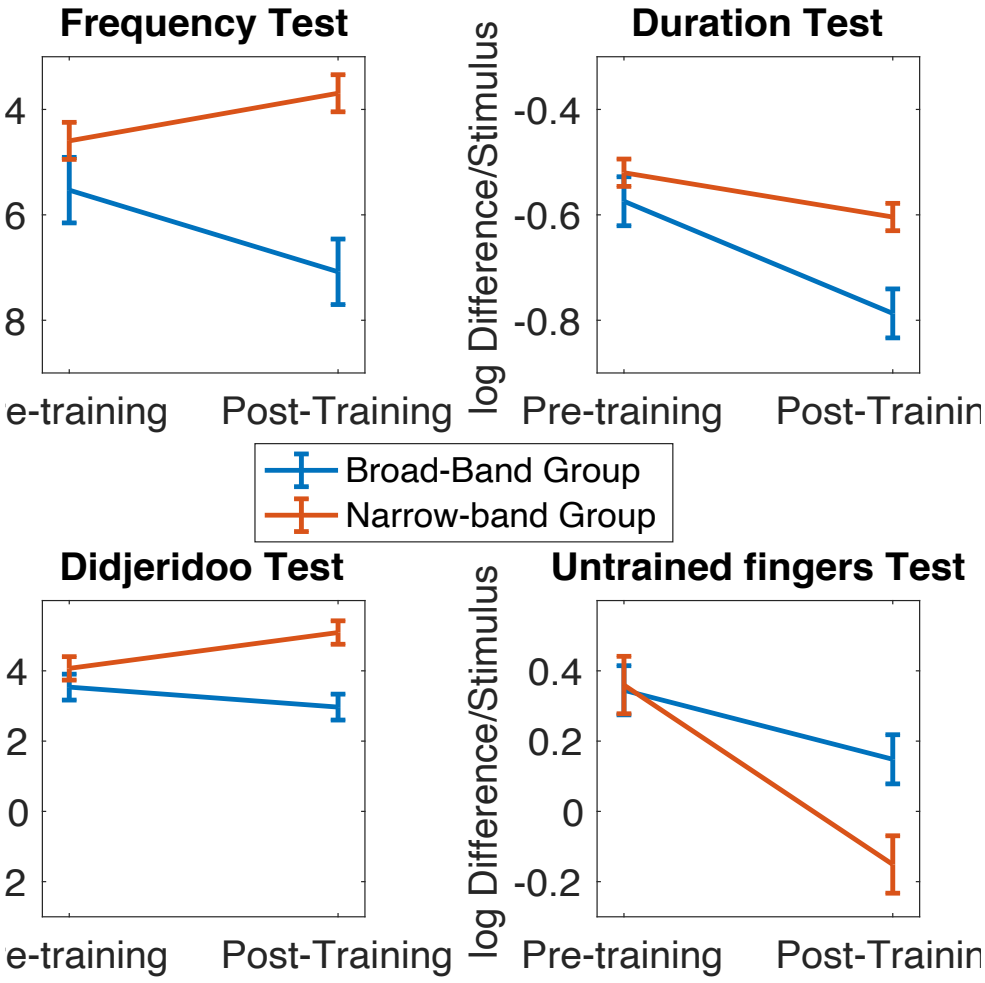

Figure 2. 3: Generalization of learning. Lower values indicate better performance. Error bars indicate within subjects SEM.

	Group	PRE-	POST-
Training	N-B	0.06 (.08)	-0.47 (.07)
	B-B	0.08 (.06)	-0.14 (.04)
Frequency Test	N-B	-0.45 (.03)	-0.36 (.03)
	B-B	-0.55 (.06)	-0.70 (.06)
Duration Test	N-B	-0.52 (.02)	-0.60 (.02)
	B-B	-0.57 (.04)	-0.78 (.04)
Didgeridoo Test	N-B	0.40 (.03)	0.50 (.03)
	B-B	0.35 (.06)	0.29 (.03)
Untrained-fingers	N-B	0.35 (.08)	-0.15 (.08)
	B-B	0.34 (.06)	0.14 (.06)

Table 2. 1: Generalization of learning summary. Shows the means and within-subjects standard error in parenthesis for the pre- and post-training threshold measures reported. N-B stands for Narrow-Band and B-B for Broad-Band Groups.

DISCUSSION

Here, we incorporate knowledge from perceptual neuroscience to investigate how to optimise the provision of successful substitutionary sensory feedback. In addition to the invasive nature of brain stimulation required (Flesher et al., 2016), a key challenge for implementing artificial tactile feedback is determining where to provide that feedback. When stimulating a nerve directly, the perceived location on the body is often displaced from the desired location on the artificial hand (Hakonen, Piitulainen & Visala, 2015). With the more common non-invasive

interfaces, the stimulation is presented on a nearby skin surface (e.g. the arm, Schofield, Evans, Carey & Hebert, 2014). Although participants show an ability to interpret substitutionary feedback to a displaced body part, the contributions to visually guided motor control are minimal (Saunders & Vijayakumar, 2011; Schofield, Evans, Carey & Hebert, 2014). This failure is likely due to the high cognitive and perceptual demand required to translate substitutionary feedback (differing in type and location) onto the artificial hand.

Using training to create new sensori-motor contingencies between the stimulated region and the artificial limb might help address these difficulties. It has been proposed that key to any type of perceptual experience –including the use of technology– depends crucially in the coupling of the sensory inflow and motor outflow as we explore the environment (O'Reagan & Noë, 2001). Research on the sensori-motor integration domain using tasks that afford exploratory behavior including tightly coupled sensory and motor components is needed to address this possibility.

For this study, we focused on the sensory stimulation component of the interaction and hypothesized that training with broad-band vibrations based on music would generalize more broadly to untrained conditions than training on narrow-band stimuli. The results of the *Frequency Test and the Duration tests are largely* consistent with this hypothesis – as we found broad-band training conferred an advantage on a separate (untrained) task assessing frequency

perception (with similar trends in the Duration test that failed to reach significance). However, neither training generalized to an untrained complex stimulus (*Didgeridoo test*). Interestingly, generalization of learning was greater for the narrow-band training than for the broad-band training to untrained digits.

The benefits seen from broad-band training may reflect the diversity of frequencies and durations present in the trained stimuli. These could promote a higher-order learning of statistical type of regularities that then transfers to their basic components (Wang et al., 2016). This type of facilitation has previously been reported in the auditory domain with musical stimuli (Brattico, Näätänen & Tervaniemi, 2002; Schellenberg, 2004; Kraus & Chandrasekaran, 2010; Larrea-Mancera, Rodríguez-Agudelo & Solís-Vivanco, 2017). However, the lack of generalization to the *Didgeridoo test* brings to question whether this was due to a difference in the component frequencies that are required for accurate discriminations compared to that in the broad-band training, or that the lack of generalization relies upon higher level components of learning. Future research will be required to differentiate between these possibilities.

Notably, the narrow-band group showed greater generalization of learning to untrained digits. This may suggest some independence between mechanisms that guide stimulus dimensions and ones that are devoted to body maps. While the greater specificity to digits in the broad-band group could represent a lower-level interpretation of learning involving refinement of receptive fields to the

trained digits (Recanzone et al., 1992; Ejaz, Hamada & Diedrichsen, 2015), an alternative explanation is that the broadband stimuli led to a more narrow focus of attention to the trained compared to the untrained digits (e.g. see Puckett, Bollmann, Barth & Cunnington, 2017). Further work will be necessary to better understand these mechanisms.

This study is a first step towards understanding factors that influence generalization of tactile perceptual learning. The present results suggest some benefits of training with broad-band stimuli. Future research will be needed to better understand the extent to which this relies upon unshared feature primitives, and/or the extent to which learning is related to higher-level features. Likewise, whether generalization to untrained digits is informative to the level of learning within the system is unclear. Understanding these mechanisms of tactile perceptual learning can have significant consequences to integrating substitutionary tactile feedback.

REFERENCES

[1] Wolpert, D. M., Ghahramani, Z., & Jordan, M. (1995). An Internal Model for Sensorimotor Integration. *American Association for the Advancement of Science*, *269*(5232), 1880–1882.

[2] Richardson, A. G., Attiah, M. A., Berman, J. I., Chen, H. I., Liu, X., Zhang, M., … Lucas, T. H. (2016). The effects of acute cortical somatosensory deafferentation on grip force control. *Cortex*, *74*(2), 1–8. https://doi.org/10.1016/j.cortex.2015.10.007

[3] Bensmaia, S. J., & Miller, L. E. (2014). Restoring sensorimotor function through intracortical interfaces: Progress and looming challenges. *Nature Reviews Neuroscience*, *15*(5), 313–325. https://doi.org/10.1038/nrn3724

[4] Kuiken, T. A., Lock, B. A., Lipschutz, R. D., Miller, L. A., Stubblefield, K. A., & Englehart, K. B. (2016). Targeted Muscle Reinnervation for Real-time Myoelectric Control of Multifunction Artificial Arms. *Journal of the American Medical Association*, *301*(6), 619–628. https://doi.org/10.1001/jama.2009.116

[5] Tan, D., Schiefer, M., Keith, M. W., Anderson, R., & Tyler, D. J. (2013). Stability and selectivity of a chronic, multi-contact cuff electrode for sensory stimulation in a human amputee. *International IEEE/EMBS Conference on Neural Engineering, NER*, 859–862. https://doi.org/10.1109/NER.2013.6696070

[6] Horch, K., Meek, S., Taylor, T. G., & Hutchinson, D. T. (2011). Object discrimination with an artificial hand using electrical stimulation of peripheral tactile and proprioceptive pathways with intrafascicular electrodes. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *19*(5), 483–489. https://doi.org/10.1109/TNSRE.2011.2162635

[7] Schofield, J. S., Evans, K. R., Carey, J. P., & Hebert, J. S. (2014). Applications of sensory feedback in motorized upper extremity prosthesis: A review. *Expert Review of Medical Devices*, *11*(5), 499–511. https://doi.org/10.1586/17434440.2014.929496

[8] Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature, 287*, 43–44.

[9] Merzenich, M., Recanzone, G., Jenkins, W., Allard, T., and Nudo, R.J. (1988) Cortical Representational Plasticity. In P. Rakic and W. Singer eds., *Neurobiology of Neocortex*, 41-67. New York, Wiley.

[10] Karni, A., Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proc Natl Acad Sci*., 88:4966-4970.

[11] Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science, 256*, 1018–1021. doi:10.1126/science .1589770

[12] Sagi, D. (2011). Perceptual learning in vision research. *Vision Research*, 51, 1552–1566. doi:10.1016/j.visres.2010.10.019

[13] Mountcastle VB, Steinmetz MA, Romo R (1990) Frequency discrimi- nation in the sense of flutter: psychophysical measurements correlated with postcentral events in behaving monkeys. *J Neurosci* 10:3032–3044.

[14] Recanzone G.H., Merzenich M.M., Jenkins W.M., Grajski K.A., Dinse H.R. (1992) Topographic reorganization of the hand representation in cortical area 3b owl monkeys trained in a frequency-discrimination task. *J Neurophysiol* 67: 1031–1056.

[15] Hernández, A., Salinas, E., Garcıa, R., & Romo, R. (1997). Discrimination in the Sense of Flutter: New psychophysical measurements in monkeys. *The Journal of Neuroscience*, *17*(16), 6391–6400.

[16] Harris, J. A., Harris, I. M., & Diamond, M. E. (2001). The topography of tactile learning in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *21*(3), 1056–61.

[17] Harrar, V., Spence, C., & Makin, T. R. (2014). Topographic generalization of tactile perceptual learning. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(1), 15–23. http://doi.org/10.1037/a0033200

[18] Dempsey-Jones, H. E., Harrar, V., Oliver, J., Johansen-Berg, H., Spence, C., & Makin, T. R. (2015). Transfer of tactile perceptual learning to untrained

neighbouring fingers reflects natural use relationships. *Journal of Neurophysiology*, *44*(0), jn.00181.2015. http://doi.org/10.1152/jn.00181.2015

[19] Sathian K, Zangaladze A. (1997). Tactile learning is task specific but transfers between fingers. *Percept Psychophys 59*: 119–128.

[20] Spengler F, Roberts TPL, Poeppel D, Byl N, Wang X, Rowley HA, Merzenich MM (1997) Learning transfer and neuronal plasticity in humans trained in tactile discrimination. *Neurosci Lett* 323:151–154.

[21] Seitz, A. R. (2017). Perceptual learning. *Current Biology*, *27*(13), R631–R636. http://doi.org/10.1016/j.cub.2017.05.053

[22] N. Kowalski, D. Depireux, and S. Shamma. (1996). Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra. *Journal ofNeurophysiology*, vol. 76, no. 5, pp. 3503–3523.

[23] Alho, K., Tervaniemi, M., Huotilainen, M., Lavikainen, H., Tiitinen, R., Ilmonemi, R., … Näätänen, R. (1996). Processing of complex sounds in the human auditory cortex as revealed by magnetic brain responses. *Psychophysiology*, *33*, 369–375.

[24] Tervaniemi, M., Schröger, E., Saher, M., Näätänen, R. (2000). Effects of spectral complexity and sound duration on automatic complex-sound pitch processing in humans: a mismatch negativity study. *Neuroscience Letters* 290 (1), 66-70.

[25] Tervaniemi, M. (2003). Musical sound processing: EEG and MEG evidence. In Peretz, I. and Zatorre, R. (eds.). *The Cognitive Neuroscience of Music*, (pp. 21-31) Oxford: Oxford University Press.

[26] Brattico, E., Näätänen, R., Tervaniemi, M. (2002). Context effects on pitch perception in musicians and non-musicians: Evidence from ERP recordings. Music Percept. 19, 199–222.

[27] Lelo de Larrea-Mancera, E. S., Rodríguez-Agudelo, Y., & Solís-Vivanco, R. (2017). Musical rhythm and pitch: a differential effect on auditory dynamics as

revealed by the N1/MMN/P3a complex. *Neuropsychologia*, *100*(January), 44–50. http://doi.org/10.1016/j.neuropsychologia.2017.04.001

[28] Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychol. Sci.*, *15*(8), 511–514.

[29] Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nat Rev Neurosci*, *11*(8), 599–605. http://doi.org/nrn2882 [pii]\n10.1038/nrn2882

[30] Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.

[31] Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

[32] Flesher, S. N., Collinger, J. L., Tyler-Kabara, E. C., Bensmaia, S. J., Gaunt, R. A., Boninger, M. L., … Weiss, J. M. (2016). Intracortical microstimulation of human somatosensory cortex. *Science Translational Medicine*, *8*(361), 1–10. https://doi.org/10.1126/scitranslmed.aaf8083

[33] Hakonen, M., Piitulainen, H., & Visala, A. (2015). Current state of digital signal processing in myoelectric interfaces and related applications. *Biomedical Signal Processing and Control*, *18*, 334–359. https://doi.org/10.1016/j.bspc.2015.02.009

[34] Saunders, I., & Vijayakumar, S. (2011). The role of feed-forward and feedback processes for closed-loop prosthesis control. *Journal of NeuroEngineering and Rehabilitation*, *8*(1), 60. https://doi.org/10.1186/1743-0003-8-60 22032545

[35] O'Regan, J., & Noë, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, 79–103. Retrieved from http://www.springerlink.com/index/nl361h8276502488.pdf

[36] Wang, R., Wang, J., Zhang, J.-Y., Xie, X.-Y., Yang, Y.-X., Luo, S.-H., … Li, W. (2016). Perceptual Learning at a Conceptual Level. *Journal of*

Neuroscience, *36*(7), 2238–2246. https://doi.org/10.1523/JNEUROSCI.2732-15.2016

[37] Ejaz N, Hamada M, Diedrichsen J. (2015). Hand use predicts the structure of representations in sensorimotor cortex. *Nat Neurosci 18*: 1034–1040.

[38] Puckett, A.M., Bollmann, S., Barth, M. & Cunnington, R. (2017). Measuring the effects of attention to individual fingertips in somatosensory cortex using ultra-high field (7T) fMRI. *NeuroImage* (1), 179-187.

CHAPTER THREE: Portable Automated Rapid Testing (PART) for auditory assessment: Validation in a young adult normal-hearing population

This chapter presents a novel and accessible tool for the assessment of the mechanical sense of hearing. We show the validity and reliability of several assessments of central auditory function in a portable platform of automated rapid testing suitable for research in perceptual learning with potential to supplement clinical practice. This work was presented at the 177th meeting for the Acoustical Society of America in Louisville, KY. The published version of the manuscript was a P&P Technical Area Selection of the Journal of the Acoustical Society of America (March, 2021), and can be found online here: https://doi.org/10.1121/10.0002108

ABSTRACT

This study aims to determine the degree to which Portable Automated Rapid Testing (PART), a freely-available program running on a tablet computer, is capable of reproducing standard laboratory results. Undergraduate students were assigned to one of three within-subject conditions that examined repeatability of performance on a battery of psychoacoustical tests of temporal fine structure processing, spectro-temporal amplitude modulation, and targets in

competition. The Repeatability condition examined test/retest with the same system, the Headphones condition examined the effects of varying headphones (passive and active noise-attenuating), and the Noise condition examined repeatability in the presence of recorded cafeteria noise. In general, performance on the test battery showed high repeatability, even across manipulated conditions, and was similar to that reported in the literature. These data serve as validation that suprathreshold psychoacoustical tests can be made accessible to run on consumer-grade hardware and performed in less controlled settings. This dataset also provides a distribution of thresholds that can be used as a normative baseline against which auditory dysfunction can be identified in future work.

INTRODUCTION

The assessment of auditory function in modern clinical audiology was translated from the laboratory in the middle of the previous century (Carhart & Jerger, 1959, Hughson & Westlake, 1944), and has remained focused on using pure-tone audiograms to evaluate audibility and speech tests to assess the ability to detect particular acoustical cues in speech (see CHABA, 1988). These clinical assessments are targeted at diagnosis of hearing impairment based on audibility and on an approach to rehabilitation that is largely defined by its reliance upon amplification via hearing aids or cochlear implants. This focus on audibility and amplification has provided little incentive for clinical care to include the assessment and rehabilitation of supra-threshold auditory processing disabilities. As a result, there are very few tools, and even fewer protocols, available for the diagnosis and/or treatment of auditory difficulties that are not accompanied by losses of audibility. The diagnostic and rehabilitative approaches that do exist are regarded as specialized tools to be used by those clinicians who work with children or adults with suspected auditory processing disorders (APDs). There is a long history of clinicians and scientists using the term APD (e.g., Iliadou et al., 2018); yet, some clinicians and researchers are uncomfortable with the term due to the potential overlap of APD with language and cognitive dysfunction (e.g., Moore, 2018). The perspective taken by this study is that regardless of the

clinical status of APD, it is undeniably the case that tests of auditory perceptual abilities (e.g., Moore et al., 2014; Eddins & Hall, 2010; Gallun et al., 2013) have the potential to shed light on complaints of hearing difficulties that are only weakly predicted by the audiogram or performance on clinical speech tests (Hoover et al., 2017; Eckert et al., 2017; Souza et al., 2018).

Clinically accessible tests of functional hearing are needed to better understand self-reported difficulties with auditory perception and poor performance on laboratory tests of auditory processing. These tests would need to be applied and validated across a population with diverse hearing abilities in order to clearly characterize which measures are most informative about the variety of hearing difficulties experienced by individual listeners or groups of listeners. Although a number of candidate tests have been developed and are relatively well studied in laboratory settings (e.g. Moore, 1987; Grose and Mamo, 2012; Bernstein et al., 2013; Gallun et al., 2014; Füllgrabe, Moore & Stone, 2015; Jakien et al., 2017; Hoover, Souza & Gallun, 2017; Hoover et al., 2019), very few of these tests have been translated into standard clinical practice. Those tests that have been translated into the clinic are generally only used by audiologists with expertise in APDs because the testing often requires specialized equipment or setup and a calibrated audiometer. Even when the tests are built into the audiometer, many audiologists have not received adequate training to feel comfortable administering, scoring, and interpreting the tests.

Tests that have moved successfully from the laboratory to the clinic include: the Staggered Spondaic Words test (SSW; Katz, 1962; Arnst, 1981), the Gaps in Noise test (GIN; Plomp, 1964; Green, 1971), the Masking Level Difference (MLD; Hirsh, 1948; Olsen, Noffsinger, Carhart, 1976), the Dichotic Digits Test (DDT; Broadbent, 1958; Musiek, 1983), the Listening in Spatialized Noise test (LISN; Cameron & Dillon, 2007; Glyde et al., 2013), the Frequency Patterns Test (FPT; Musiek & Pinheiro, 1987; Musiek, 1994), and the Dichotic Sentences Test (DST; Fifer et al., 1983). In addition, the Screening test for auditory processing (SCAN; Keith, 1995) is a battery of assessments that incorporates multiple auditory processing abilities. While these and other tests have been used successfully both in the laboratory and in the clinic to identify auditory processing dysfunction (e.g., Gallun et al., 2012; 2016; Hoover et al., 2017), none of them are portable, automated, or rapid. They all require specialized equipment, such as an audiometer, and demand a trained audiologist to administer (most take at least 30 minutes) and score them by hand. The goal of this research project is to supplement these well-established tests with a low-cost, portable test system that could be used to administer a key set of basic auditory processing tests that is scored automatically and requires minimal clinical involvement. The assessments should each be rapid enough that clinicians and clinical researchers could tailor the length of the test battery to the time available. Moreover, portable automated rapid testing could play an

essential role in gathering the datasets necessary to better characterize the auditory processing abilities and difficulties of individual listeners relative to the expected abilities of other listeners of a similar age with similar audiometric thresholds. Without this information, the clinician will continue to have difficulty appropriately identifying and remediating the auditory processing dysfunction they observe in their patients.

To address this gap, several state-of-the-art psychometric tests currently used in the laboratory to research central auditory processes have been translated into the application PART (Portable Automatic Rapid Testing) developed by the University of California Brain Game Center (https://braingamecenter.ucr.edu). PART can run both on mobile devices (e.g. iPad, iPhone, Android) and standard desktop computers (MacOS, Windows) and is currently freely available on the Apple App Store, the Google Play Store, and the Microsoft Store. PART has proven capable of accurately reproducing precise acoustic stimuli on an iPad (Apple, Inc., Cupertino, CA) with Sennheiser 280 Pro headphones (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) at output levels set by the built-in calibration routine (Gallun et al., 2018).

The psychophysical test battery evaluated here was designed to reflect a description of the central auditory system inspired by current research in psychoacoustics and auditory neuroscience (e.g., Stecker & Gallun, 2012; Bernstein et al., 2013; Depireux, Simon, Klein & Shamma, 2000). This test

battery is comprised of three sub-batteries, each with supporting evidence of clinical utility: temporal fine structure processing, spectro-temporal amplitude modulation, and targets in competition. These three groups of tests address different stages of auditory processing in the central nervous system that together mediate our ability to parse the auditory scene (Bronkhorst, 2015; Gallun and Best, in press). We note that the test battery reported in this manuscript represents only a small subset of PART's functionality and the PART platform facilitates a wide range of psychoacoustical tests.

Temporal Fine Structure (TFS) coding is assumed to rely upon the precision of phase-locking in populations of auditory nerve fibers responding to movements of the cochlear partition (Pfeiffer & Kim, 1975). The fine temporal information carried by the auditory nerve serves as the input to both the binaural system (see Stecker and Gallun, 2012) and the monaural pitch system (see Winter, 2005). Further refinement of this and other spectral and temporal information carried by the auditory nerve is responsible for the spectro-temporal modulation (STM) sensitivity observed in the inferior colliculus (Versnel, Zwiers & Opstal, 2009) and auditory cortex (Kowalski, Depireux & Shamma, 1996). TFS sensitivity has been evaluated psychophysically using both monaural and binaural stimuli (Grose & Mamo, 2012; Gallun et al., 2014; Hoover et al., 2019). Neither the audiogram, nor most conventional speech tests, evaluate the detection of frequency modulation, or use spatialization of auditory signals; yet, it

has been found that TFS measures are a good predictor of speech understanding in competition (Füllgrabe, Moore & Stone, 2015) and are suitable tests for age-related temporal processing variability (Grose & Mamo, 2012; Gallun et al., 2014; Füllgrabe, Moore & Stone, 2015). In this study, diotic frequency modulation was used to assess monaural TFS sensitivity, and dichotic frequency modulation was used to assess binaural TFS sensitivity. A temporal gap detection test (inter-tone burst delay) was also used to assess the sensitivity of temporal processes (Gallun et al., 2014). Because gap discrimination can be performed either using TFS information or by using envelope information carried by the auditory nerve (and refined by later processing), it is important to note that it is presently unclear which cue(s) are being evaluated, or even whether or not gap discrimination evaluates the same cues among different listeners. Nevertheless, these three tests have been proposed previously as measures of TFS with potential clinical utility (Hoover et al., 2019), and so, that category label is retained here for ease of reference.

The preferential tuning of auditory cortical neurons to modulation, both over time and across frequency, has resulted in an increased focus on the potential explanatory power of spectro-temporal modulation perception (Kowalski, Depireux & Shamma, 1996; Theunissen, Sen & Doupe, 2000; Shamma, 2001; Schonwiesner & Zatorre, 2009). All natural sounds can be characterized as a pattern of spectro-temporal modulation (Theunissen, Sen &

Doupe, 2000; Theunissen & Elie, 2014) and the relationship between sinusoidal spectro-temporal modulation and speech stimuli has been appreciated for some time (e.g., van Veen and Houtgast, 1985). This has led to a number of studies exploring sensitivity to spectral-, temporal-, and spectro-temporal modulation (STM) both for non-speech stimuli (e.g. Whitefield & Evans, 1965) and for speech stimuli (Bernstein et al., 2013; Mehraei et al., 2014; Venezia et al., 2019) as central processes that exist beyond basic audibility (Gallun & Souza, 2008). Studies using STM in participants with supra-threshold hearing loss have found that an extra 40% of the variance of speech-in-noise performance can be accounted for by these evaluations beyond the 40% accounted for by the audiogram alone (Bernstein et al., 2013; Mehraei et al., 2014). Thus, this study included tests for temporal-, spectral- and STM sensitivities, all of which are largely absent from the clinic.

Because the accurate identification of an acoustic target in competition is considered fundamental to auditory perception and scene analysis beyond peripheral audibility (Shinn-Cunningham, 2008; Moore, 2014; Bronkhorst, 2015), tests were included that assess the capacity of the system to select relevant information and suppress test-irrelevant interference. The notched-noise method (Patterson, 1976; Moore & Glasberg, 1990) evaluates the detection of a tone presented in competition with noise either with or without a spectral notch around the target frequency. This test allows the evaluation not only of peripheral

frequency selectivity but also frequency processing efficiency (Patterson, 1976; Moore & Glasberg, 1990; Stone et al., 1992; Bergman et al., 1992). To address auditory scene analysis including speech and binaural processing, spatial release from masking (SRM; Marrone et al., 2008; Gallun et al., 2013; Jakien et al., 2017; Jakien & Gallun, 2018) was assessed using the Coordinate Response Measure (CRM) corpus (Bolia et al., 2000). Following the methods of Gallun et al. (2013) speech understanding was assessed both with speech maskers colocated with the target speech in simulated space, as well as with the maskers separated from the target by 45 degrees in simulated space. These tests independently assess speech understanding in competition under different stimulus conditions and the difference between the scores on the two provides a measure of the ability of an individual listener to benefit from spatial differences between target and masking stimuli.

The purpose of this study was to determine the degree to which this preliminary PART battery is capable of reproducing standard laboratory results in a population of young, normal-hearing adults. To this end, the reliability of threshold estimation (test-retest) and the degree to which estimates obtained from PART approximate those reported in the literature for the same tests were both evaluated. Additionally, to address the robustness of results to different listening conditions, we evaluated the extent to which test measures were consistent across the use of different headphone types and under different

ambient noise conditions. Ultimately, the goal of this work is to generate a normative dataset that could be used in a range of contexts from research to the clinic.

To accomplish these goals, data were collected from young normal-hearing students under similar conditions to previous validation work from our group (Gallun et al., 2018) with repeated tests using circumaural headphones (Repeatability condition), using both passive and active noise-attenuating headphones in a silent environment (Headphone condition), and in the presence of recorded cafeteria noise (Noise condition). First, the results addressing measurement reliability (test-retest) are presented. Second, the relation to the relevant literature is examined. Third, the effects of the experimental manipulations involving headphones and background noise are estimated. Overall, results show that PART produces repeatable threshold estimates consistent with those that have been reported previously in the laboratory across different listening conditions. These data serve as validation that accessible auditory hardware (consumer-grade tablet and headphones) can be used to test auditory function with sufficient precision to reproduce the thresholds obtained using laboratory-grade equipment. This dataset also provides a distribution of thresholds that can now be used as a normative baseline against which auditory dysfunction can be identified in future work.

METHODS

Participants

Listeners were 150 undergraduate students from the University of California, Riverside (42 male, *M* age = 19.6 years, *SD* = 2.31 years), who received course credit for their participation. All participants reported normal hearing and vision, and no history of psychiatric or neurological disorders. They provided signed informed consent as approved by the University of California, Riverside Human Research Review Board. In alignment with our goal to evaluate “normal” auditory processing, we rejected thresholds that deviated more than 3 *SD* from the mean of each assessment from the results presented in the main manuscript. A full data set with the thresholds of all participants is included in a form suitable for further analysis, and analyses and plots with the full dataset are included in the Supplemental Materials (see Fig S1 & table ST1).

Materials

All procedures were conducted using standard iPad tablets (Apple, Inc., Cupertino, CA) running the PART Portable Automatic Rapid Testing (PART) application with stimuli delivered via either Sennheiser 280 Pro headphones (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany), which are rated to have a 32 dB passive noise attenuation with an 8 Hz to 25 kHz frequency

response, or Bose (active) noise cancelling Quiet Comfort 35 wireless headphones (Bose Corporation, Framingham, MA) set to the high noise cancelling setting. Output levels were calibrated for the Sennheiser headphones using an iBoundary microphone (MicW Audio, Beijing, China) connected to another iPad running the NIOSH Sound Level Meter App (SLM app; https://www.cdc.gov/niosh/topics/noise/app.html), as described in Gallun et al., 2018). The SLM app and iBoundary microphone system were calibrated with reference to measurements made with a Head and Torso Simulator with Artificial Ears (Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark) in the anechoic chamber located at the VA RR&D National Center for Rehabilitative Auditory Research (NCRAR). Similar testing of the Bose system revealed that the method used, which did not involve changing the calibration settings when the headphones were changed, resulted in an overall reduction in the mechanical output level at the ear of 14 dB, but no distortions in the time or frequency domain. The levels described here and used throughout the study refer to the calibrated Sennheiser system.

Procedure

In each session, participants sat in a chair inside a double-walled sound-treated room and listened through a set of headphones connected to an iPad running PART. Tests were self-administered with text-based instructions delivered within

the PART application. Responses were collected via digital buttons presented on the iPad touch screen. Most tasks employed a 2-cue 2-alternative forced choice (2-Cue 2-AFC) procedure where four intervals are presented in an audio-visual sequence with inter-stimulus-intervals (ISI) of 250 ms (Fig. 3.1A top-left). The first and last stimuli were standard cues, and participants made a choice between the two alternatives presented in the second and third intervals (Fig. 3.1A top-right). Participants responded by touching the second or third square on the screen. The selected square then flashed either green (correct) or red (incorrect) as response feedback (Fig. 3.1A bottom) before proceeding to the next trial (1 sec ITI). This 2-cue 2-AFC task, which is identical to the one used in Souza et al. (2020), has the advantage that, unlike a two-interval or three-interval task, the target is always preceded and followed by a standard stimulus. This allows the task to be performed by comparing information either forward or backward in time. This is important as it is known that sensory comparisons are more difficult if they must be performed to a following standard rather than to a preceding standard, especially for older listeners (Gallun et al., 2012). A 2-cue 2-AFC design thus helps ensure that if in the future differences are found between the normative data reported here and data from other patient groups, that difference will be less likely to reflect the influences of attention or memory and more likely to reflect actual differences in the ability to make sensory comparisons. The one task that differed in procedure was the Spatial Release from Masking task (SRM)

that uses a colored number grid to respond and has a fixed progression of difficulty (see details below).

The tasks using the 2-cue 2-AFC procedure adjusted difficulty using a two-stage 2-down 1-up staircase procedure. The first stage used large steps for 3 reversals before moving on to the second stage that used smaller steps (1/5 the size of the first stage) and terminated after 6 reversals. Further, to help ensure that after incorrect responses participants were provided with easier exemplars, steps up were larger than steps down with a 2:1 step-size ratio in the Repeatability condition, and 1.5:1 step-size ratio in the Headphone and Noise conditions. Thresholds were estimated from the geometric mean of the second-stage reversals. A general schematic of the adaptive staircase procedures is included in Figure 3.1B. This combination of up-down rule and step-size ratio results in a threshold estimate that asymptotically targets the stimulus level corresponding to 81.7% correct for 2:1 and 77.5% correct for 1.5:1, comparable to a 79.4% targeted by a typical 3-down 1-up staircase with equal steps up and down (Levitt, 1971; García-Pérez, 2001; see Supplemental Materials for the comparison across procedures). While unequal step sizes are common in audiometric testing (ANSI 3.21, 2004; ISO 8253-1, 2010), there are few who have followed the suggestion of García-Pérez (2011) in adopting the use of unequal step sizes when designing efficient staircase methods. The goal is to minimize the influence of task and listener factors that can result in thresholds

deviating from the asymptotic target point (García-Pérez, 1998; 2001). Designing optimal methods for the clinical translation of laboratory procedures is a continued area of research by our group (e.g., Hoover et al., 2019). The exploration of different ratios of unequal step sizes reported here represents an initial foray into this question.

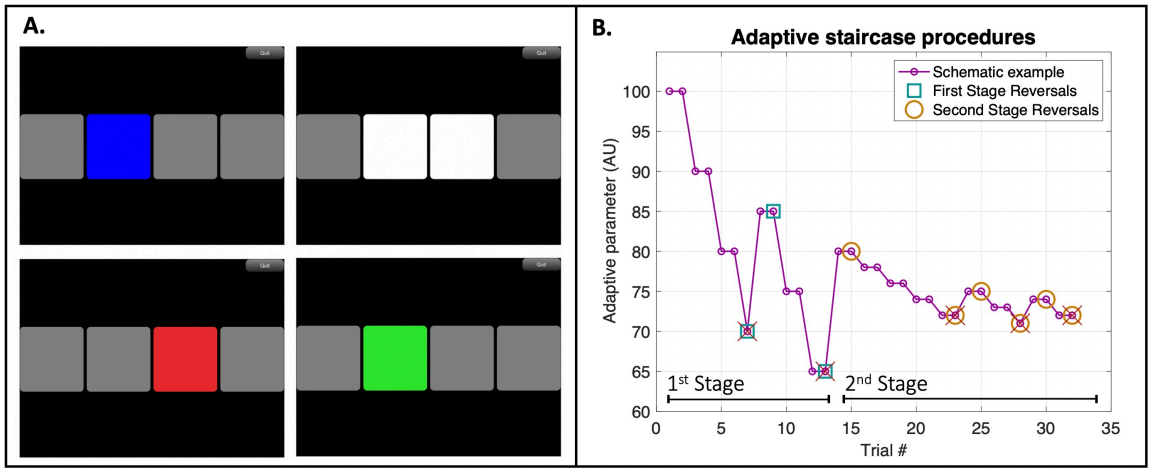

Figure 3. 1: Diagram of task and adaptive procedure. In A, each panel represents a screenshot taken from PART while on a 2-cue, 2-alternative forced-choice test. Each box lit up sequentially in blue emitting a sound (top-left). After all intervals were played the 2 alternatives in the middle became available for response (top-right). Feedback is shown by color code (red = wrong; bottom panels). In B, we present a schematic example of the adaptive staircase procedures used. The difference in the magnitude of steps between staircase stages and the unequal step sizes going up/down can be easily observed in this example. Incorrect trials are marked with crosses and reversals are marked with either squares (1st stage) or circles (2nd stage). Arbitrary units were selected as adaptive parameter values for descriptive purposes only.

Each session of the experiment began with a monitored screening test which presented 10 trials of a 2 kHz tonal target signal at 45 dB SPL in the environmental settings relative to each condition. In cases where participants failed to respond accurately on at least 9 of the 10 trials, instructions were repeated in isolation from other participants to ensure that the task was properly understood. All of those participants who needed to restart the testing reported that they did not realize that the tone would be presented at a fairly low level. Once properly prepared for the stimuli to be at 45 dB SPL, all participants were able to detect the 2 kHz tone with at least 90% accuracy. At this point, all participants moved on to complete two assessments involving the detection of the same 2 kHz tone, but now presented in noise maskers with or without a spectral notch (described in detail below). Participants then were pseudo-randomly assigned to complete the remaining eight assessments (details described below) in three blocks of testing organized by test type: Temporal Fine Structure (TFS; 3 assessments); Spectro-temporal Modulation (STM; 3 assessments); and the second half of Targets in Competition (SRM; 2 assessments). All assessments were preceded by 5 non-adaptive practice trials at a high point in their respective staircase where target stimuli were easily detectable. Participants were encouraged to take small breaks between testing blocks. All three test blocks were given each session. The ten assessments in the testing block took around 5 minutes each, resulting in test sessions of around

50 minutes. The second session was always conducted on a different day, no longer than a week after the first. Test sessions involved up to three participants seated next to each other in a single room, listening and responding independently. In general, listeners received minimal instructions regarding the proper placement of the headphones and adherence to the brief written instructions automatically delivered by PART. The full verbal and written instructions given to each listener are provided in the Supplemental Materials.

Stimuli

Visual examples of the stimuli used in each assessment are shown in Fig. 3.2.

Temporal Fine Structure (Fig. 3.2A)

Temporal Gap

This gap discrimination task (Gallun et al., 2014; Hoover et al., 2019) compares a target signal that consisted of a diotically presented temporal gap placed between two 0.5 kHz tone bursts of 4 ms played at 80 dB SPL to standards that consisted of both tone bursts sequentially with no gap between them. The adaptive parameter was an inter-tone burst delay with an initial value of 20 ms. The staircase adapted on an exponential scale with first stage step-size (down) of $2^{1/2}$ and second stage step-size (down) of $2^{1/10}$ with a minimum of 0 ms and a maximum of 100 ms.

Diotic Frequency Modulation

This FM detection task (Grose & Mamo, 2012; Whiteford & Oxenham, 2015; Whiteford et al., 2017; Hoover et al., 2019) compares a target diotic frequency modulation rate of 2 Hz to standards that consisted of a pure tone carrier frequency randomized between 460 and 550 Hz, each presented at 75 dB SPL for 400 ms. Randomization of the carrier frequency of standards ensures that the test cannot be successfully conducted by a simple pitch cue. The adaptive parameter was modulation depth with an initial value of 6 Hz. The staircase adapted on an exponential scale with first stage step-size (down) of $2^{1/2}$ and second stage step-size (down) of $2^{1/10}$ with a minimum of 0 and a maximum of 10,000 Hz.

Dichotic Frequency Modulation

This FM detection task (Grose & Mamo, 2012; Hoover et al., 2019) uses a stimulus first developed by Green et al. (1976), that creates a continuously shifting interaural phase difference (IPD) in the target interval. The task compares a target signal consisting of a frequency modulation rate of 2 Hz that is inverted or is anti-phasic between the ears to standards that consisted of a pure tone carrier frequency randomized between 460 and 550 Hz, each presented at 75 dB SPL for 400 ms. The adaptive parameter was modulation depth (which determines the size of the IPD) with an initial value of 3 Hz. The staircase adapted on an exponential scale with first stage step-size (down) of $2^{1/2}$ and

second stage step-size (down) of $2^{1/10}$ with a minimum of 0 and a maximum of 10,000 Hz.

Spectro-Temporal Sensitivity (Fig. 3.2B)

All stimuli for these tasks involved a broadband noise that was either unmodulated (the standard) or modulated temporally, spectrally, or spectro-temporally, depending on the task (described below). The unmodulated standard consisted of flat-frequency broad-band noise with a frequency range of 0.4 to 8 kHz. Stimuli were generated in the frequency domain using the maximum number of components allowed by a 44.1 kHz sampling rate with random amplitude and phase values, presented at 65 dB SPL for 500 ms. Modulation was applied on a logarithmic amplitude scale (dB) and modulation depth was measured from the middle of the amplitude range to the peak amplitude. The stimuli were generated between trials using the algorithm developed by Stavropoulos et al. (2021).

Temporal Modulation

The TM detection task (Viemeister, 1979) compares a target with sinusoidal temporal amplitude modulation (AM) at a rate of 4 Hz to the unmodulated standard. The adaptive parameter was modulation depth in dB. The staircase adapted linearly in dB with first stage step-size (down) of 0.5 dB and second stage step-size (down) of 0.1 dB with a minimum of 0.2 dB Hz and a maximum of 40 dB.

Spectral Modulation

The SM detection task (Hoover, Eddins & Eddins, 2018) compares a target with a sinusoidal spectral modulation with random phase at a rate of 2 cycles per octave (c/o) to an unmodulated standard. The adaptive parameter was modulation depth in dB, which was adaptively varied as in the TM task.

Spectro-Temporal Modulation

This STM detection task (Bernstein et al., 2013; Mehraei et al., 2014) uses similar stimuli to the TM & SM tasks described above but compares a target with both 2 cycles per octave (c/o) spectral modulation and 4 Hz AM to standards that consisted of flat-frequency broadband noise. The resulting spectro-temporal modulation (STM) was randomly assigned to move upward or downward in frequency over time on each trial. The adaptive parameter was modulation depth in dB, which was varied as in the TM and SM tasks.

Targets in competition (Fig. 3.2C)

No-Notch Condition

This abbreviated notch-noise method is adapted from Moore (1987) and measures the ability of the listener to detect a target 2 kHz pure tone presented at 45 dB SPL in only one of the four intervals. The masking noise, which occurred on all intervals, consisted of 10,000 sinusoidal components distributed exponentially (-3 dB/octave) centered on the target frequency with a bandwidth of

1600 Hz (1.2 to 2.8 kHz) presented for 500 ms. The adaptive parameter was the RMS level of the noise, measured in dB. The staircase started with a noise level of 35 dB SPL and adapted on a linear scale with first stage step-size (down) of 6 dB SPL and second stage step-size (down) of 2 dB SPL with a minimum of 25 and a maximum of 90 dB SPL.

Notch Condition

This condition was identical to the no-notch condition, with the exception that a spectral notch of 0.8 kHz was introduced, increasing the bandwidth of the masker such that it covered two frequency ranges: 0.8-1.6 kHz and 2.4-3.2 kHz, leaving a 0.8 kHz notch centered on 2 kHz, which was the frequency of the target to be detected. The adaptive parameter was the RMS masker level, which again had a starting value of 35 dB SPL. The staircase adapted in the same manner as in the no-notch condition. This condition is equivalent to a notch width of 0.2 times the center frequency of 2 kHz measured from center to the nearest edge of the noise as described by Moore (1987). The difference in threshold with the no-notch condition can be taken as an index of frequency (spectral) resolution.

SRM Colocated

The three-talker speech-on-speech masking method of Marrone et al. (2008), as adapted for progressive tracking by Gallun et al. (2013), was used to measure the ability of listeners to identify keywords of a target sentence in the presence of two masking sentences. Using a color/number grid (4 colors by 8 numbers)

participants identified two keywords (a color and a number) by selecting the position indicated by the keywords spoken by the target talker, which was a single male talker from the Coordinate Response Measure corpus (CRM, Bolia et al., 2000) presented from directly in front of the listener in a virtual spatial array. Target sentences all included the call sign “Charlie” and two keywords: a number and a color. Targets were fixed at an RMS level of 65 dB SPL. The target was presented simultaneously with two maskers, which were male talkers uttering sentences with different call signs, colors and numbers in unison with each other and with the target. All three sentences were presented from directly in front of the listener (colocated). Progressive tracking included 20 trials in which the maskers progressed in level from 55 dB SPL to 73 dB SPL in steps of 2 dB every two trials as reported in Gallun et al. (2013), resulting in 2 responses at each of 10 target-to-masker ratios (TMRs). Threshold TMR was calculated following Gallun et al. (2013), by subtracting the number of correct responses from 10 dB, resulting in values between 10 dB for no correct responses to -10 dB for all correct responses. Negative TMR thresholds indicate threshold performance (roughly 50% correct) could be achieved when the target was at a lower level than the maskers, while positive thresholds indicate that the maskers needed to be lower in level than the target.

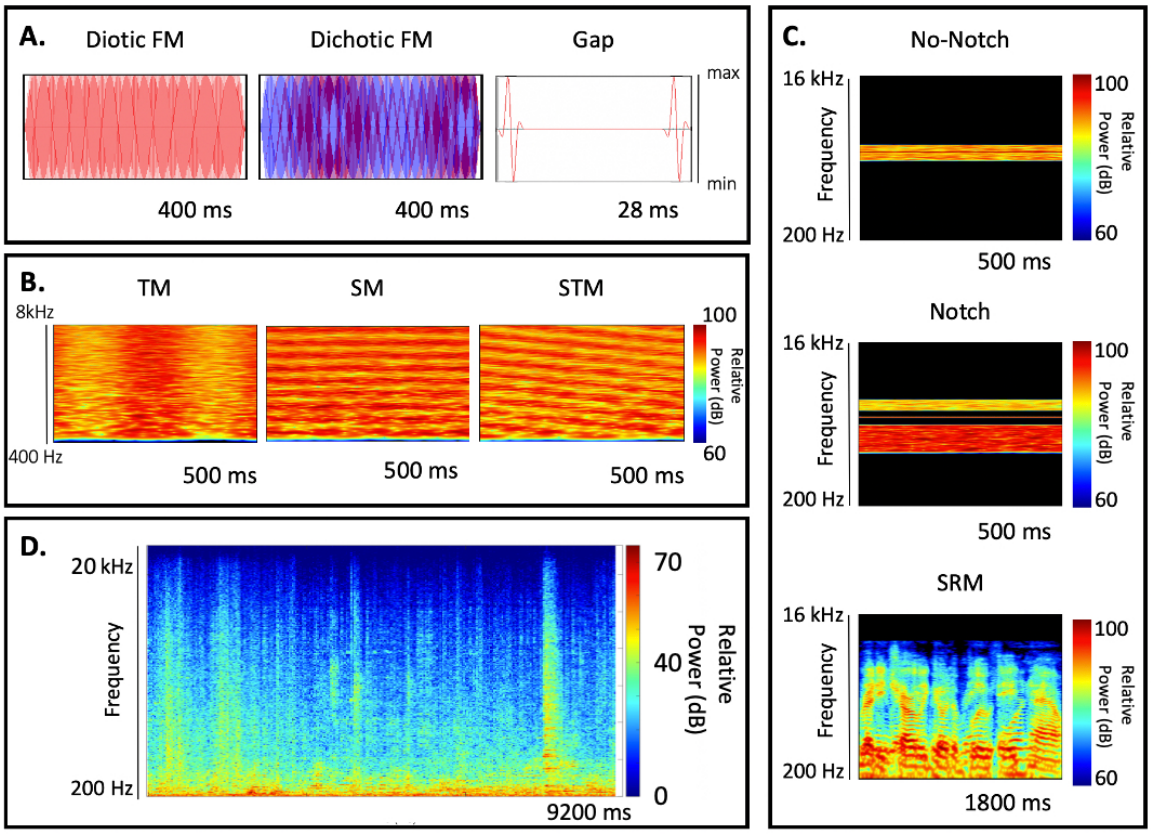

Figure 3. 2: Visual representations of the stimuli employed. Each assessment is grouped by sub-battery as shown in sub-panels A-C. Amplitude envelopes are shown for the TFS sub-battery shown in A and spectrograms are provided for the rest of the assessments shown in B and C. A representative nine-second segment of the cafeteria noise utilized for the Noise condition is shown in sub-panel D. The total recording had a duration of 11 minutes and was played in a continuous loop during testing.

SRM Separated

The stimuli were identical to those in the colocated condition, with the exception that the maskers were presented from 45 degrees to the left and right of the target talker. Responses were again given in the context of a color/number grid (4 colors by 8 numbers) and participants had to select the position indicated by the target signal. Masker level again progressed every other trial from 55 dB SPL to 73 dB SPL in steps of 2 dB as reported in Gallun et al. (2013) and threshold TMR was again estimated by subtracting the number correct from 10 dB. The Spatial Release metric was estimated by subtracting the threshold in the Separated test from the threshold in the Colocated test, resulting in values between -20 dB and 20 dB, with 0 dB indicating no SRM, positive values indicating improvements in performance with spatial separation, and negative values indicating reduced performance with spatial separation.

Experimental Design

The study consisted of 3 different conditions targeted to evaluate the repeatability of PART procedures in a variety of settings. These conditions were run sequentially on three different groups of participants.

Repeatability condition

The first 51 students enrolled were tested with Sennheiser 280 Pro headphones for both sessions and used 2:1 up/down step-size ratio in the staircase.

Headphone condition (in silence)

The next 51 participants enrolled were tested with different headphones (Sennheiser 280 Pro vs Bose Quiet Comfort 35) with the order counter-balanced between participants and used a 1.5:1 up/down step-size ratio.

Noise condition

The next 48 participants enrolled were tested using the same procedure as in the Headphone condition, but with recorded cafeteria noise played at 70 dB SPL. The noise was recorded in a local coffee shop, edited to remove silent gaps between recordings and transient recording noise at the beginning and ends of the recordings, and then bandpass filtered between 20 and 20,000 Hz. The coffee shop noise contained a large number of sound sources at all times, including both speech and environmental sounds. A spectrogram of a representative segment is shown in Fig. 3.2D. Sound files, after processing, were 11 minutes in duration and were played on a loop through two loudspeakers placed 30 cm apart from each other, and positioned in the center of the back of the test room, between 5 and 6 meters behind the three listeners.

RESULTS

Results are divided into sections for the purpose of clarity. First, the results for each test, session, and condition are presented (Section A). Then, issues of test-retest reliability are addressed (Section B) and the consistency of results for each test in comparison to previously reported measures are described (Section C). Finally, the effects of headphones and noise are addressed (Section D). The full data set is provided in the Supplemental Materials for transparency and to encourage re-plotting, comparison with future and past data, and/or re-analysis.

Overview

An overview of the results can be seen in Figure 3.3, which plots data for each test for each participant in each Condition and Session. This figure shows the relationship between estimated thresholds in Session 1 and 2, and substantial overlap of performance between Conditions. This interpretation is consistent with summary statistics for each test shown as a function of Condition and Session in Table 3.1. Due to the high consistency between Conditions, the “main effects” were first analyzed by collapsing the data across Condition (Sections B,C) before addressing the effects of Condition (Section D). By combining data across conditions, a large normative dataset could be constructed, consisting of ~150 participants per test. Consistent with this goal of showing a normative sample,

outlier rejection was performed by removing all data that exceeded three standard deviations from the mean for any condition of any task. The implications of this decision are addressed in the Discussion. Supplementary Materials are provided that demonstrate the robustness of results to different choices of outlier rejection as well as replotting data from Figure 3.3 with outliers included and labeled.

Test-Retest Reliability

Test-retest reliability for the two sessions performed for each assessment in each experimental condition was evaluated using three metrics: Limits of Agreement (LoA, Altman & Bland, 1983; Bland & Altman, 1999), correlation, and t-tests. Each of these measures provides a different, but complementary, perspective on test reliability. The LoA analysis is considered a gold standard analysis as it provides information regarding both agreement and bias (e.g. systematic difference between sessions). Correlations are included to provide a measure of within-subject consistency that ignores systematic effects of session, which can be important for research studies that seek to correct for effects of session. Lastly, t-tests were calculated as a function of session to help determine the reliability of effects of session.

Test	*Repeatability M (SD)*	*Headphone M (SD)*	*Noise M (SD)*	*All Cond. M (SD)*	Units
Gap	2.4 (2.9)	2.17 (3.2)	2.83 (3.41)	2.46 (3.15)	Gap length
S2	1.8 (3.34)	2.08 (3.04)	3.008 (3.07)	2.26 (3.18)	(ms)
Dichotic FM	0.55 (2.09)	0.54 (2.18)	0.45 (2.44)	0.51 (2.23)	Modulation Depth
S2	0.57 (2)	0.47 (2.37)	0.52 (2.77)	0.52 (2.37)	(Hz)
Diotic FM	7.07 (1.57)	6.55 (1.88)	5.48 (1.78)	6.35 (1.75)	Modulation Depth
S2	6.71 (1.65)	5.65 (1.76)	5.82 (1.89)	6.05 (1.77)	(Hz)
TM	1.58 (.77)	1.64 (.82)	1.42 (.85)	1.55 (.81)	Modulation depth
S2	1.58 (.85)	1.46 (.85)	1.21 (.84)	1.42 (.85)	(dB)
SM	1.57 (.61)	1.55 (.75)	1.71 (.82)	1.61 (.72)	Modulation depth
S2	1.33 (.63)	1.37 (.75)	1.64 (.92)	1.44 (.78)	(dB)
STM	0.94 (.17)	.97 (.59)	.93 (.53)	0.95 (.46)	Modulation depth
S2	0.97 (.49)	.98 (.55)	.94 (.62)	0.96 (.55)	(dB)
No-Notch	-11.28 (1.38)	-11.82 (1.92)	-11.15 (2.14)	-11.43 (1.84)	Target-to-masker
S2	-11.74 (1.71)	-12.88 (2.39)	-12.16 (2.48)	-12.26 (2.24)	ratio (dB)
Notch	-31.4 (2.27)	-32.26 (4.52)	-31.29 (3.82)	-31.67 (3.64)	Target-to-masker
S2	-31.91 (3.02)	-32.71 (3.61)	-32.77 (3.31)	-32.44 (3.32)	ratio (dB)
SR Colocated	2.33 (1.36)	2.07 (1.67)	1.92 (2.77)	2.12 (1.96)	Target-to-masker
S2	2.01 (1.51)	1.84 (1.96)	1.34 (2.74)	1.76 (2.08)	ratio (dB)
SR Separated	-4.58 (2.64)	-3.58 (2.93)	-3.57 (4.23)	-3.91 (3.32)	Target-to-masker
S2	-5.36 (2.94)	-5.5 (2.64)	-4.19 (3.86)	-5.04 (3.2)	ratio (dB)
Spatial Release	6.94 (2.78)	5.66 (2.9)	4.57 (3.72)	5.8 (3.24)	Sep - Co
S2	7.34 (3.4)	7.35 (2.71)	4.67 (3.41)	6.58 (3.37)	(dB)

Table 3. 1: Normative summary. Mean thresholds and standard deviations for the 10 assessments utilized plus the derived spatial release metric across all three conditions and their aggregate. Data are presented in PART's native measurement units except for the targets-in-competition tests, which have been converted to TMR. The first row of each test shows session 1 and the second session 2 (S2).

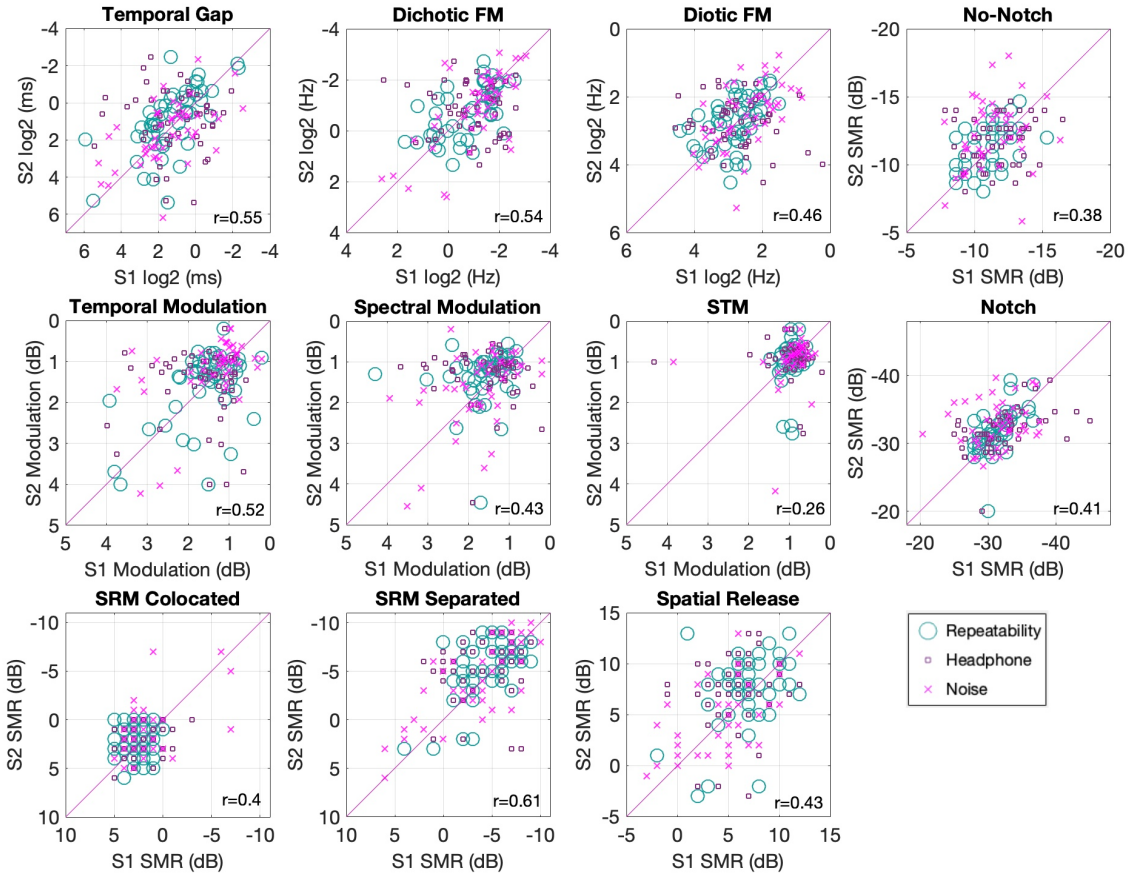

Figure 3. 3: Test re-test correlations. Scatter plots of Session 1 vs Session 2 for the 10 assessments. All axes are oriented to show better performance values away from the origin. Correlations are indicated in the lower right of each panel. Different markers illustrate the different conditions.

Test re-test reliability using limits of agreement (LoA)

For these data, LoA was considered to be a more informative measure of reliability than the normalized correlation, as between-subject variability for a sample that consisted solely of young listeners without hearing problems was anticipated to be small. Correlations, although more common in the auditory literature, are known to depend heavily on between-subject variability and measurement range (Altman & Bland, 1983; Bland & Altman, 1999). LoA plots for the estimated thresholds for each assessment are shown in Fig. 3.4 and the statistics are shown in Table 3.2. This analysis is based on the evaluation of performance across sessions (mean of test and re-test) as a function of their difference. LoA plots can be used to determine the extent to which learning effects are present, which would represent shifts towards better performance across sessions, the region where 95% of the difference between test and re-test is expected to lie, and whether these statistics hold for different levels of performance (homoscedasticity).

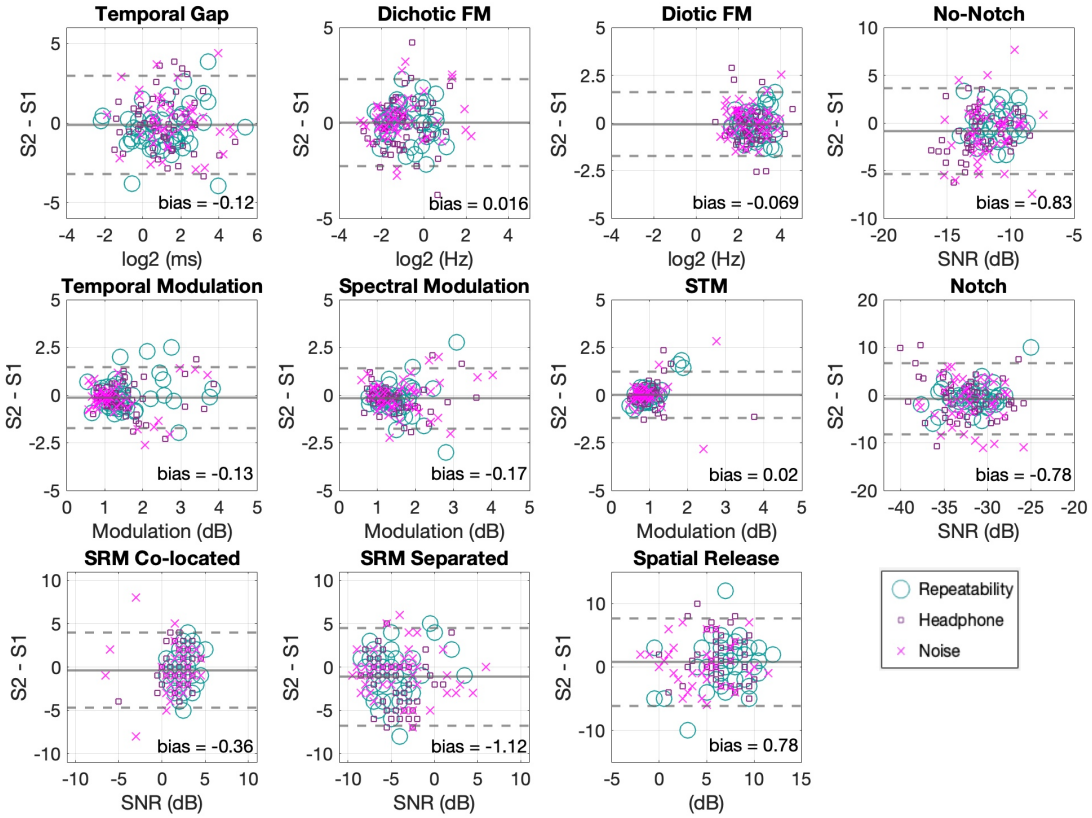

Figure 3. 4: Test re-test limits of agreement. The mean threshold of both sessions is plotted against their difference showing the limits of agreement between sessions for all tests. The solid lines indicate the mean difference between sessions. Dotted lines indicate the 95% limits of agreement. Solid lines that fall below zero indicate better performance on session 2 (except the spatial release metric). Different markers illustrate the different conditions.

Test	*Bias*	*Limits of Agreement*	Units	*r (p)*	*t (p)*	*Cohen's d*	*df*
Gap	-0.12	[-3.2 to 2.96]	log2 (ms)	.55 (<.01)*	0.94 (.34)	-0.07	145
DichoticFM	0.01	[-2.2 to 2.2]	log2 (Hz)	.53 (<.01)*	-0.17 (.86)	0.01	147
DioticFM	-0.06	[-1.73 to 1.59]	log2 (Hz)	.46 (<.01)*	0.97 (.33)	-0.08	144
TM	-0.13	[-1.73 to 1.47]	M (dB)	.52(<.01)*	1.94 (.054)	-0.15	145
SM	-0.16	[-1.75 to 1.41]	M (dB)	.58 (<.01)*	2.46 (.01)*	-0.22	140
STM	0.01	[-1.2 to 1.2]	M (dB)	.26 (<.01)*	-0.29 (.76)	0.03	136
No-Notch	-0.83	[-5.3 to 3.6]	TMR (dB)	.38 (<.01)*	4.37 (<.01) *	-0.4	144
Notch	-0.77	[-8.2 to 6.6]	TMR (dB)	.40 (<.01)*	2.44 (.01)*	-0.22	142
Colocated	-0.36	[-4.7 to 3.9]	TMR (dB)	.39 (<.01)*	1.95 (.052)	-0.17	142
Separated	-1.12	[-6.7 to 4.5]	TMR (dB)	.61 (<.01)*	4.47 (<.01) *	-0.34	147
SpatialR	0.78	[-6.1 to 7.6]	dB	.47 (<.01)*	-2.62 (<.01) *	0.23	140

Table 3. 2: Test re-test statistics. Limits of agreement and within-subject significance testing for the 10 assessments utilized at two time points. Negative values on the bias column indicate better performance on the second session except on the Spatial Release metric, which is the only scale in which larger magnitudes indicate better performance. * indicates significance at α = .05.

In order to facilitate visual inspection and comparisons across different tests, TFS tests were transformed from PART's output units (either Hz or ms) to $\log_2$ units and target-in-competition tests were converted to target-to-masker ratios (subtracting the target level from PART's masker level outputs). The mean across sessions is plotted on the x-axis to give a single point estimate for each participant in terms of their estimated threshold, thus showing between-subject variability of threshold estimation. The difference between sessions is plotted on the y-axis to give a single point estimate of the magnitude of deviation between sessions, thus showing within-subject variability of the estimated threshold. The mean of these differences is plotted as a straight line across the x-axis and its distance from zero (zero = perfect agreement) represents the main point estimate

of the measurement's systematic bias across sessions. The 95% limits of agreement (±1.96 *SD* (difference between sessions)) are plotted as dotted lines and indicate an estimate of the region in which we may expect to observe 95% of the within-subject, between-session differences of threshold estimation.

As can be observed in Fig. 3.4 and Table 3.2, the mean difference between sessions was close to zero in all of the tests, indicating little systematic bias. The 95% limits of agreement for the frequency modulation tests were at modulation rates of approximately ± 2 $\log_2$ (Hz) (or between 0.2 and 4.8 Hz). For the gap detection task, the LoA was approximately ± 3 $\log_2$ (ms) (or between 0.1 and 7.7 ms). The LoA for the modulation detection tasks were approximately ± 3 dB. For the speech tasks, the LoA are target-to-masker ratios of approximately ± 8 dB for the targets in competition tests. Precise values are reported in Table 3.2. As can be seen in Fig. 3.4, the distribution of the threshold estimates had no salient asymmetries and session differences were similar across different levels of performance (symmetry along the abscissa). It is worth noting, however, that more spread can be identified at worse performance levels for some individuals in some tests. This applies both to the full set of data points in each plot and also to the subset of each showing the different conditions. There was little systematic bias between sessions (symmetry along the ordinate) suggesting similar measurement error for both sessions and that the poorer performance cases were expressed without a clear bias towards either session. This analysis

demonstrates the range of alignment to be expected between different threshold estimates within subjects, and indicates that PART produces minimally biased estimates at the group level (see Table 3.2 for relevant statistics).

Correlations Between Sessions

Table 3.2 shows statistics including the strength of association (Pearson r) between sessions. Significant correlations were observed for all the assessments. Overall, the relatively low correlation magnitudes reflect the warning of Altman and Bland (1983) that correlations are less informative to quantify reliability than LoA plots when performance is distributed across relatively narrow ranges of threshold estimates, as was to be expected for young listeners without hearing problems. This is particularly clear in the case of the STM assessment (in the same scale as the SM and TM assessments) where the range of threshold values obtained was quite restricted. In this context, the reduced between-subject variability in relation to a particular within-subject variance will have an impact on r values, decreasing their magnitude.

Repeated-Measures T-tests

To supplement the LoA plots as a test for whether learning, or other factors, gave rise to systematic changes in performance, thresholds were compared between sessions across all three conditions using repeated-measures t-tests (see Table 3.2). While there were statistically significant differences between sessions in the spectral modulation detection test and the tone-in-noise tests, these changes

were quite small, with magnitudes of less than 1 dB. The speech intelligibility test in the separated condition showed a significant difference of greater than 1 dB, which is consistent with the 1.58 dB difference previously reported by Jakien et al. (2017).

Comparison with previously published results

While the above analyses demonstrate reliability of these PART assessments, it is possible that the rapid methods, the presence of noise, or consumer-grade equipment would result in deviations from the results expected based on the published literature. This section thus compares the thresholds reported in Table 3.1 for all conditions averaged across sessions to those previously reported in the literature. This "grand mean" threshold estimate includes thresholds from all of our 150 participants minus rejected outliers and is included in Table 3.3. As outlier rejection was conducted by removing any points that fell more than three standard deviations above or below the mean, the number of outliers rejected and from which conditions is also reported in this section (also in table ST1), so that this can be considered in the comparisons. Overall, threshold estimates align with previous reports within 1.6 *SD* (see Table 3.3) and the number of outliers rejected was roughly consistent with the statistical expectation.

Temporal Fine Structure (TFS)

Sensitivity to temporal processing was assessed with three different tests, temporal gap detection, dichotic FM, and diotic FM. For temporal gap detection, 4 cases were rejected as outliers (Headphone condition 2; Noise condition 2), leaving threshold values that closely resemble those found in the literature (*M* = 2.36 ms , *SD* = 3.16). For example, Schneider et al. (1994) reported thresholds of 3.8 ms (right ear) and 3.5 ms (left ear) on average using 2 kHz tone-bursts similar to the ones we used, however, their stimuli were delivered monaurally. Moreover, Hoover et al. (2019) reported thresholds of 1.45 ms using 0.75 kHz tone-bursts. Gallun et al. (2014) used the most similar stimuli (tone-bursts of 2 kHz) and obtained thresholds of 1.2 ms on average. All three of these studies used monaural presentation of their stimuli. Despite the differences in stimulus frequency and presentation style, all of these estimates lie within half a *SD* from the PART dataset. The fact that the published data report smaller thresholds and the second run appeared to produce smaller thresholds in this study suggest that the differences with the published literature might be removed by providing additional practice in the form of multiple measurements as opposed to the single track on each test session used here.

For the frequency modulated tests (dichotic & diotic FM), thresholds were higher than those previously reported in the literature. For the dichotic FM test, 2 cases were rejected as outliers (Repeatability condition 2). Thresholds in Hz (*M* =

0.52, *SD* = 2.29) are around 1 *SD* higher (on a logarithmic scale) than the 0.2 Hz found by Grose & Mamo (2012), the 0.15 Hz reported by Whiteford & Oxenham (2015), and the 0.19 Hz reported by Hoover et al. (2019).

For diotic FM, 5 cases were rejected as outliers (Repeatability condition 1; Headphone condition 3; Noise condition 1). Thresholds in Hz (*M* = 6.19, *SD* = 1.76), were about 2 *SD* higher than reports of Grose & Mamo (2012) of 1.9 Hz, Whiteford & Oxenham (2015) of .75 Hz, those of Moore & Sek (1996) of 1.12 Hz, and those of Hoover et al (2019) of 1.85 Hz, after conversion to Hz using the method of Witton et al. (2000) where appropriate. These differences in both FM tests are likely due to the difference in stimulus durations employed, which in the literature vary between 1000 ms (Moore & Sek, 1996) and 2000 ms (Whiteford & Oxenham, 2015). Here, the duration was set to 400 ms. This choice was based on the results of Palandrani et al. (2019), who showed that FM detection thresholds improve with the square-root of stimulus duration. This predicts diotic FM thresholds of 3.6 Hz for the listeners here, if durations that were comparable to Grose & Mamo (2012) had been used. Even after this correction, however, the thresholds obtained were about 1 *SD* worse on average (on a logarithmic scale) after conversion to Hz using the method of Witton et al. (2000) where appropriate. As with the temporal gap assessment, it would not be surprising if repeated testing resulted in reduced thresholds, more similar to those reported in the literature.

Spectro-Temporal Modulation (STM

Sensitivity to spectro-temporal modulation was assessed with three different tests, spectro-temporal modulation (STM), spectral modulation (SM) and temporal modulation (TM). It is difficult to make exact comparisons with previously reported results in the literature without making a variety of transformations and ignoring several differences in methodology. The most important issue is the measurement of modulation depth. Measurement depends on the scale (log or linear), the reference points (peak-to-valley or peak-to-midpoint), and the order in which the modulation operations are performed among other factors (see Isarangura et al., 2019). In this case, PART generated stimuli that were modulated on a logarithmic amplitude scale (dB), with modulation depth measured from the middle of the amplitude range to the peak amplitude. This differs from the method used by others, such as Hoover et al. (2018), who measured applied modulation that was sinusoidal on a dB scale but measured the amplitude as the difference from the maximum to the minimum, rather than the midpoint. Still more different was Bernstein et al. (2013), who applied sinusoidal modulation on a linear scale and also measured the modulation depth from the maximum to the minimum. When the amplitude scale is linear, threshold is expressed by transforming the modulation depth (m), which varies between 0 and 1, into dB units using the value 20 times log(m), which means that a fully modulated signal has a value of 0 dB, and a modulation depth

of 0.01 has a value of -40. These differ from the values used to express threshold using a log amplitude scale, and thus Equation 1 (below) was used to convert the thresholds obtained with PART to 20log(m) dB units as detailed in Isarangura et al. (2019).

$$20Log_{10}\left(\frac{10^{\left(\frac{m}{10}\right)}-1}{10^{\left(\frac{m}{10}\right)}+1}\right) \qquad \text{(Equation 1)}$$

For STM at 4 Hz and 2 c/o, 13 cases were rejected as outliers (Repeatability condition 2; Headphone condition 3; Noise condition 8). STM thresholds obtained (*M* = 0.95 (M) dB, *SD* = 0.51) were converted using equation 1 to (*M* = -19.28 20log(m) dB, *SD* = 4.67) and closely resembled those previously reported in the literature. They were within a *SD* from those reported by Gallun et al. (2018) for five different testing sites (range -21.74 to -18.42 dB) and for Chi et al. (1999) (-22 dB). The obtained thresholds for STM are about 2 *SD* better than those reported by Bernstein et al. (2013) (-14 dB). It is unclear why outlier rejection was higher in this task than in others, but this may indicate that this is an ability on which some listeners are particularly poor. It is worth noting that the Supplemental Materials, where these statistics are reported without outlier rejection, still shows better performance than in the literature.

For SM at 2 c/o, 9 cases were rejected as outliers (Repeatability condition 1; Headphone condition 4; Noise condition 4). SM modulation depth thresholds (*M* = 1.52 dB, *SD* = 0.75) were converted using equation 1 to (*M* = -15.34 20log(m) dB, *SD* = 4.23) were better by about 1 *SD* than those reported by

Hoover, Eddins & Eddins (2018) (-11.08 dB), and those reported by Davies-Venn, Nelson & Souza (2015) (about -11 dB). These differences might be due to differences in modulation depth generation patterns or modulation depth metrics employed (see Isarangura et al., 2019). Further, stimulus parameters like those of the noise carrier bandwidth or presentation level, and test parameters such as tracking procedure varied across studies and so might account for the slight differences found. One reason to suspect that these methodological differences influenced performance is the fact that our listeners often outperformed the more practiced listeners in the other studies. Again, a higher number of outliers were observed, but as can be seen in the Supplemental Material, this did not account for the better performance in this study.

For TM at 4 Hz, 4 cases were rejected as outliers (Repeatability condition 1; Noise condition 3). TM thresholds (*M* = 1.49 (M) dB, *SD* = 0.83), converted using equation 1 to (*M* = -15.99 20log(m) dB, *SD* = 4.34), were within half a *SD* of those reported by Viemeister (1979) of -18.5 dB for four observers.

Target Identification in Competition

Tone Detection in Noise with and without a Spectral Notch

These tests evaluated the ability to detect a 2 kHz pure tone in competition with broad-band noise either overlaying the target signal (no-notch condition) or with an 800 Hz spectral notch or protective region without noise (notch condition). We rejected 5 cases as outliers (Headphone condition 1; Noise condition 4) from the

No-Notch test and obtained thresholds of *M* = -11.85, *SD* = 2. In the case of the Notch test, we rejected 7 cases (Headphone condition 1; Noise condition 6) and obtained thresholds of *M* = -32.06, *SD* = 3.5. The notched-noise procedure has been widely used for the analysis of frequency selectivity in the cochlea (see Moore, 2012). Because of this, the emphasis of the literature has been on calculating detailed information about the shape of the auditory filter, and specific thresholds associated with each condition are typically not reported. In one of the few examples where thresholds are described directly, Patterson (1976) reported an average distance between the equivalent of our no-notch and notch conditions of about 24 dB for four participants. This is comparable to the mean distance we obtained here of 20.2 dB (SD = 2.9), and some of our participants did indeed produce thresholds similar to those of the four well-practiced listeners described in Patterson (1976). That some of our listeners were able to obtain the same thresholds as the well-trained listeners described in Patterson (1976) suggests that this may be a task for which training plays a fairly small role. This conclusion is supported by the finding, shown in Table 3.2, that the the thresholds in the first run were on average less than 1 dB higher than the thresholds obtained on the second run.

Speech-on-speech Competition

These tests evaluated the discrimination of speech in the face of speech competition using variants of the Spatial Release from Masking (SRM) test

described by Gallun et al. (2013). Two conditions were used, one where the speech-based competition was colocated in virtual space with the target speech (colocated) and one where the speech-based competition was located ±45 degrees away from the target (separated) in simulated space. All values are reported in target-to-masker ratio (TMR) dB units. In the case of the colocated condition, 7 cases were rejected as outliers from the Noise condition. Interestingly, these cases are mainly due to performance that was better than average by more than 3 *SD* (see Supplemental Materials Fig S1), which has been observed previously for the occasional younger listener with normal hearing. Colocated thresholds (*M* = 1.94 dB, *SD* = 2.03) closely resemble those reported by Gallun et al. (2018) across two testing sites (1.85 & 1.96 dB). Performance was slightly worse than predicted by the normative functions of Jakien & Gallun (2018), which are based on linear regression to the data from a variety of listeners varying in age and hearing loss. For a 20 year old with a pure-tone average (PTA) of 5 dB HL, which seems appropriate for this sample, colocated thresholds averaged across two runs are predicted to be 1.2 dB, which is within 1 *SD* of what is observed.

In the Separated condition, 2 cases were rejected as outliers (Repeatability condition 1; Noise condition 1). Separated thresholds (*M* = -4.47 dB, *SD* = 3.31) closely resembled those reported by Gallun et al. (2018) across two testing sites (-4.33 & -4.62 dB), all of which were higher on average than the

predictions of the equation of Jakien & Gallun (2018), which predicts a threshold of -6.7 dB, which is still within 1 *SD* of those observed.

The difference between the separated and the colocated conditions, a metric indicating the spatial release from masking effects, showed spatial release values (*M* = 6.19 dB, *SD* = 3.32) that again closely resembled the ones reported by Gallun et al. (2018) across two testing sites (6.19 & 6.57 dB) and were within 1 *SD* of the predictions of the regression equation of Jakien & Gallun (2018), which predicts 8.3 dB. Of note, the spatial release from masking magnitudes reported here are smaller than those reported for other similar tests already used in the clinic like the LiSN-S (Cameron & Dillon, 2007) which do not use synchronized concurrent masking and thus allow for better ear glimpsing (Brungart & Iyer, 2012) to contribute to the effect of release from masking (Glyde et al., 2013).

Assessment	*Grand Average M (SD)*	*Closest laboratory test*	*Distance in SD*
Gap	2.36 (3.16) ms	Gallun et al., 2014	–
Dichotic FM	0.52 (2.3) Hz	Grose & Mamo, 2012; Hoover et al., 2019	–
Diotic FM	6.2 (1.76) Hz	Grose & Mamo, 2012; Hoover et al., 2019	– –
TM	1.49 (.83) M (dB)	Viemeister, 1979	–
SM	1.52 (.75) M (dB)	Hoover, Eddins & Eddins, 2005	+ +
STM	0.95 (.51) M (dB)	Gallun et al., 2018	- / +
No-Notch	-11.85 (2.09) TMR (dB)	Patterson, 1976	– –
Notch	-32.06 (3.5) TMR (dB)	Patterson, 1976	– –
SR Colocated	1.94 (2.03) TMR (dB)	Jakien et al., 2017; Gallun et al., 2018	- / +
SR Separated	-4.47 (3.31) TMR (dB)	Jakien et al., 2017; Gallun et al., 2018	- / +
Spatial Release	6.19 (3.32) (dB)	Jakien et al., 2017; Gallun et al., 2018	- / +

Table 3. 3: Comparison to known laboratory measures. Summary of the similarities of the grand average thresholds estimated in the present study using PART and matched psychophysical tests from previous research. Plus or minus signs indicate values that are better or worse than previous reports respectively. The number of signs indicates increases in terms of *SD*s, one sign indicates < 1 *SD* and two signs indicate between 1 & 2 *SD.* Cases with both a plus and a minus sign indicate that different conditions or experiments reported previously are < 1 *SD* above and below the threshold estimates in this study.

The Effects of Headphones and Noise

To address the effects of headphone types with and without noise-attenuation technology and external noise conditions, “main effects” were evaluated by collapsing across tasks. To do so, composite scores were computed by normalizing each individual assessment relative to its mean and standard deviation (a z-score transform), calculating a z-score for each listener in each

assessment, and then averaging these normalized values across the 10 assessments for each participant. Z-scores and composite scores are included in the data set that is available as part of the Supplemental Materials.

LoA plots, Pearson correlations, and t-tests were then computed for the composite scores. Results are reported for each condition separately, divided by headphone type. To test differences across experimental manipulations, a mixed-model Analysis of Variance (ANOVA) was used to compare composite scores across the factors of interest. The data for each condition as a function of test type are available in the Supplemental Materials.

The internal reliability of the composite score was assessed by calculating Cronbach's α, which gave a score of 0.75. This indicates that the composite has strong internal reliability and thus it can be appropriate to use it as a summary score. Of note, the composite score is not an attempt to reduce central auditory processing to a single construct. Rather, this measure is intended to address the effects of these experimental manipulations in an efficient manner across all the assessments in the battery.

Fig. 3.5 shows the 95% limits of agreement for the standardized composite scores of the whole sample across three experimental conditions (panel on the left). This analysis showed close to zero bias (< 0.01), and limits of agreement of [-0.87, 0.88] which indicate that 95% of repeated estimates of the composite score of the battery used in this study are expected to lie within 1 SD

from each other in young listeners without hearing problems. In addition, Fig. 3.5 also shows a scatterplot of session 1 vs 2 for the composite scores (panel on the right). This composite showed stronger association between sessions than each of the individual assessments (r = 0.65 p < .001, [95% CI = 0.55, 0.73]) and represents an alternative estimate to the LoA regarding the reliability of the battery as a whole and not of its individual assessments.

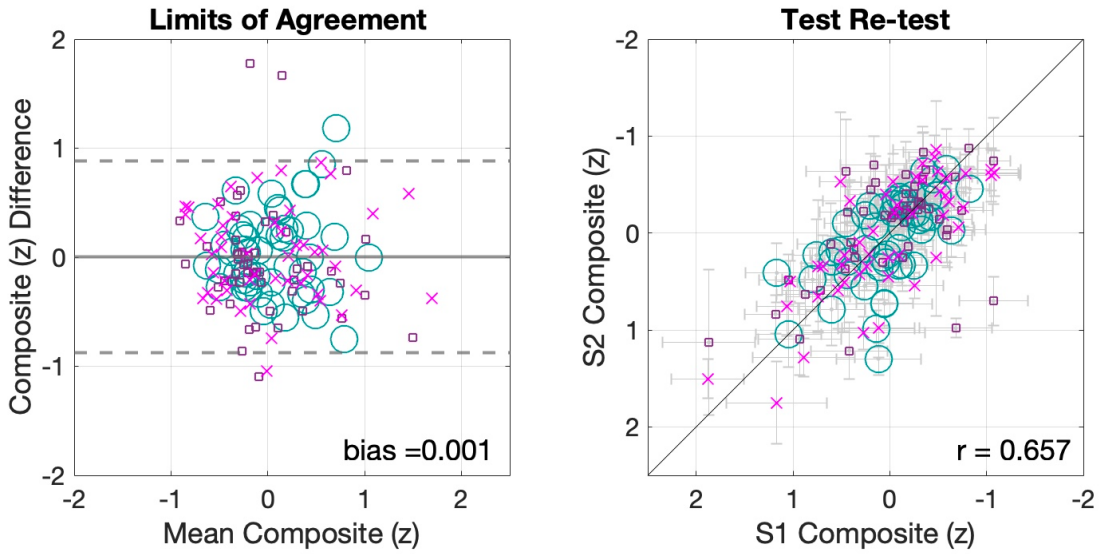

Figure 3. 5: All composite scores. Calculated within-subject and compared across all three conditions using different markers. Panel on the left shows the limits of agreement (see Altman & Bland, 1983). Panel on the right shows scatterplot with its correlation. The horizontal error bars indicate SEM for session 1 while the vertical bars reflect session 2 SEM.

Threshold Differences across Conditions

To address how composite scores changed as a function of listening condition, the composite score is plotted separately for each condition (Fig. 3.6). In all three experimental conditions, composite scores showed minimal bias (Repeatability condition = -0.01; Headphone condition = -0.05; Noise condition = 0.07), limits of agreement that resemble the aggregate sample's composite around 1 *SD* (Repeatability condition [-0.72, 0.79]; Headphone condition [-0.98, 0.81]; Noise condition [-0.8, 0.86]), and similar strength of association between scores of session 1 and 2 with (r = .601, p < .001) for Repeatability condition (standard); (r = .639, p < .001) for Headphone condition (silence); and (r = .774 p < .001) for Noise condition (noise). These correlations are within the 95% confidence intervals of the general aggregate composite r value.

To formally test the hypothesis of zero bias in threshold estimation between sessions for the different listening conditions, a series of t-tests was conducted. These tests failed to find significant differences in any of the conditions (Repeatability condition, $t_{(50)}$ = -0.34, p = .73, *Cohen's d* = -0.04; Headphone condition, $t_{(50)}$ = 0.3, p = .76, *Cohen's d* = 0.04); Noise condition, $t_{(47)}$ = -0.17, p = .86, *Cohen's d* = -0.02). Finally, as an additional test of significance, a 3 x 2 repeated measures ANOVA with the within-subject factor Session and the between-subjects factor Condition was conducted to assess the overall effects of repeated measurements across a range of conditions. Again, no statistically

significant effects were found for either Session ($F_{(1,147)}$= 0.004, p = .94 η^2 < 0.01) nor for Condition ($F_{(2,147)}$= 0.92, p = .4, η^2 = *0*.01), and with no significant interaction ($F_{(2,147)}$= 0.11, p = .88, η^2 < 0.01).

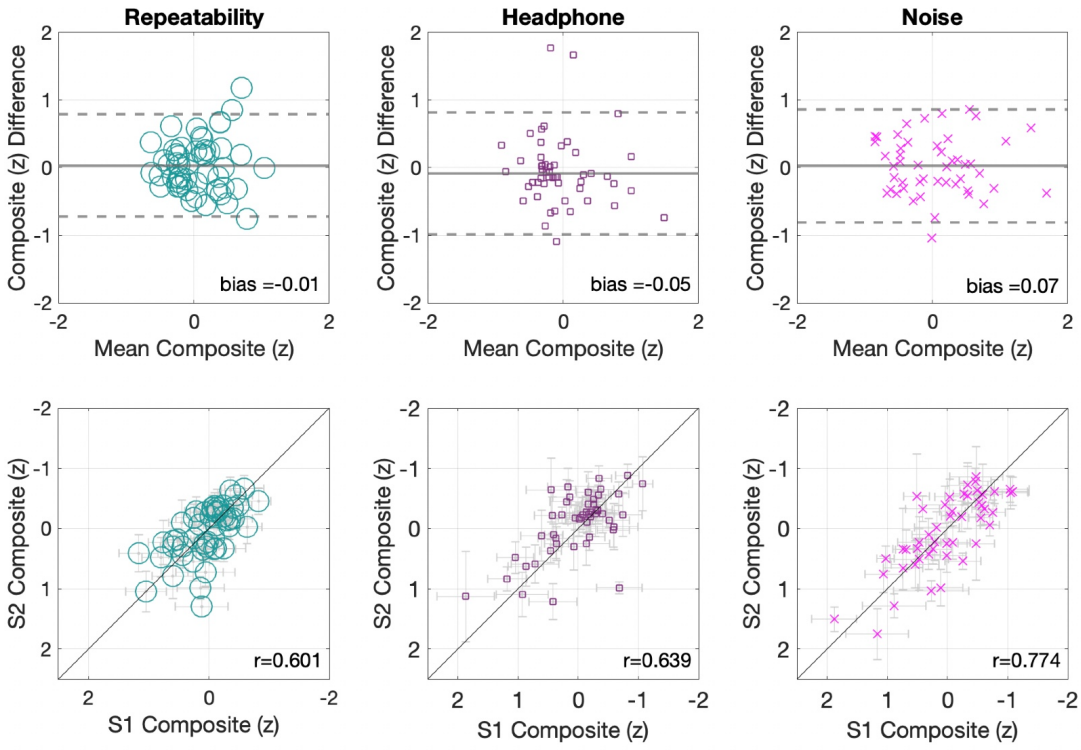

Figure 3. 6: Composite scores. Separated by condition. Top panels show the limits of agreement plots. Bottom panels show the composite scatterplots for each condition. The horizontal error bars indicate SEM for session 1 while the vertical bars reflect session 2 SEM.

Headphone Comparison

To examine the effects of headphone type and the presence of environmental noise, data are presented from the Headphone condition (both headphone types in silence) and the Noise condition (two headphone types in noise). Fig. 3.7 shows the limits of agreement between sessions as well as the scatter plots that show their association for the Headphone and Noise conditions separately. The data are plotted for each assessment in Supplemental Figures S2 and S3. The agreement analysis between the estimated thresholds using either set of headphones again showed minimally biased estimates (Headphone condition (silence) 0.003; Noise condition (noise) 0.082) and similar limits of agreement near 1 *SD* (Headphone condition (silence) [-0.94, 0.94]; Noise condition (noise) [-0.67, 0.84]) as reported in the general aggregate. Composite correlations were also similar to what is reported above, with the correlation for Headphone condition (silence) (r = .509 p < .001) suffering due to a reduced between-subject variability. A stronger association between measures was found in Noise condition (noise) where performance between-subjects is increased in relation to the within-subject variation (r = .74, p < .001).

The stability of threshold estimates across headphones in different environmental noise conditions is a notable result, as not only were the headphones different, but also, they shared the same output from the iPad, which was calibrated according to the Sennheiser 280 Pro—and not the Bose—

headphone's mechanical output levels as detailed in the Methods section. After calibration, an output level of 80 dB SPL (using the Sennheiser 280 Pro as recorded with a Brüel & Kjær Head and Torso Simulator with Artificial Ears in a VA RR&D NCRAR anechoic chamber) resulted in a level of 66 dB SPL for the Bose Quiet Comfort 35, with the high noise-cancelling setting engaged as used in all testing sessions (73 dB SPL with the noise-cancelling setting turned off). In order to allow the headphone effects to be examined without modification, and to avoid recalibration of the iPad between test sessions in the experiment, the settings that produced an 80 dB SPL output for the Sennheiser were used also for the Bose headphones. This meant that even in a silent environment, all of the stimuli were attenuated by 14 dB when Bose headphones were used.

Table 3.4 shows the mean thresholds and SDs for each type of headphone in each condition and assessment. Table 3.5 shows the within-subject LoA, correlations between headphones used, and repeated measures t-tests that formally test differences between the estimated thresholds with each headphone for each condition and assessment separately. The data associated with these statistical tests are plotted in Supplemental Figures S2 and S3. To test for differences in threshold estimation as a function of headphone type, t-tests were used to compare between headphone types in each condition. Of note, since headphone type was counterbalanced across sessions, these analyses were averaged across sessions. No statistically significant effects were observed

in either condition (Headphone condition (silence), $t_{(50)}$ = -0.03, p = .97, *Cohen's d* = -0.005; Noise condition (noise) $t_{(47)}$ = 1.45, p = .15, *Cohen's d* = 0.21).

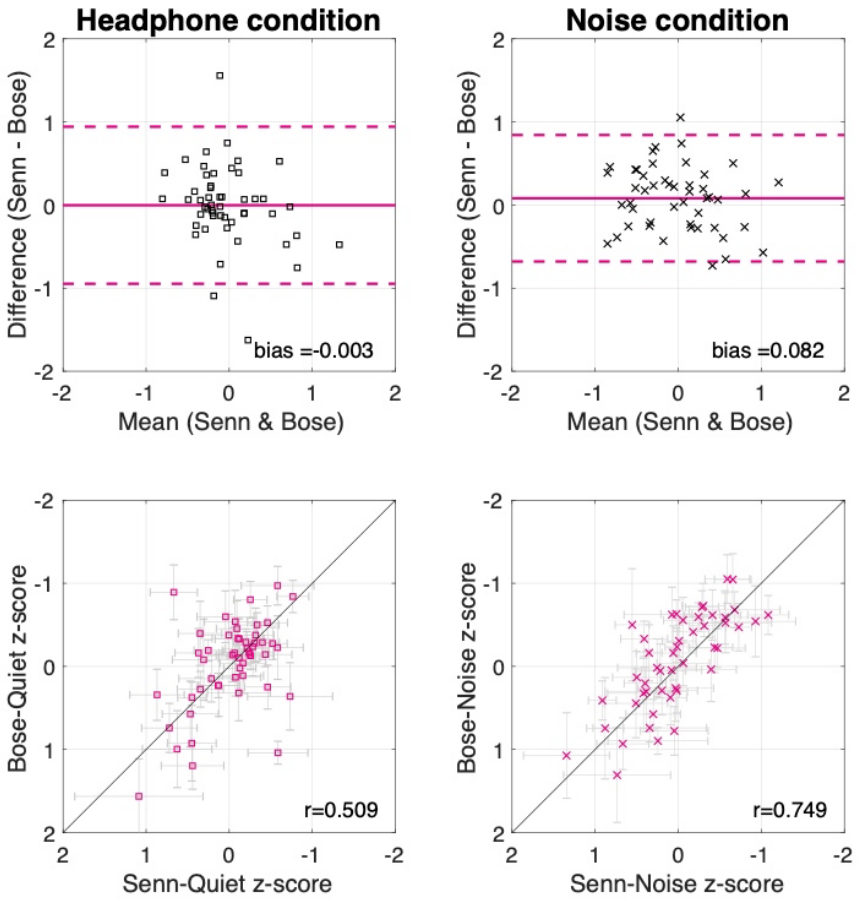

Figure 3. 7: Headphone composite. Top panels show the limits of agreement plots. Bottom panels show the composite scatterplots relating headphone type used. The horizontal error bars indicate performance with Sennheiser 280 pro headphones and the vertical errorbars indicate performance with the Bose Quiet Comfort 35 headphones.

As an additional test of significance, a 2 x 2 repeated measures ANOVA with the within-subject factor Headphone and the between-subjects factor Condition was conducted to assess headphone effects across experimental conditions. Again, no statistically significant effects were found for either Headphone ($F_{(1,97)}$= 0.8, p = .37 η^2< 0.01) nor for Condition ($F_{(1,97)}$= 0.01, p = .91, η^2< 0.01), and with no significant interaction ($F_{(1,97)}$= 0.9, p = .34, η^2< 0.01).

In summary, the data failed to show any systematic effect of headphone type when participants were tested in either silent or noisy environments. These composite analyses further support the reliability of PART and suggest that it may be achieved with or without active noise cancelling technology and in presence of moderate environmental noise. These results also suggest that even a 14 dB difference at the mechanical output level did not produce noticeable differences in performances for these undergraduate students with hearing in the normal range.

Test	*Headphone* *M (SD)*	*Noise* *M (SD)*	Units
Gap Senn	2.38 (2.92)	3.18 (3.15)	Gap length (ms)
Bose	1.89 (3.29)	2.67 (3.31)	
Dichotic FM Senn	0.49 (1.96)	0.48 (2.74)	Modulation Depth (Hz)
Bose	0.52 (2.59)	0.49 (2.48)	
Diotic FM Senn	5.91 (1.71)	5.88 (1.85)	Modulation Depth (Hz)
Bose	6.26 (1.94)	5.43 (1.81)	
TM Senn	1.54 (.75)	1.32 (.8)	Modulation depth (dB)
Bose	1.57 (.91)	1.3 (.9)	
SM Senn	1.49 (.72)	1.68 (.88)	Modulation depth (dB)
Bose	1.43 (.79)	1.66 (.86)	
STM Senn	0.94 (.51)	0.94 (.59)	Modulation depth (dB)
Bose	1.009 (.62)	0.93 (.56)	
No-Notch Senn	-12.25 (2.19)	-11.66 (2.1)	Signal-to-masker ratio (dB)
Bose	-12.46 (2.27)	-11.66 (2.61)	
Notch Senn	-32.33 (3.19)	-31.93 (2.56)	Signal-to-masker ratio (dB)
Bose	-32.64 (4.83)	-32.13 (4.48)	
SR Co-located Senn	1.84 (2.22)	1.87 (2.46)	Signal-to-masker ratio (dB)
Bose	2.07 (1.3)	1.39 (3.04)	
SR Separated Senn	-4.82 (3.25)	-3.51 (4.04)	Signal-to-masker ratio (dB)
Bose	-4.27 (2.59)	-4.25 (4.05)	
Spatial Release Senn	6.66 (3.05)	4.47 (3.54)	SR (Sep - Co) (dB)
Bose	6.35 (2.8)	4.77 (3.59)	

Table 3. 4: Headphone summary. Mean thresholds and standard deviations for the 10 assessments utilized plus the derived spatial release metric across both conditions that used different headphones. Data is presented in PART's native measurement units except for the targets-in-competition tests that have been converted to TMR. The first row of each test shows thresholds obtained with the Sennheiser 280 Pro system and the second with the Bose Quiet Comfort 35 system.

Test	*Bias*	*Limits of Agreement*	Units	*r (p)*	*t (p)*	*Cohen's d*	*df*
Gap	-0.33	[-3.63 to 2.97]	Log2	.47 (<.01)*	1.37 (.17)	-.2	48
(Noise)	-.25	[-3.36 to 1.67]	(ms)	.56 (<.01)*	1.07 (.28)	-.14	45
DichoticFM	0.07	[-2.62 to 2.77]	Log2	.35 (.01)*	-0.37 (.7)	.06	50
(Noise)	0.02	[-2.21 to 2.25]	(Hz)	.66 (<.01)*	-0.12 (.9)	.01	47
DioticFM	0.08	[-1.88 to 2.05]	Log2	.34 (.01)*	-0.58 (.56)	.96	47
(Noise)	-0.11	[-1.72 to 1.49]	(Hz)	.56 (<.01)*	0.94 (.34)	-.12	46
TM	0.03	[-1.58 to 1.64]	M	.53(<.01)*	-0.28 (.77)	.03	50
(Noise)	-0.01	[-1.7 to 1.67]	(dB)	.49 (<.01)*	0.13 (.89)	-.02	44
SM	-0.06	[-1.55 to 1.43]	M	.5 (<.01)*	0.55 (.58)	-.08	46
(Noise)	-0.01	[-1.78 to 1.75]	(dB)	.47 (<.01)*	0.11 (.9)	-.01	43
STM	0.06	[-1.19 to 1.32]	M	.38 (<.01)*	-0.71 (.47)	.11	47
(Noise)	-0.004	[-1.45 to 1.45]	(dB)	.19 (.23)	0.03 (.97))	-.007	39
No-Notch	-0.2	[-5.02, 4.6]	SMR	.39 (<.01)*	0.59 (.55)	-.09	49
(Noise)	0	[-6.1 to 6.1]	(dB)	.14 (.34)	0 (1)	0	43
Notch	-0.31	[-8.72 to 8.09]	SMR	.49 (<.01)*	0.51 (.6)	-.07	49
(Noise)	-0.2	[-9.52 to 9.11]	(dB)	.17 (.25)	0.27 (.78)	-.05	41
Co-located	0.23	[-3.77 to 4.24]	SMR	.42 (<.01)*	-0.82 (.41)	.12	50
(Noise)	-.48	[-5.89 to 4.91]]	(dB)	.51 (<.01)*	1.13 (.26)	-.17	40
Separated	0.54	[-6.31 to 7.4]	SMR	.29 (.03)*	-1.12 (.26)	.18	50
(Noise)	-.74	[-6.32 to 4.83]	(dB)	.75 (<.01)*	1.79 (.07)	-.18	46
SpatialR	-0.31	[-8 to 7.37]	SMR	.107 (.45)	0.57 (.57)	-.1	50
(Noise)	0.3	[-6.28 to 6.88]	(dB)	.55 (<.01)*	-0.56 (.57)	.08	39

Table 3. 5: Headphone statistics. Limits of agreement and significance testing for the 10 assessments comparing headphones used in two conditions. The first row shows the Headphone condition and the second the Noise condition. Positive values on the bias column indicate better performance with the Sennheiser system except on the Spatial Release metric which is the only scale in which larger magnitudes indicate better performance. * indicate significance at $\alpha = .05$.

DISCUSSION

This study examined the validity and reliability of a battery of ten assessments that evaluate different aspects of central auditory function using the Portable Automatic Rapid Testing (PART) application applied to young adult listeners without reported hearing problems. Overall, results show that thresholds can be obtained that are highly consistent across sessions and that are very similar to those reported in laboratory settings obtained with more traditional equipment and, in some cases, with extended testing and training (see Table 3.3). Furthermore, results from the Repeatability condition were replicated in the Headphone and Noise conditions, demonstrating that PART produces consistent threshold estimates across a variety of settings and equipment. Overall these results suggest that the PART platform can provide valid measurements across a range of listening conditions.

An important utility for this study is that it provides an initial normative data set for young adult listeners for the PART tasks reported, and eventually as a reference for patient populations. However, substantial work is required before PART will be appropriate for clinical use. For example, while the tasks included in this first battery were chosen based upon prior literature suggesting possible sensitivity in understanding listening disorders, these data do not capture variations in age and do not include effects of differences in hearing threshold

(see Jakien and Gallun, 2018). Future work will involve developing a similar normative data set for this battery. With such a dataset it would be possible to determine which combinations of tests are most sensitive in distinguishing between different disorders. Likewise, the reliability of measures needs to be established in different populations that may have more difficulty with the procedures than the college student population reported here.

Further, although learning effects are among the smallest effects observed in this data set, they must be explored in relevant patient populations and potentially accounted for when interpreting test results. In both cases, our results suggest that either repeated measures or adjustments to adaptive procedures will be required to increase reliability in patient populations. Still, the fact that thresholds similar to those found in the literature can be obtained on a large number of tests within a short period of time, using consumer-grade technology, provides optimism that PART will be useful in the clinic.

The criteria we used for outlier rejection was justified by the goal of creating a normative dataset, but it is important to note that the field holds a variety of different views regarding outlier rejection. The current choice is simply definitional, in that ‘normative’ refers to a normal distribution, and thus, it is appropriate to reject data falling far outside of this distribution. Nevertheless, there was a minimal effect on population estimates of means and standard deviations, whether or not outliers are excluded (see Supplemental Table ST1).

Supplemental Figure S1, which shows the data with outliers circled, reveals that the main impact of including outliers is to make it more difficult to see the normal range of the dataset. Another important question is whether or not it is possible to say something meaningful about which listeners gave data that was then rejected. For example, are they impaired in auditory processing, or do they represent typical variation of the larger population? It seems unlikely that these participants had any significant hearing loss as they all self-reported to have no hearing difficulties, which is considered a reasonable indication for normal hearing (Vermiglio, Soli & Fang, 2017), and were able to detect a 45 dB SPL 2 kHz pure tone which assured an audibility minimum criteria. Moreover, as indicated in Supplemental Table ST1, most outlying cases were not consistent across sessions, and as can be seen in Supplemental Figure S1 are within the normal range in one of the testing sessions. This suggests that many of the outliers were either inattentive, unmotivated, confused or otherwise non-compliant in one of the sessions. Further, some of the outliers were actually in the supra-normal hearing range, again evidence against outliers being indicative of hearing impairment. Still, while it is reasonable to suggest that outliers do not represent normative or dominantly systematic effects, they are still a concern and do need to be considered when contemplating clinical implications, especially of a single test. In particular, it is important to keep in mind the expected probability

(ranging from 1-8% depending on the test) of getting an unreliable test result when using these tests on this platform.

Another issue of concern is the extent to which performance may change systematically across testing sessions. LoA plots, correlational analyses, and statistical tests of session effects all provide complementary information on changes in performance over time. In general, systematic bias of threshold estimates across sessions was minimal, and similar to the bias observed across levels of performance (see LoA plots). The expected differences among measures across sessions are estimated in the limits of agreement between sessions. While significance testing revealed some differences between sessions in some of the tests, the effect sizes we found are small and typically comparable to the smallest step sizes used in the procedures. Further, these can now be considered as test re-test effects in future work. Consistent with this, reliability was further quantified using Pearson *r*, which is not sensitive to systematic effects of testing session. Although correlation is highly reactive to small changes in between-subject variability (see Supplementary Table ST1), and cases with reduced between-subject variability, it presents complementary information regarding the relation among within-subject and between-subject variabilities. When the assessments with smaller *r* values in complement to the LoA plots are examined, it can be seen that in most cases the limits of agreement closely resemble other tests with higher *r* values. This is an indication that *r* is

decreasing due to restricted ranges of good performance as was to be expected in this sample of normal listeners.

It is notable that thresholds were consistent across different external noise conditions (Repeatability & Headphone vs Noise conditions) and across different types of headphones (Repeatability vs Headphone & Noise conditions) (see Figs. 3.6 & 3.7). Also of note, the correlations were higher for the condition with external cafeteria noise without an increase on the limits of agreement. This is important because a test platform that is portable, automatic, and rapid can only be successfully exploited if it is able to provide accurate measurements that can be collected in a variety of potentially less than optimal settings. Here, we have shown that PART was able to obtain estimates of central auditory function that resemble those found under laboratory conditions, despite using untrained listeners tested in settings including that resembling a typical, moderately noisy, university cafeteria. PART should be considered as a supplemental tool in the clinic which can be used to collect valuable information about a person's hearing capabilities with little need for supervision from a clinician. These results also suggest that this system and these tests are robust to the presence of moderate noise and substantial variability in sound output levels.

This study also compared the use of headphones with an active noise-cancelling technology to those with passive attenuation. We considered it worthwhile to test this technology because it is now widely available, but little is

known about the advantages and disadvantages it could represent for auditory testing. We failed to find a statistically significant effect between threshold estimates obtained for the different headphones under both silent and noisy listening conditions. In other words, estimated thresholds were similar for the Sennheiser 280 Pro in silence (Headphone condition), and this lack of difference manifested similarly in noisy conditions (Noise condition). This suggests that the passive attenuation provided by the Sennheiser 280 Pro is sufficient to obtain reliable measurements in less than optimal external noise conditions outside of the sound booth. Also, it suggests that the differences between the headphones, including the active noise-cancelling algorithm, are not changing the signal in any way that results in significant reductions in performance. Perhaps the noise-cancelling signal processing was inactive or operating at low frequencies that did not affect performance. In any case, threshold estimation held constant across the headphone technologies used with a single calibration profile (the same output from the iPad). These data serve as verification that relatively inexpensive auditory hardware can be used to test auditory function in a variety of settings with sufficient precision to provide clinical evidence of central auditory function in individual listeners.

PART can thus appropriately be considered as a valid platform for testing several aspects of central auditory processing. It is robust to moderate levels of ambient noise and small variants in equipment and procedure. The reported data

can now be used as a normative baseline against which auditory dysfunction can be identified in future work. However, clinical research will be needed to determine how thresholds vary as a function of age and different degrees of hearing loss. The reliability analysis reported here applies only to young listeners with normal hearing. Future work will need to address whether threshold estimates from PART can be reliably obtained for older listeners with varying degrees of hearing loss, and to determine the extent to which the measured reliability in this work is adequate for identifying central auditory processing deficit. This next step is feasible considering that the PART platform is highly accessible given its relatively low cost in terms of expense (it only requires a computer tablet and headphones), time (the whole battery of 10 assessments takes less than 1 hr), human resources (it runs the assessments automatically, one after another, including instructions and breaks), and that it can be used in range of environmental settings suitable for testing (from the anechoic chamber as in Gallun et al. (2018) to noisy cafeteria conditions). Thus, PART has the potential to provide a supplementary tool to gather the quantity and variety of psychophysical measures of auditory function that will allow us to translate laboratory findings into the clinic to inform clinical practice.

REFERENCES

ANSI/ASA. (**2004**). *Methods for manual pure-tone threshold audiometry (ANSI S3.21-2004).* (American National Standards Institute, New York).

Arnst, D. J. (**1981**). "Errors on the Staggered Spondaic Word (SSW) Test in a Group of Adult Normal Listeners," Ear and Hearing, **2**(3), pp. 112–116.

Altman, D. G., & Bland, J. M. (**1983**). "Measurement in Medicine: The Analysis of Method Comparison Studies," The Statistician, **32**(3), pp. 307.

Bergman, M., Najenson, T., Korn, C., Harel, N., Erenthal, P., & Sachartov, E. (**1992**). "Frequency selectivity as a potential measure of noise damage susceptibility," British Journal of Audiology, **26**(1), pp. 15–22.

Bernstein, J. G., Mehraei, G., Shamma, S., Gallun, F. J., Theodoroff, S. M. and Leek, M. R. (**2013**). "Spectro-temporal modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners," J Am Acad Audiol **24**(4), pp. 293-306.

Bland, J. M., & Altman, D. G. (**1999**). "Measuring agreement in method comparison studies," Statistical Methods in Medical Research, **8**(2), pp. 135–160.

Bolia, R. S., Nelson, W. T., Ericson, M. A. and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," Journal of the Acoustical Society of America **107**(2), pp. 1065-1066.

Broadbent, D.E. (**1958**), *Perception and communication.* (Pergamon, New York).

Bronkhorst, A. W. (**2015**). "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," Attention, Perception, & Psychophysics, **77**(5), pp. 1465–1487.

Brungart, D. S. and Iyer, N. (**2012**). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers," J Acoust Soc Am **132**(4), pp. 2545-2556.

Bunch, C. C. (**1929**). "Age Variations in Auditory Acuity," Archives of Otolaryngology - Head and Neck Surgery, **9**(6), pp. 625–636.

Cameron, S. and Dillon, H. (**2007**). “Development of the Listening in Spatialized Noise-Sentences Test (LISN-S),” Ear Hear **28**(2), pp. 196-211.

Carhart, R., and Jerger, J. F. (**1959**). “Preferred Method For Clinical Determination Of Pure-Tone Thresholds,” Journal of Speech and Hearing Disorders, **24**(4), pp. 330–345.

CHABA. (**1988**). “Speech understanding and aging,” J. Acoust. Soc. Am. 83, pp. 859–895.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., Shamma, S., Chi, T., … Shamma, S. (**1999**). “Spectro-temporal modulation transfer functions and speech intelligibility,” The Journal of the Acoustical Society of America **106**, pp. 2719–2732.

Davies-Venn, E., Nelson, P., & Souza, P. (**2015**). “Comparing auditory filter bandwidths, spectral ripple modulation detection, spectral ripple discrimination, and speech recognition: Normal and impaired hearing,” The Journal of the Acoustical Society of America, **138**(1), pp. 492–503.

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (**2001**). “Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex,” Journal of Neurophysiology, **85**(3), pp. 1220–1234.

Eckert, M. A., Matthews, L. J., & Dubno, J. R. (**2017**). “Self-Assessed Hearing Handicap in Older Adults With Poorer-Than-Predicted Speech Recognition in Noise,” Journal of Speech, Language, and Hearing Research, **60**(1), pp. 252-261.

Eddins, D. A. and Hall, J. W. (**2010**). “Binaural processing and auditory asymmetries,” In *The Aging Auditory System* (Springer, 135-165).

Fifer, R. C., Jerger, J. F., Berlin, C. I., Tobey, E. A., & Campbell, J. C. (**1983**). “Development of a dichotic sentence identification test for hearing-impaired adults,” Ear and Hearing, **4**(6), pp. 300–305.

Füllgrabe, C., Moore, B. C. and Stone, M. A. (**2015**). “Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition,” Frontiers in Aging Neuroscience, **6**(347) pp. 1–25.

Gallun, F. and Souza, P. (**2008**). “Exploring the role of the modulation spectrum in phoneme recognition,” Ear Hear, **29**(5), pp. 800–813.

Gallun, F.J., Diedesch, A.C., Beasley, R. (**2012**). “Impacts of age on memory for auditory intensity,” Journal of the Acoustical Society of America, **132**(2), pp. 944–956.

Gallun, F. J., Diedesch, A. C., Kampel, S. D., & Jakien, K. M. (**2013**). “Independent impacts of age and hearing loss on spatial release in a complex auditory environment,” Frontiers in Neuroscience, 7, pp. 1–11.

Gallun, F. J., McMillan, G. P., Molis, M. R., Kampel, S. D., Dann, S. M. and Konrad-Martin, D. L. (**2014**). “Relating age and hearing loss to monaural, bilateral, and binaural temporal sensitivity,” Front Neurosci **8**(172), pp. 1–14.

Gallun, F. J., Seitz, A., Eddins, D. A., Molis, M. R., Stavropoulos, T., Jakien, K. M., … Srinivasan, N. (**2018**). “Development and validation of Portable Automated Rapid Testing (PART) measures for auditory research,” **33**(175). Paper 2pPPb15.

Gallun, F. J., Best, V. (***In Press***). “Age-Related Changes in Spatial Hearing, Segregation, and Realistic Listening Situations,” In *The Aging Auditory System* (Springer Handbook of Auditory Research).

García-Pérez, M. A. (**1998**). “Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties,” Vision research, **38**(12), pp. 1861-1881.

García-Pérez, M. A. (**2001**). “Yes-no staircases with fixed step sizes: psychometric properties and optimal setup,” Optometry and Vision Science, **78**(1), pp. 56-64.

García-Pérez, M. A. (**2011**). “A cautionary note on the use of the adaptive up–down method,” The Journal of the Acoustical Society of America, **130**(4), pp. 2098-2107.

Glyde, H., Buchholz, J. M., Dillon, H., Cameron, S., Hickson, L. (**2013**). “The effect of better-ear glimpsing on spatial release from masking,” J Acoust Soc Am **134**(4), pp. 2937-2945.

Green, D. M. (**1971**). “Temporal auditory acuity,” Psychological Review, **78**(6), pp. 540–551.

Green, D.M. (**1976**). *An Introduction to Hearing*. (Wiley, New York).

Grose, J. H., and Mamo, S. K. (**2010**). “Processing of temporal fine structure as a function of age,” Ear and Hearing, **31**, pp. 755-760.

Hirsh, I. J. (**1948**). “The influence of interaural phase on interaural summation and inhibition,” Journal of the Acoustical Society of America, **20**, pp. 536–544.

Hoover, E. C., Gallun, F. J., & Eddins, D. A. (**2019**). “Challenging standard practices in adaptive psychophysics,” The Journal of the Acoustical Society of America, **145**(3), pp. 1758–1758.

Hoover, E. C., Pasquesi, L., & Souza, P. (**2015**). “Comparison of Clinical and Traditional Gap Detection Tests,” Journal of the American Academy of Audiology, **26**(6), pp. 540–546.

Hoover, E. C., Souza, P. E., & Gallun, F. J. (**2017**). “Auditory and cognitive factors associated with speech-in-noise complaints following mild traumatic brain injury,” Journal of the American Academy of Audiology, **28**(4), pp. 325–339.

Hoover, E. C., Eddins, A. C., & Eddins, D. A. (**2018**). “Distribution of spectral modulation transfer functions in a young, normal-hearing population,” The Journal of the Acoustical Society of America, **143**(1), pp. 306–309.

Hoover, E. C., Kinney, B. N., Bell, K. L., Gallun, F. J., & Eddins, D. A. (**2019**). “A Comparison of Behavioral Methods for Indexing the Auditory Processing of Temporal Fine Structure Cues,” Journal of Speech, Language, and Hearing Research : JSLHR, **62**(6), pp. 2018–2034.

Hughson, W., & Westlake, H. (**1944**). “Manual for program outline for rehabilitation of aural casualties both military and civilian,” Transactions of the American Academy of Ophthalmology and Otolaryngology, **48**(Suppl), pp. 1–15.

Iliadou, V., Chermak, G. D., Bamiou, D.-E., Rawool, V. W., Ptok, M., Purdy, S., … Musiek, F. E. (**2018**). “Letter to the Editor: An Affront to Scientific

Inquiry," Response to: Moore, D. R. (2018) Editorial: Auditory Processing Disorder. Ear and Hearing, **39**(6), pp. 1236–1242.

Isarangura, S., Palandrani, K., Stavropoulos, T., Seitz, A., Hoover, E. C., Gallun, F. J., and Eddins, D. A. (**2019**). "The effects of modulator shape and methods for expressing modulation depth on spectral modulation detection thresholds," The Journal of the Acoustical Society of America, **145**(3), pp. 1722-1722.

ISO. (**2010**). *Acoustics–audiometric test methods–part 1: Basic pure tone air and bone conduction threshold audiometry (ISO 8253-1:2010)*. (International Organization for Standardization, Geneva, CH).

Jakien, K.M. and Gallun, F.J. (**2018**) "Normative data for a rapid, automated test of spatial release from masking," American Journal of Audiology, **27**, pp. 529–538.

Jakien, K. M., Kampel, S. D., Stansell, M. M., and Gallun, F. J. (**2017**). "Validating a rapid, automated test of spatial release from masking," American Journal of Audiology, **26**(4), pp. 507–518.

Katz, J. (**1962**). "The use of staggered spondaic words for assessing the integrity of the central nervous system," The Journal of Auditory Research, **2**, pp. 327–337.

Keith, R. W. (**1995**). "Development and standardization of SCAN-A: test of auditory processing disorders in adolescents and adults," Journal of the American Academy of Audiology, **6**(4), pp. 286–292.

Kowalski, N., Depireux, D. A., & Shamma, S. A. (**1996**). "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," Journal of Neurophysiology, **76**(5), pp. 3503–3523.

Levitt, H. (**1971**). "Transformed Up-Down Methods in Psychoacoustics," The Journal of the Acoustical Society of America, **49**(2B), pp. 467–477.

Marrone, N., Mason, C.R., Kidd, G. (**2008**) "The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms," The Journal of the Acoustical Society of America **124**, pp. 3064–3075.

Mehraei, G., Gallun, F. J., Leek, M. R. and Bernstein, J. G. (**2014**). "Spectrotemporal modulation sensitivity for hearing-impaired listeners: Dependence on carrier center frequency and the relationship to speech intelligibility," J Acoust Soc Am, **136**(1), pp. 301–316.

Moore, B. C. J. (**1987**). "Distribution of auditory-filter bandwidths at 2 kHz in young normal listeners," Journal of the Acoustical Society of America, **81**(5), pp. 1633–1635.

Moore, B. C. J., & Glasberg, B. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hearing Research, **47**, pp. 103–138.

Moore, B. C., M. A. Stone, C. Füllgrabe, B. R. Glasberg and S. Puria (**2008**). "Spectro-temporal characteristics of speech at high frequencies, and the potential for restoration of audibility to people with mild-to-moderate hearing loss," Ear and Hearing **29**(6), pp. 907-922.

Moore, B. C. J. (**2012**)*. An introduction to the psychology of hearing.* (Brill, Leiden, the Netherlands).

Moore, D. R., Edmondson-Jones, M., Dawes, P., Fortnum, H., McCormack, A., Pierzycki, R. H., & Munro, K. J. (**2014**). "Relation between speech-in-noise threshold, hearing loss and cognition from 40-69 years of age," *PLoS ONE*, ***9***(9), pp. 1–10.

Moore, D. R. (**2018**). "Guest Editorial," Ear and Hearing, **39**(4), pp. 617–620.

Musiek, F.E. (**1983**). "Assessment of central auditory dysfunction: the dichotic digit test revisited," Ear Hear, **4**(2), pp. 79-83.

Musiek, F.E., and Pinheiro, M.L. (**1987**). Frequency Patterns in Cochlear, Brainstem, and Cerebral Lesions. Audiology, 26(2), pp. 79-88.

Musiek, F. E. (**1994**). "Frequency (pitch) and duration pattern tests," Journal of the American Academy of Audiology, **5**(4), pp. 265–268.

Olsen, W. O., Noffsinger, D., & Carhart, R. (**1976**). "Masking Level Differences Encountered in Clinical Populations," Audiology, **15**(4), pp. 287–301.

Palandrani et al. (**2019**). "The effects of duration on monaural and binaural temporal fine structure coding," Assoc. Res. Otolaryngol. Abs.:318.

Patterson, R. D. (**1976**). “Auditory filter shapes derived with noise stimuli,” The Journal of the Acoustical Society of America, **59**(3), pp. 640–654.

Pfeiffer, R. R., and Kim, D. O. (**2005**). “Cochlear nerve fiber responses: Distribution along the cochlear partition,” The Journal of the Acoustical Society of America, **60**(4), pp. 966–966.

Plomp, R. (**1964**). “Rate of Decay of Auditory Sensation,” The Journal of the Acoustical Society of America, **36**(2), pp. 277–282.

Schimmel, O., Van de Par, S., Breebaart, J. and Kohlrausch, A. (**2008**). “Sound segregation based on temporal envelope structure and binaural cues,” Journal of the Acoustical Society of America, **124**(2), pp. 1130–1145.

Schonwiesner, M. and Zatorre, R. J. (**2009**). “Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI,” Proc Natl Acad Sci U S A, **106**(34), pp. 14611-14616.

Shamma, S. (**2001**). “On the role of space and time in auditory processing,” Trends Cogn Sci **5**(8), pp. 340–348.

Shinn-Cunningham, B. G. (**2008**). “Object-based auditory and visual attention,” Trends Cogn Sci **12**(5), pp. 182–186.

Snell, K. B., Mapes, F. M., Hickman, E. D. and Frisina, D. R. (**2002**). “Word recognition in competing babble and the effects of age, temporal processing, and absolute sensitivity,” Journal of the Acoustical Society of America **112**(2), pp. 720-727.

Souza, P., Hoover, E., Blackburn, M., & Gallun, F. (**2018**). “The Characteristics of Adults with Severe Hearing Loss,” Journal of the American Academy of Audiology, **29**(8), pp. 764–779.

Souza, P., Gallun, F.J., Wright, R. (**2020**) “Predicting speech-cue weighting in older adults with impaired hearing,” Journal of Speech Language and Hearing Research, **63**(1), pp. 334–344.

Stavropoulos, T.A, Isarangura, S., Hoover, E.C, Eddins, D.A, Seitz, A.R., Gallun F.J. (**2021**) “Exponential spectro-temporal modulation,” *The Journal of the Acoustical Society of America, 149(3), 1434–1443. https://doi.org/10.1121/10.0003604*

Stecker, G. C., and Gallun, F. J. (**2012**). “Binaural Hearing, Sound Localization, and Spatial Hearing,” In: Tremblay, K.L., Burkard, R.F. (Eds.), *Translational Perspectives in Auditory Neuroscience: Normal Aspects of Hearing* (Plural Publishing, San Diego, pp. 383–433).

Stone, M. A., Glasberg, B. R., & Moore, B. C. J. (**1992**). “Technical note: Simplified measurement of auditory filter shapes using the notched-noise method,” British Journal of Audiology, **26**(6), pp. 329–334.

Theunissen, F. E., Sen, K. and Doupe, A. J. (**2000**). “Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds,” The Journal of Neuroscience **20**(6), pp. 2315-2331.

Theunissen, F. E., & Elie, J. E. (**2014**). “Neural processing of natural sounds,” Nature Reviews Neuroscience, **15**(6), pp. 355–366.

Tremblay, K., Piskosz, M., & Souza, P.E. (**2003**). “Effects of age and age-related hearing loss on the neural representation of speech cues,” Clinical Neurophysiology, **114**, pp. 1332-1343.

van Veen, T. M., and Houtgast, T. (**1985**). “Spectral sharpness and vowel dissimilarity,” J. Acoust. Soc. Am. **77**, pp. 628–634.

Venezia, J. H., Martin, A. G., Hickok, G., & Richards, V. M. (**2019**). “Identification of the spectrotemporal modulations that support speech intelligibility in hearing-impaired and normal-hearing listeners,” Journal of Speech, Language, and Hearing Research, **62**(4), pp. 1051–1067.

Vermiglio, A. J., Soli, S. D., and Fang, X. (**2018**). “An argument for self-report as a reference standard in audiology,” Journal of the American Academy of Audiology, **29**(3), pp. 206–222.

Versnel, H., Zwiers, M. P., and van Opstal, A. J. (**2009**). “Spectrotemporal Response Properties of Inferior Colliculus Neurons in Alert Monkey,” Journal of Neuroscience, **29**(31), pp. 9725–9739.

Viemeister, N. F. (**1979**). “Temporal modulation transfer functions based upon modulation thresholds,” Journal of the Acoustical Society of America, **66**(5), pp. 1364–1380.

von Békésy, G. (**2005**). “Hearing Theories and Complex Sounds,” The Journal of the Acoustical Society of America, **35**(4), pp. 588–601.

Whitefield, I. C., & Evans, F. (**1965**). “Responses of Auditory Cortical Neurons to Stimuli of changing Frequency,” J. Neuroph, **28**, pp. 655–672.

Whiteford, K. L., & Oxenham, A. J. (**2015**). Using individual differences to test the role of temporal and place cues in coding frequency modulation. The Journal of the Acoustical Society of America, **138**(5), pp. 3093–3104.

Whiteford, K. L., Kreft, H. A., & Oxenham, A. J. (**2017**). “Assessing the Role of Place and Timing Cues in Coding Frequency and Amplitude Modulation as a Function of Age,” JARO - Journal of the Association for Research in Otolaryngology, **18**(4), pp. 619–633.

Witton, C., Green, G. G. R., Rees, A., & Henning, G. B. (**2000**). “Monaural and binaural detection of sinusoidal phase modulation of a 500-Hz tone,” The Journal of the Acoustical Society of America, **108**(4), pp. 1826–1833.

Winer, J. A. (**2006**). “Decoding the auditory corticofugal systems,” Hearing Research, **212**(1–2), pp. 1–8.

Winter I.M. (**2005**). “The Neurophysiology of Pitch,” In Plack C.J., Fay R.R., Oxenham A.J., Popper A.N. (eds) *Pitch.* Springer Handbook of Auditory Research, vol 24. (Springer, New York).

APPENDIX I: Chapter 3 Supplemental Materials

Outlier Analysis

Table ST1 is provided to demonstrate that outlier rejection, which was largely for ease of presentation has a limited impact on estimates of normative ranges for the assessments (compare columns 4 to 5 of table ST1). Figure S1 plots the data from all participants for all tests, with outliers indicated as circled symbols. Table ST1 also provides measures of how many outliers were observed in each condition of each task and the degree to which a given participant provided consistent thresholds, using a metric of a score greater than 2 standard deviations beyond the mean on both sessions, as shown in columns 2 and 3 in Table ST1. This reveals that for those participants for whom data were excluded, thresholds were typically within the normal range in one of the sessions. This is also visible in Figure S1.

Assessment	Outliers (by condition)	Consistent cases <|2SD| ΔSession	M (SD) Full-data	M (SD) No-outliers	Units
Gap	4 (0, 2, 2)	0	2.49 (2.9)	2.36 (2.75)	ms
DichoticFM	2 (2, 0, 0)	2 (2, 0, 0)	0.53 (2.1)	0.52 (2)	Hz
DioticFM	5 (1, 3, 1)	2 (0, 2, 0)	6.3 (1.7)	6.1 (1.6)	Hz
TM	4 (1, 0, 3)	2 (0, 0, 2)	1.59 (.93)	1.49 (.73)	(M) dB
SM	9 (1, 4, 4)	4 (0, 1, 3)	1.71 (.99)	1.52 (.63)	(M) dB
STM	13 (2, 3, 8)	4 (0, 1, 3)	1.18 (.86)	0.95 (.4)	(M) dB
No-Notch	5 (1, 0, 4)	1 (1, 0, 0)	-11.6 (2)	-11.8 (1.7)	SMR dB
Notch	7 (0, 1, 6)	0	-30.9 (6.1)	-32 (2.9)	SMR dB
SR Co-located	7 (0, 0, 7)	5 (0, 0, 5)	1.51 (2.5)	1.94 (1.6)	SMR dB
SR Separated	2 (1, 0, 1)	0	-4.34 (3.1)	-4.47 (2.9)	SMR dB
Spatial Release	9 (1, 0, 8)	5 (0, 0, 5)	5.86 (3)	6.19 (2.79)	dB

Table ST 1: Outlier analysis. Cases on each assessment, consistency across sessions, and impact on mean thresholds and standard deviations for the 10 assessments and the spatial release metric.

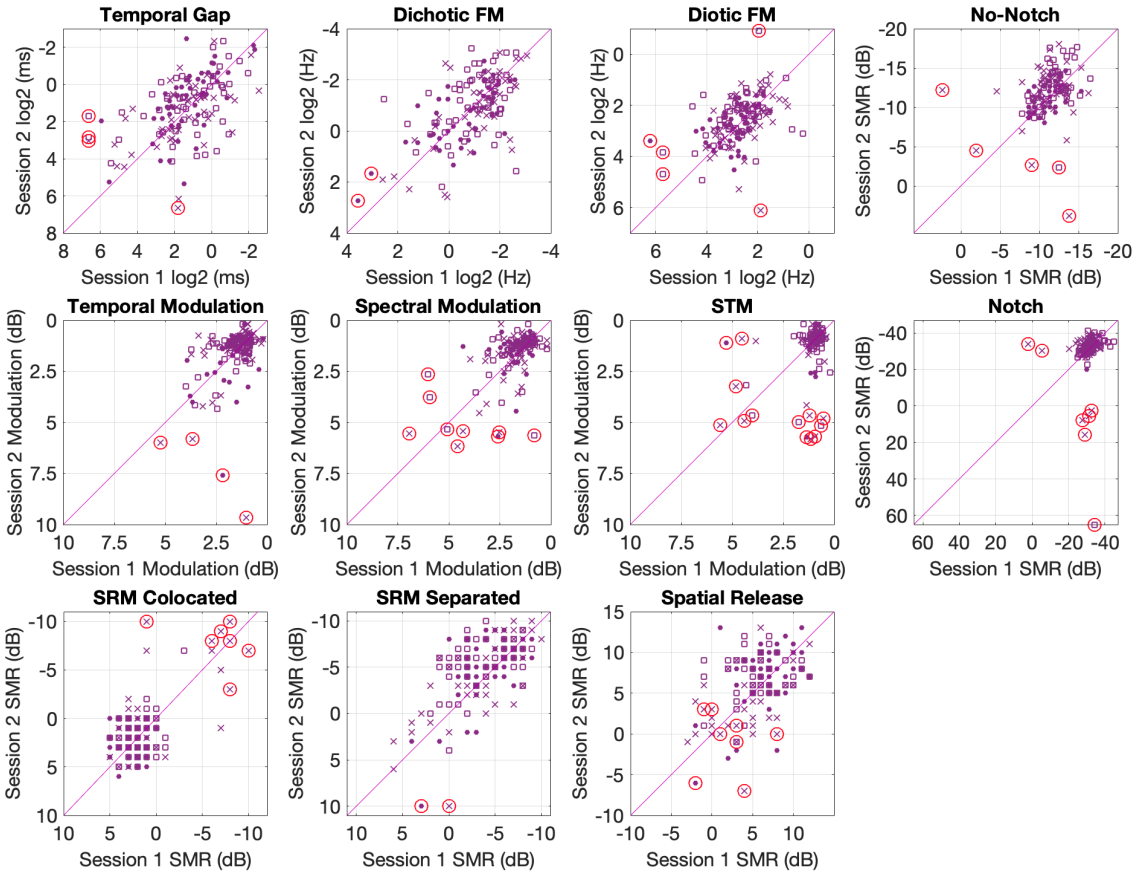

Figure S 1: Outliers. Scatter plots of Session 1 vs Session 2 for the 10 assessments used for all three conditions. Filled circles represent the Repeatability condition, open squares represent the Headphone condition, and crosses represent the Noise condition. Cases flagged as outliers (±3 *SD*) and removed from main analysis are marked with a surrounding circles. All axes are oriented to show better performance values away from the origin. The diagonal is plotted to ease evaluation of differences between sessions. Dots above this line indicate better performance in session 2.

Headphone effects

Figures S2 (Limits of Agreement) and S3 (scatterplots) show the within-subject effects of headphone type.

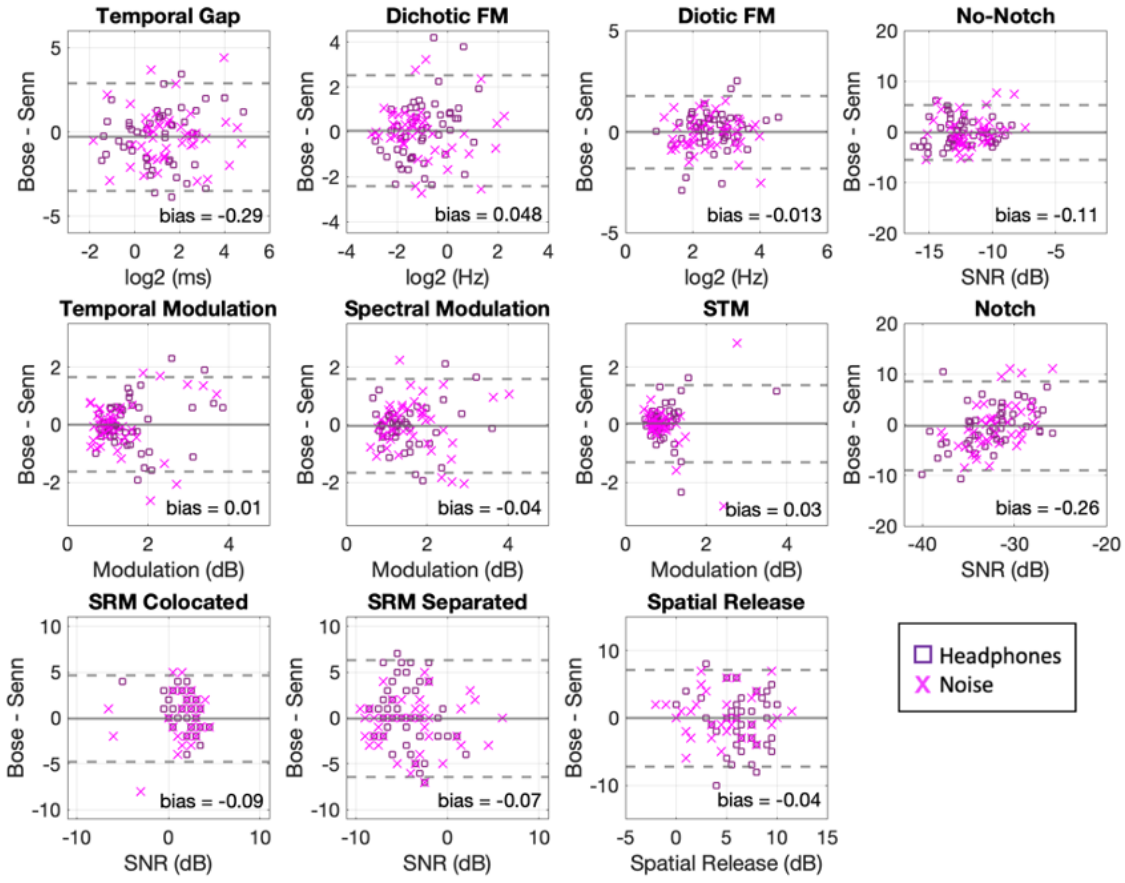

Figure S 2: Limits of agreement plots. Estimated thresholds across two sessions using different headphone types in conditions 2 (squares) and 3 (crosses). The solid lines indicate the mean difference between headphone type. Dotted lines indicate the 95% limits of agreement. The red circle indicates the mean threshold for each test centered at zero difference between headphones. Solid lines below zero indicate better performance on the Bose headphones with active noise attenuation for all the plots except the spatial release metric, for which higher values indicate better performance.

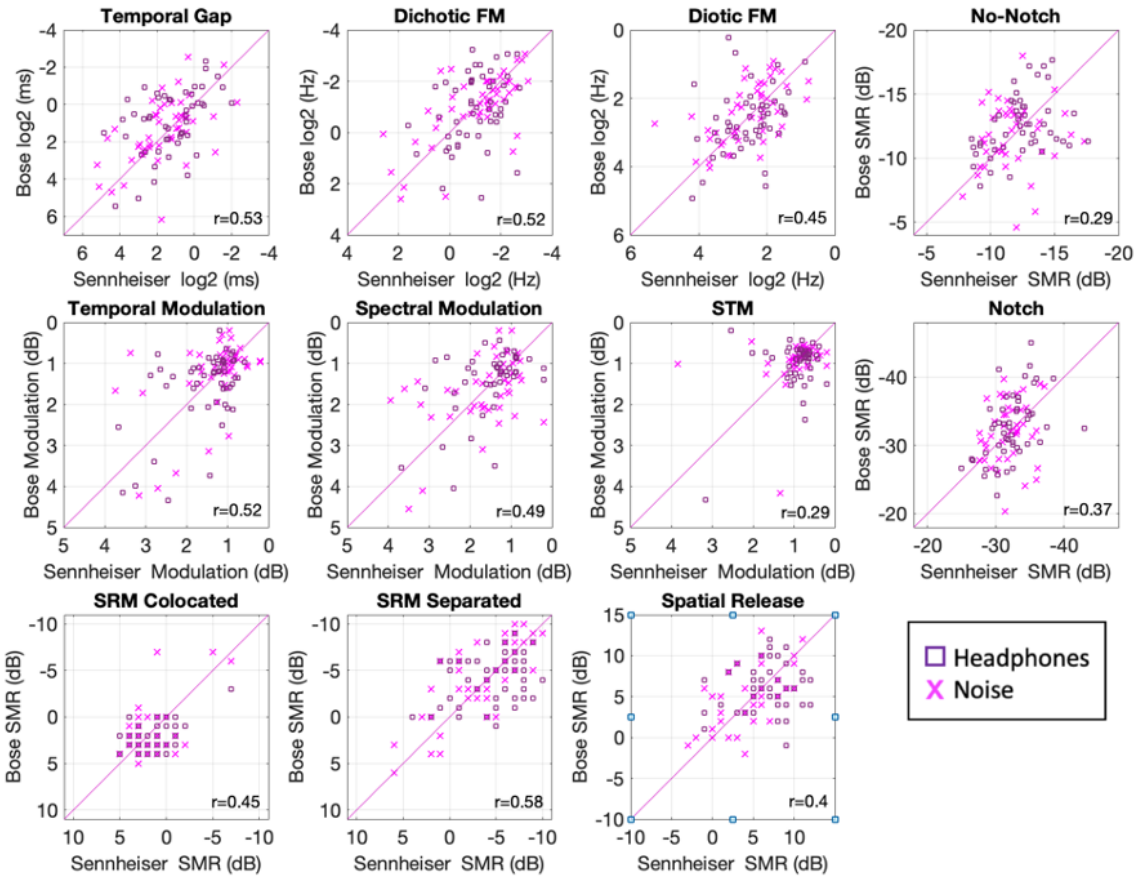

Figure S 3: Scatter plots relating headphone types. For the 10 assessments in conditions 2 (squares) and 3 (crosses). All axes are oriented to show better performance values away from the origin. Correlations are indicated in the lower right of each panel. The diagonal is plotted to ease evaluation of differences between headphone types. Dots above this line indicate better performance with active noise attenuation.

Effects of Staircase Parameters

Here we examine how variance of parameters of the staircase procedure impacted performance. The ratio of step-sizes up:down in the adaptive staircases was 2:1 for the Repeatability Condition and 1.5:1 for the other two (Headphone & Noise). This manipulation did not manifest salient differences in terms of the estimated thresholds between conditions, however it did have an impact on the number of trials required to achieve a threshold estimate. This is shown in Figure S4, in which the mean number of trials are plotted per task per experiment.

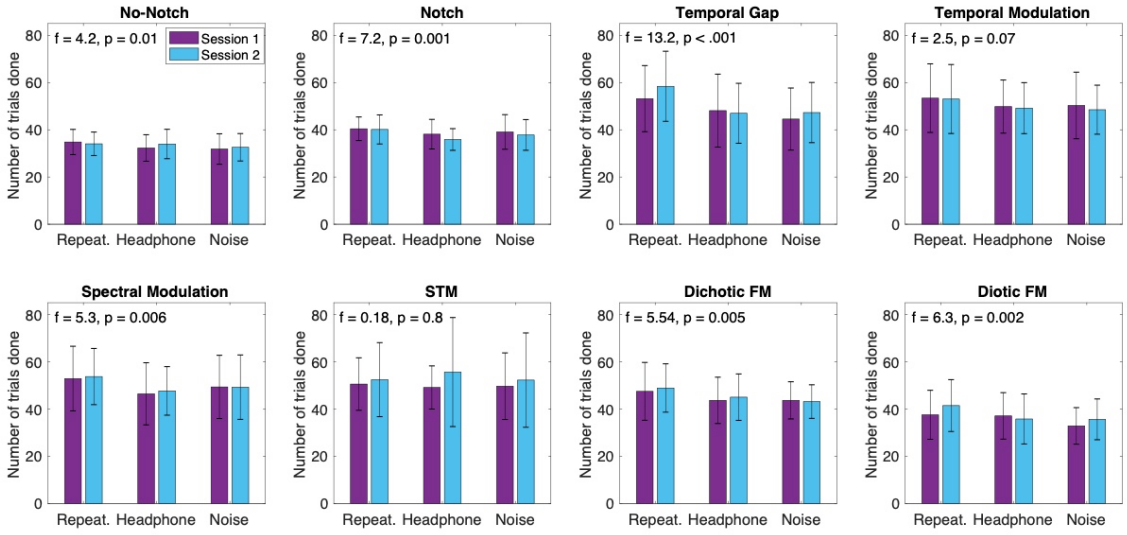

Figure S 4: Summary statistics for number of trials. Mean and standard deviations of the number of trials presented per task for each Experiment. Statistics from a one-way ANOVA with the between-subject factor Condition (3 levels) are displayed in the top of each graph.

To address the effects specific to the staircase, a series of independent samples t-tests were conducted between the number of trials needed to achieve a threshold estimate in the Repeatability vs Headphone conditions. These two conditions are the most similar (apart from the headphone differences in one of the sessions of the Headphone condition) and so are the best place for examining the difference between step size ratios. It would have been possible to conduct this comparison only on the conditions using the same headphones, but this would have reduced the size of the data set in half. The number of trials for the adaptive staircases are included in the Supplemental dataset to permit alternative analyses to be conducted. Differences that met statistical significance were obtained in a number of the assessments, supporting the hypothesis that tests with the 1.5:1 staircase were more likely to finish in fewer trials. The effect sizes were relatively small, as differences are no more than 6 trials on average and sum up to 25.8 trials on average for the whole battery. On the other hand, there were no significant differences between the number of trials for any of the tests of the Headphone and Noise conditions, both of which used the 1.5:1 ratio. These results indicate the change in step-size from 2:1 to 1.5:1 resulted in staircases that were slightly more efficient but equally reliable in threshold estimates. It is, of course, the case that every study to which data were compared in the main manuscript, and most in the literature, used equal step-sizes (Levitt, 1971), so further research will be required to determine whether

there is an advantage, or cost, related to using the uneven step-sizes chosen in the current study.

Instructions

Informed consent was obtained by a research assistant or experimenter before testing started. All participants were given demographic surveys and then heard the following instructions, read by a research assistant or experimenter:
"We are developing a set of tests with diverse types of sounds that will help us better understand your hearing abilities. As we age (or by injury) our hearing abilities decline, and it is important to better understand the nature of this loss so we can deal with diagnosis and rehabilitation. During your session, which will take roughly 45 minutes, you will be listening to sounds and responding to them on an iPad. Your participation will aid research aimed to help improve people's quality of life, please take it seriously and give it your best effort."

Participants then were taken to the test room, where they were given an iPad and set of headphones. They were verbally instructed to sit down and put the headphones on with the left phone aligned with the left ear. All were then read the following instructions:

"Now you will be tested on several different tasks. The instructions will be shown to you each time a task begins. The game is programed to challenge your hearing limits, so it will get harder to solve as you go. Please try to respond to the

best of your ability, the better you are able to perform, the quicker the program will find your limit and the task will end. This experiment is divided in four chunks, now you will start the first one, when it finishes, the instructions will ask you to call me so I can write down your scores. Please follow the instructions on the iPad carefully. Do you have any questions? I will stay in here with you for the first few trials in case you have any further questions. Good luck!"

The first thing that appeared to them on the iPad was an instruction screen that read:

"Welcome to PART! In this experiment, you will be responding to different series of sounds to test your hearing. In this first example, 4 squares will be presented, and will light up sequentially one after another. Your job is to find and touch the square that makes a sound as it lights up."

These instructions were followed by a familiarization/screening task involving ten trials in which they were required to detect a 2kHz tone presented in one of the two test intervals, with silence presented during the other three intervals. Performance was monitored and instructions were repeated if necessary. Some participants did not anticipate that the noise would be near detection threshold and so did not hear anything. They were re-instructed, the testing was restarted, and then all were able to detect at least 9 of the ten targets. At this point, all participants completed conditions No-Notch and Notch

from the Targets in Competition sub-battery, before which the following instructions were shown on the iPad screen:

“Next, you will be looking for the same bip sound, but this time, noise will play on every square. Your job is to find the square that makes a bip sound in addition to the noise. The program will try to find the amount of noise necessary for you to be unable to find the bip sound, just try your best and guess if uncertain. Hint: the first and last squares never contain the bip.”

Participants then were randomly assigned to complete the second half of the Signals in Competition sub-battery (SRM assessments), the STM sub-battery or the TFS sub-battery.

STM

The STM sub-battery used similar descriptions for all three assessments it contained. It started with the following instructions:

“One of these sounds is not like the others... On every trial, four squares will be presented on screen. They will light up and emit a sound, one at a time. The first and the last squares will always carry some type of "ordinary" sound. One of the two squares in the middle will have a "special" modification, the other will carry "ordinary" sound. Your task is to identify which of the squares in the middle carries the "special" sound.”

TFS

The TFS sub-battery contained the following instructions for Dichotic FM:

“One of these sounds wobbles... On every trial, four squares will be presented on screen. One of the two squares in the middle will have a "special" modification, the sound they emit will seem to wobble between your ears! Your task is to identify which of the squares in the middle carries the "special" wobbling sound.”

The following are instructions for Diotic FM:

“Now the sound modification will be slightly different. Instead of wobbling between your ears, one of the sounds will fade in and out. Can you detect the modified “special” sound?”

The following are instructions for the Gap detection:

“Next, each of the squares will carry two clicks so close to each other they sound like only one. One of the squares in the middle will carry a pair of clicks with a bigger separation between them so they will sound slightly different. Your task is to identify the square with the different pair of clicks.”

Targets in Competition

The second half of the Signal in Competition sub-battery contained the SRM assessments. Participants were shown the following instructions:

“In each trial, you will hear a person call the name Charlie followed by the directions to go to a specific combination of color and number. Press the button on the grid that corresponds to such directions.”

The remaining assessments contained similar instructions, and small breaks were provided between assessments. Each sub-battery would end with an instructions screen with the following text or its equivalent:
“You have completed this part of the evaluation, please let an experimenter know you are done :)”.

An experimenter or research assistant would load the next sub-battery to continue testing or finish the session.

CHAPTER THREE B: Portable Automated Rapid Testing (PART) of auditory processing abilities in young normally-hearing listeners: A remotely administered replication with participant-owned devices.

This chapter presents a replication which was first intended for a special issue on testing hearing in times of the global pandemic where the possibility of using the assessments validated in the previous chapter in remote conditions by exploiting the participant's own devices and headphones, and PART's availability across a number of platforms. This work can be considered a natural extension of the findings of the previous chapter to highly variable conditions in terms of device, headphone and environmental settings. An additional group of participants is currently being tested to help disentangle these separate sources of variance. The work in its current form can be found published online as a preprint here: https://psyarxiv.com/9u68p/.

ABSTRACT

The COVID-19 pandemic has raised awareness of the need for robust and reliable remote testing of auditory function. Here, we examine how the recently introduced Portable Automatic Rapid Testing (PART) system–validated to produce precise psychoacoustical data in consumer hardware [Larrea-Mancera

et al., JASA, 2020]—performs when data are collected remotely on participant-owned uncalibrated smart-phones. To accomplish this, we compare data collected remotely, to a published dataset that was collected in a lab-based sample using standardized calibrated hardware. Performance was examined in a group of 40 participants with PART assessments administered via a video-call. Results largely matched the normative dataset collected in the laboratory with, on average, slightly worse performance and similar repeatability; however, the rate of outlying performance did increase for some assessments suggesting that some testing settings may not be appropriate for adequate data collection in some cases. These data suggest the feasibility of remote auditory testing on participants' own devices for suprathreshold tests of auditory processing. Future work is needed to better determine the adequacy of different remote settings for reliable psychoacoustic data collection or clinical use.

INTRODUCTION

With the COVID-19 pandemic, valid and reliable remote evaluation of hearing ability has become an essential need for acoustical researchers and clinicians (see the Technical Committee on Psychological and Physiological Acoustics (PP) of the Acoustical Society of America wiki on remote testing https://www.spatialhearing.org/remotetesting/). The advantages of remote testing are manifold and reach beyond the immediate needs of the pandemic. Reliable psychoacoustical testing outside of the confines of the lab would produce lasting transformative changes in data collection that could increase accessibility and inclusion for both basic research and clinical auditory assessment. This transformation is within reach as new developments in technology afford high quality audio generation software that can be run across a number of platforms that range from mobile phones to tablets and personal computers. *PART* (Portable Automatic Rapid Testing; Gallun et al., 2018), which was developed by the University of California Brain Game Center (https://braingamecenter.ucr.edu) and is freely available for a wide range of mobile and desktop/laptop operating systems, has substantial potential to address this growing need.

Larrea-Mancera et al. (2020) reported data from *PART* collected in a laboratory setting on measures of spatial, spectral, and temporal sensitivity using both synthetic stimuli (tones and noises) and speech. Data were collected

running *PART* on a calibrated platform consisting of an iPad (Apple, Cupertino CA, USA) and Sennheiser 280 Pro headphones (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany). Thresholds obtained were consistent with those obtained using laboratory-grade equipment in prior research, even when measured in moderate levels of background noise and with some headphone variability. These results suggest that *PART* may provide an appropriate platform for remote testing, although there are numerous challenges to the success of such a suggestion including the wide range of computers/smart-phones and headphones that participants may have access to, lack of calibration, as well as concerns of ambient sound levels and other distractors in people's homes.

In the present study, we adapted *PART* to be run by participants on their own uncalibrated devices and tested the repeatability of performance measurement in a home environment and how distributions of data compared to a normative dataset collected in the more controlled laboratory setting. Experimental sessions were conducted via video-calls and participants were instructed to download *PART* onto their own device and completed the assessment battery using their own headphones. We present thresholds that are about half a standard deviation worse than what would be expected from the normative dataset (Larrea-Mancera et al., 2020), but with comparable test repeatability. These results suggest that, at least for suprathreshold tests of hearing, that lack of calibration may lead to a relatively small offset in

performance, but with sufficient repeatability that would still be informative to basic research and clinical screening procedures.

METHODS

The methods of Larrea-Mancera et al. (2020) were replicated as closely as possible. Modifications were limited primarily to those necessary for remote data collection. Due to a programming error, the notch-noise task was run with the wrong parameters in the no-notch condition, and so this test is excluded from analyses. All other tests (9 of 10) were exact replications of the previous study in terms of parameters, instructions, and order of presentation. Since we did not have access to calibration of the devices and headphones used in the study, we do not have an exact measurement of the output levels presented to participants in dB SPL. However, given that the system still specifies desired dB values, which would be accurate once calibrated, we refer to the levels presented as “nominal dB” to facilitate comparisons across datasets collected with PART (calibrated and otherwise).

Additionally, we included an audibility sub-battery where failure to achieve a minimum level of performance in two out of three tests (40 nominal dB for 2 kHz and broad-band noise detection; 50 nominal dB for single talker discrimination from the Coordinate Response Measure (CRM) corpus) (Bolia et

al., 2000) was considered rejection criteria from analysis. Outlying values beyond 3 standard deviations (SD) from the normative dataset (Larrea-Mancera, 2020) were also excluded from analysis unless indicated otherwise.

Participants

Listeners were 40 undergraduate students from the University of California, Riverside (24 female, *M* age = 20.1 years, *SD* = 2.2 years), who received course credit or were payed $10/hr for participation. Four participants whose audibility thresholds failed to reach the minimum cutoff in two out of three tests were excluded from analysis. All participants reported normal hearing and vision, and no history of psychiatric or neurological disorders and provided electronic informed consent as approved by the University of California, Riverside Human Research Review Board.

Materials

Participants completed the entire experiment remotely and all procedures were conducted using their own uncalibrated smart-phones or tablets (34 iOS and 6 Android devices) and with their own headphones or earbuds (categorized for descriptive purposes as either in-ear or on-ear). Of the 6 participants using Android devices, two used in-ear and four used on-ear headphones. Of the 34 participants using iOS devices, 25 used in-ear and 9 used on-ear headphones.

To streamline the user experience, a variant of PART (called *BGC Science* that uses the same code-base as PART), was employed. This allowed participants to enter a server code that configures the appropriate test battery for each of them. Upon completion of the test session, data were encrypted and securely transmitted to an Amazon Web Services server (Amazon Web Services, Seattle, WA, USA). No participant identifiers were transmitted or stored with the data.

Assessments

We used the same task structures as in Larrea-Mancera et al. (2020). Responses for the speech in competition tasks (described below) that use CRM corpus (Bolia et al., 2000) were collected using a colored number grid (5 colors and numbers 0-9). All other tasks were collected using a 2-cue, 2-alternative forced choice (2-Cue 2-AFC) task where four squares were presented on screen and lit up sequentially in synchronicity with the sounds. Target sounds were presented in one of the two squares in the middle of the sequence and represent the alternatives of response. The first and last squares were presented as cues. This structure ensures the target stimulus was always preceded and followed by a standard stimulus.

Minimum Audibility

Pure tone detection in quiet

A progressive tracking algorithm was used in which a 2 kHz tone was presented for 100 ms at a level of 70 nominal dB (based on assumed electroacoustic relationships) and was thereafter reduced in level every 3 responses in steps of 5 nominal dB, until the tone level of 5 nominal dB was reached, or 3 errors occurred in the space of six trials.

Broad-band noise detection in quiet

A progressive track was used to reduce the level of a white noise stimulus, with starting value of 70 nominal dB, and 5 nominal dB (or 3 errors in 6 trials) stopping rule as for the pure tone detection in quiet.

Single talker speech identification

A single talker was chosen randomly from the first three male talkers in the CRM corpus (Bolia et al., 2000) and one of the stimulus sentences spoken by that talker was presented on each trial. The sentences all included the call sign "Charlie" and two keywords: a number and a color. Responses were collected by selecting a single button a graphical display composed of a grid of colored numbers. Performance was measured based on correctly identifying both keywords, and the level of the sentence was adjusted using the starting value and stopping rule as for the pure tone detection in quiet.

Temporal Fine Structure

Temporal Gap

This gap discrimination task, based on Hoover et al. (2019) uses the 2-Cue 2-AFC procedure where two 4-ms tone bursts (cropped Gaussians) were presented diotically at a level of 80 nominal dB on each of the four intervals. Tone bursts, had a carrier frequency of 0.5 kHz. Target intervals contained a brief silent gap between the end of the first burst and the beginning of the second. The gap between bursts in the target interval was initially 20 ms in duration and a two-stage adaptive tracking algorithm was used where gap duration was adapted on an exponential scale with first-stage descending steps of $2^{1/2}$ and second-stage descending steps of $2^{1/10}$ with a minimum gap duration value of 0 ms and a maximum duration of 100 ms.

Diotic Frequency Modulation

This FM detection task, also based on the method of Hoover et al. (2019), which was modified from the work of Grose and Mamo (2012). Every interval a 400-ms tone with a carrier frequency between 460 and 550 Hz was presented at a level of 75 nominal dB. Participants were instructed to detect the interval in which a tone is presented diotically (identical at the two ears) with a frequency modulation rate of 2 Hz on the carrier. Carrier frequencies were roved across intervals to discourage the performance of the task by simply detecting spectral energy outside of the standard carrier frequency range. A two-stage adaptive tracking

algorithm was used on an exponential scale with descending first-stage step sizes of $2^{1/2}$ and second-stage step sizes of $2^{1/10}$ starting at 6 Hz with a minimum value of 0 Hz and a maximum value of 10,000 Hz.

Dichotic Frequency Modulation

This FM detection task (Hoover et al., 2019); Grose and Mamo, 2012), used the same stimuli and tracking methods as in the diotic FM task, but in this case the FM was inverted between the ears, so that as the frequency increased at one ear it decreased at the other. This stimulus results in a time-varying interaural phase difference (IPD) that is perceived as a tone of fixed frequency varying in interaural location when the modulation rate is sufficiently small relative to the carrier frequency, such as the 2 Hz used here (Witton et al., 2000). The same procedure and adaptive tracking was used as in the diotic FM task, as well as the same starting modulation range of 6 Hz.

Spectro-Temporal Sensitivity

All three of these tasks involved a standard 500-ms noise band with a width of 0.4 kHz to 8 kHz presented at a level of 65 nominal dB. In each task described below, the 2-Cue 2-AFC procedure was used and participants identified the target interval by the presence of temporal, spectral, or spectrotemporal modulation. The adaptive parameter (modulation depth) was applied on a logarithmic amplitude scale (dB) and measured from the middle of the amplitude range to the peak amplitude as described in Stavropoulos et al. (2021).

Modulation depth is expressed throughout the manuscript as M (dB). The adaptive tracking procedure used large descending steps of 0.5 nominal dB and small descending steps of 0.1 nominal dB, with a minimum modulation depth of 0.2 nominal dB and a maximum modulation depth of 40 nominal dB.

Temporal Modulation

The TM detection task, based on Viemeister (1979), required participants to detect sinusoidal temporal amplitude modulation (AM) at a rate of 4 Hz.

Spectral Modulation

The SM detection task, based on Hoover, Eddins & Eddins (2018), required the detection of sinusoidal spectral modulation with random phase at a rate of 2 cycles per octave (c/o).

Spectro-Temporal Modulation

The STM detection task, based on Bernstein et al., (2013), involved the detection of a stimulus that included both sinusoidal temporal amplitude modulation (AM) at a rate of 4 Hz and sinusoidal spectral modulation with random phase at a rate of 2 c/o.

Targets in competition

Notch Noise test

Targets were 500 ms, 2 kHz pure tones presented at a level of 45 nominal dB in the presence of two noise bands with frequency ranges of 0.8-1.6 kHz and 2.4-

3.2 kHz, leaving a 0.8 kHz notch centered on the target signal, as described by Moore (1987). The 2-Cue 2-AFC procedure was used adapting on RMS level of the noise, starting at nominal dB of 35 with two-stage (6 nominal dB for 3 reversals and then 2 nominal dB for 6 reversals), 2-down, 1-up adaptive tracking with a 1.5:1 up/down step-size ratio as described in Larrea-Mancera et al. (2020). Thresholds were estimated from the geometric mean of the last six reversals. Masker levels were not allowed to exceed 90 or go below 25 nominal dB.

SRM Colocated

This three-talker speech-on-speech masking task uses the CRM corpus described in the audibility task (Bolia et al., 2000) and the presentation and scoring developed by Gallun et al. (2013). Target sentences all included the call sign “Charlie” and two keywords: a number and a color, fixed at an RMS level of 65 nominal dB. The target was presented simultaneously with two maskers, which were male talkers uttering sentences with different call signs, colors and numbers in unison with each other and with the target. All three sentences were presented from directly in front of the listener (colocated). Progressive tracking included 20 trials in which the maskers progressed in level from 55 to 73 nominal dB in steps of 2 nominal dB every two trials, with no stopping rule, as described in Gallun et al. (2013).

SRM Separated

The procedure, stimuli, and progressive tracking were identical to those in the colocated condition, with the exception that the maskers were presented from 45 degrees to the left and right of the target talker, after spatialization with generic head-related transfer functions as described in Gallun et al. (2013).

Spatial Release from Masking Metric

Obtained from subtracting the masker level threshold obtained in the Colocated condition from that obtained in the Separated condition. Positive values indicate benefit from spatial cues in nominal dB.

Procedure

Each session began with a video call where participants completed a screening to ensure that sounds from the left and right channels were heard in the left and right ears respectively, and then completed the Audibility sub-battery followed by the Notch tests in the Targets in Competition sub-battery (2 assessments). Following this, procedures were identical to those described in Larrea-Mancera et al. (2020), where participants conducted the Temporal Fine Structure (TFS; 3 assessments); Spectro-temporal Modulation (STM; 3 assessments); and the speech tests in the Targets in Competition sub-battery (2 assessments) in pseudo-random order. Participants were encouraged to take short breaks between testing blocks. Each of the ten assessments took about 5 minutes to

complete, resulting in test sessions of around 1 hour. The second session was identical to the first and was always conducted on a different day no longer than a week after the first session.

RESULTS

Results are divided into sections for the purpose of clarity. First, to evaluate thresholds for each assessment across two sessions is compared to the normative dataset from (Larrea-Mancera et al. (2020) (Section A). Then, the repeatability of measurements across sessions was evaluated by comparing differences across sessions between both studies (Section B). Sections A and B include an outlier rejection criteria of ± 3 SD away from the mean using the normative dataset parameters. Section C explores the variance obtained with and without outliers. This section also includes an analysis regarding the repeatability of our measures including outliers and an identification of outliers and their reliability is provided. Finally, the effects of headphones and device are explored including outliers (Section D).

Are remotely administered thresholds comparable to those collected in lab?

Figure 3b.1 presents the distributions of thresholds obtained in the two sessions of the remote replication study ("Home", lighter distributions) and replotted from the normative data sets of Larrea-Mancera et al. (2020) ("Lab", darker distributions). Symbols of different types are used to indicate the different types of hardware used by participants (combinations of iOS and Android devices, and in-ear or on-ear headphones). A mixed-effects ANOVA with the between-subject factor of Experiment (HOME vs LAB) and within-subjects factor Session (ONE vs TWO) was run on each of the nine assessments used and the spatial release metric. Relevant statistics including *p* values are presented in Table 3b.1. There were small but statistically significant main effects of Experiment ($p \leq .01$) across each of the assessments except for the spatial release metric.

These effects are approximately half a standard deviation in magnitude (*Cohen's D* from .38 to .65). Effect sizes expressed not in standard deviations, but in the original units, correspond to changes in threshold for the remote testing of 2.2 ms for the Gap task, 0.3 Hz for the Dichotic FM, 2 Hz for the Diotic FM, from 0.32 to 0.61 Modulation nominal dB for the Spectro-temporal modulation assessments, and between -1.49 to 1.86 TMR (nominal dB) differences in the targets in competition assessments (see Table 3b.1). These results suggest that, on average, remote administration with personal equipment in home

environments yields worse performance than the laboratory conditions tested in the original study, with the exception of the Notch task, which was slightly better.

While these differences are statistically significant, with small to medium effect sizes, this performance is still within the range of that found in the published literature. While this suggests that there may be a systematic offset in performance in the uncontrolled settings, which is not particularly surprising, the key question is the extent to which these effects are repeatable across sessions and that differences between subjects are reliable across these sessions. If so then this information can still provide utility in comparing between different individual's hearing abilities.

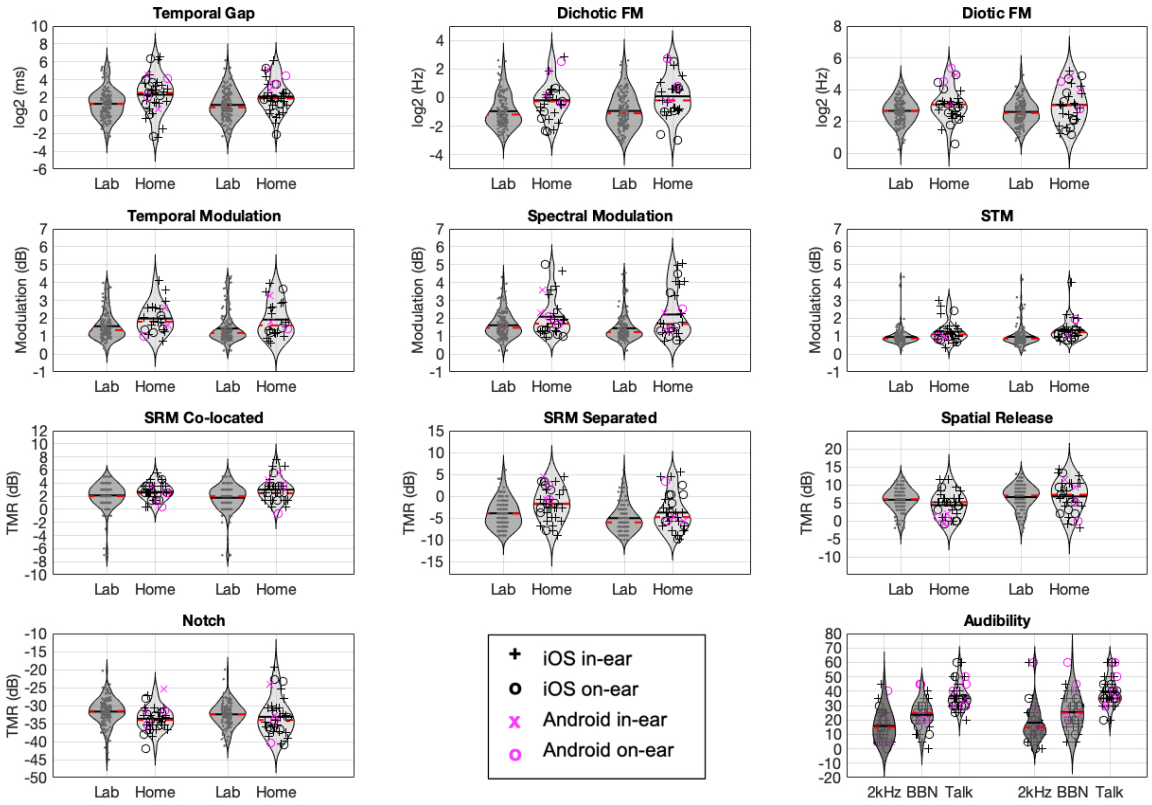

Figure 3b. 1: Threshold comparison across studies. Plots represent the density functions and the spread of datapoints for Larrea-Mancera et al. (2020) (darker) and this study (lighter distributions). The dashed line inside each density function represents the median and the solid line the mean of each distribution (session). In most cases, the solid and dashed lines are completely overlapping.

Assessment	*Mean (SD) Session 1*	*Mean (SD) Session 2*	*Mean Diff.*	*Replication f (p)*	*df*	η_p^2	*Cohen's D*
Gap (lab) @ home	2.46 (3.15) 4.5 (4.59)	2.26 (3.18) 4.66 (3.97)	2.22 ms	11.74 (<.01) *	183	0.06	0.5
DichoticFM (lab) @ home	0.51 (2.23) 0.82 (2.48)	0.52 (2.37) 0.98 (2.68)	0.37 Hz	13.96 (<.01) *	179	0.07	0.55
DioticFM (lab) @ home	6.35 (1.75) 8.16 (2.005)	6.05 (1.77) 8.24 (2.25)	2 Hz	7.7 (<.01) *	177	0.04	0.41
TM (lab) @ home	1.55 (0.81) 1.92 (0.82)	1.42 (0.85) 1.92 (1.005)	0.43 M(dB)	7.69 (<.01) *	170	0.04	0.42
SM (lab) @ home	1.61 (.72) 2.06 (1.07)	1.44 (.78) 2.21 (1.34)	0.61 M(dB)	18.22 (<.01) *	170	0.09	0.65
STM (lab) @ home	0.95 (0.46) 1.24 (0.61)	0.96 (0.55) 1.31 (0.64)	0.32 M(dB)	13.96 (<.01) *	165	0.07	0.56
Notch (lab) @ home	-31.67 (3.64) -33.77 (3.35)	-32.44 (3.32) -33.32 (5.37)	-1.49 TMR(dB)	6.65 (.011)*	177	0.03	0.38
Co-located (lab) @ home	2.12 (1.96) 2.89 (1.58)	1.76 (2.08) 3.2 (1.96)	1.1 TMR(dB)	13.97 (<.01) *	181	0.07	0.55
Separated (lab) @ home	-3.91 (3.32) -1.81 (3.68)	-5.04 (3.2) -3.4 (4.37)	1.86 TMR(dB)	11.9 (<.01) *	183	0.06	0.51
Spatial R. (lab) @ home	5.8 (3.24) 4.43 (3.38)	6.58 (3.37) 6.41 (4.44)	0.77 dB	2.17 (.28)	176	0.01	0.22

Table 3b. 1: Comparative Statistics. The mean and standard deviation (SD) obtained in each experiment (Lab on top, this study (Home) on bottom row of each assessment) are displayed for the ten assessments that are comparable across the two experiments. The mean difference column shows the difference between the averaged sessions (1&2) of each experiment and gives an estimate of effect size in the measured units. The replication column shows results for the mixed-model ANOVAs (main effect of Experiment) of each assessment. Effect sizes in terms of variance explained by which Experiment produced the data are provided for the ANOVA in terms of partial eta^2 and Cohen's D.

Are remotely administered thresholds repeatable and reliable?

To investigate the extent to which the differences observed between sessions when testing involved uncalibrated equipment and a home environment were similar to those obtained in the laboratory (Larrea-Mancera et al., 2020), the interaction term (Experiment*Session) of the ANOVA detailed in the previous

section was examined (see Table 3b.2). None of the interactions between Experiment and Session were statistically significant, indicating similar repeatability to that observed in the normative study. In order to visualize the distribution of the difference between sessions in both studies, difference scores were calculated by subtracting the threshold obtained in Session 1 from that obtained in Session 2 in both studies. Since this same subtraction is used in the limits of agreement analysis (Bland & Altman, 1999) shown in Larrea-Mancera et al. (2020), to establish the limits at which 95% of the differences between sessions are expected to occur, the limits of agreement are shown as dotted lines (Figure 3b.2). Again, different symbols are used to indicate the classes of personally-owned hardware used. Of note, the interaction term shown in Table 3b.2 is statistically equivalent to subjecting the difference scores to a one-way ANOVA with Experiment as factor (LAB vs HOME).

Assessment	*Δ Session Lab*	*Δ Session Home*	*Repeatability F (p value)*	*df*	η_p^2	*Cohen's D*
Gap	-0.2 ms	0.16 ms	0.32 (.57)	183	0.002	0.08
DichoticFM	0.01 Hz	0.16 Hz	1.2 (.27)	179	0.007	0.16
DioticFM	-0.3 Hz	0.08 Hz	0.25 (.61)	177	0.001	0.07
TM	-0.13 M (dB)	0.007 M (dB)	0.57 (.45)	170	0.003	0.11
SM	-0.17 M (dB)	0.15 M (dB)	2.99 (.08)	170	0.01	0.26
STM	0.01 M (dB)	0.07 M (dB)	0.21 (.64)	165	0.001	0.06
Notch	-0.77 TMR (dB)	0.45 TMR (dB)	2.63 (.107)	177	0.01	0.25
Co-located	-0.36 TMR (dB)	0.31 TMR (dB)	2.96 (.08)	181	0.01	0.25
Separated	-1.13 TMR (dB)	-1.59 TMR (dB)	0.52 (.47)	183	0.003	0.1
Spatial Release	0.78 dB	1.98 dB	2.73 (.2)	176	0.005	0.24

Table 3b. 2: Repeatability Statistics. Differences between the mean thresholds in session 1 and session 2 are shown for each experiment (Lab (Larrea-Mancera et al., 2020) and Home (this study). Except for Spatial Release, negative values indicate an improvement from session 1 to 2. The repeatability column shows results for the mixed-model ANOVAs interaction term (Experiment*Session) of each assessment. For all assessments, the F values are not statistically significant ($p > .05$). Partial eta^2 shows an estimate of the variance captured by the interaction, and Cohen's D expresses the size of the difference in units of standard deviation.

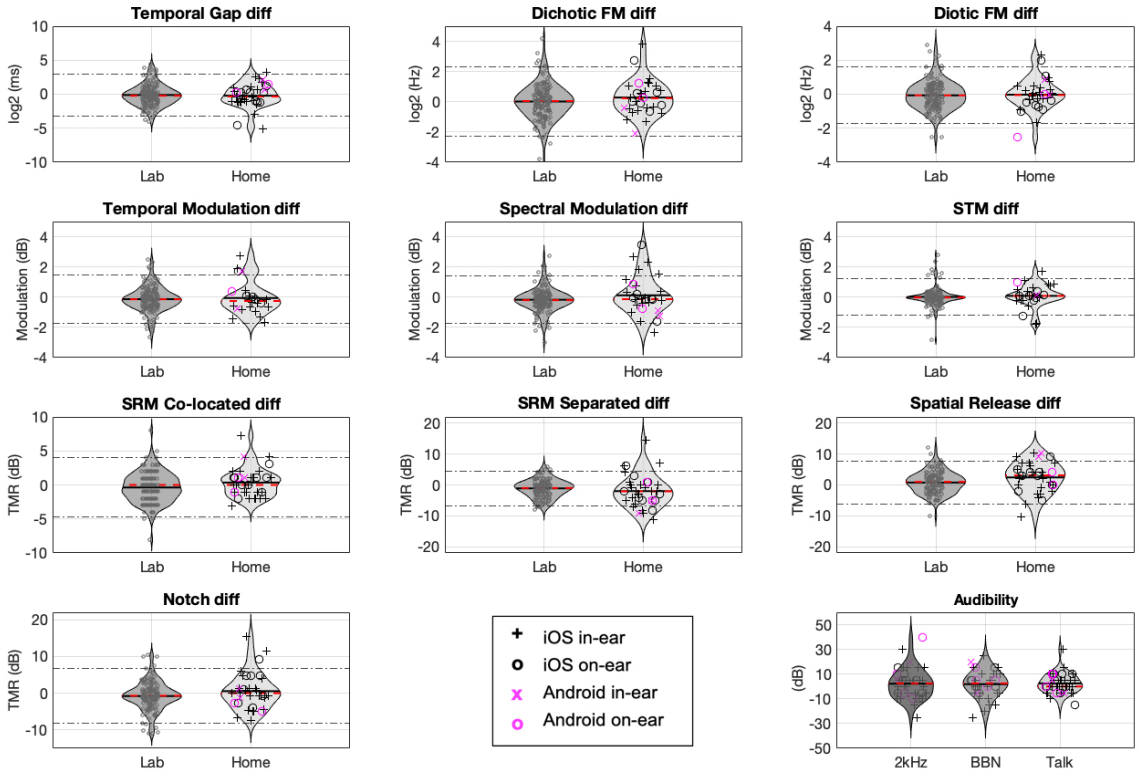

Figure 3b. 2: Repeatability across studies. Plots represent the density function of the differences between sessions (session 2 – session 1) in Larrea-Mancera et al. (2020) (darker) and this study (lighter distributions). The dashed line inside each distribution represents the median and the solid line represents the mean of each study. Dark dotted lines represent the smallest step difference above and below perfect reliability (zero). The dotted lines depicted in the background represent the 95% limits of agreement extracted from the normative dataset.

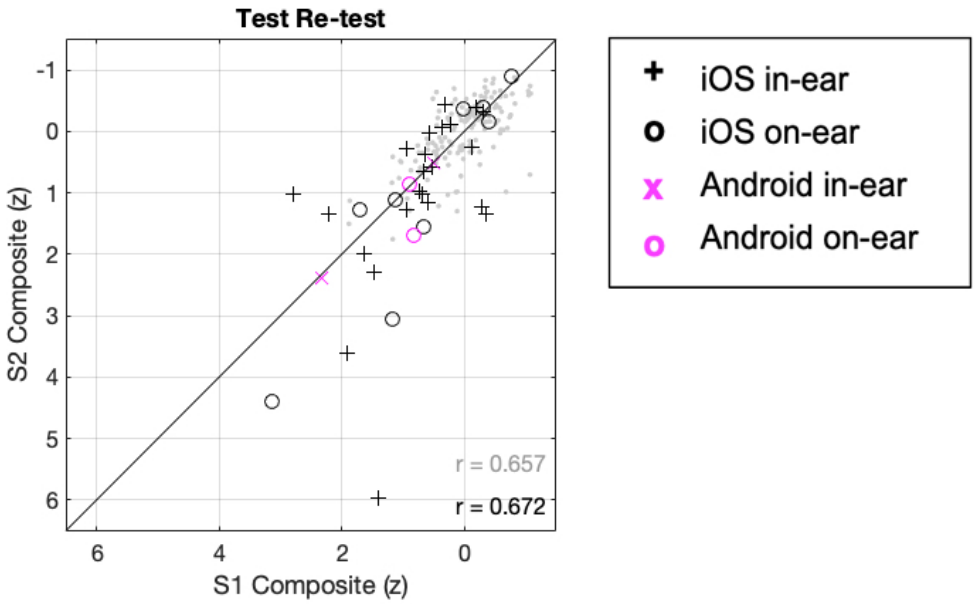

Figure 3b. 3: Composite score correlations. The correlation across session for composite scores of both experiments are plotted for the different device and headphone combinations used in this study. The normative dataset composite scores are depicted in light grey and different markers are used for the remotely collected data to indicate different headphone and device combinations. R values (gray for normative dataset) are significant ($p < .001$).

To further analyze the reliability of remote measures and further show the differences found in section A correspond to a systematic offset of the thresholds calculated remotely, we conducted normalized correlations on composite scores across sessions and present them in Figure 3b.3. A composite score was calculated for each subject on each session as the average of z-scores in all 9 assessments, using the means and standard deviations from the normative dataset (Larrea-Mancera et al., 2020). Figure 3b.3 depicts the strength and direction of association between scores obtained in each session. We compared

the obtained Pearson r's *(@lab r* = 0.65; *@home r* = 0.67) with a Fisher z-test and found they were not statistically significantly different (z = -0.19, p = 0.42).

Variance exploration and outlier analysis

The statistical analyses presented so far indicate that there is a small to medium, but reliable, worsening of performance when calibrated equipment and a laboratory test environment is replaced by the participant's own uncalibrated phone and headphones, and that reliability across test sessions is unaffected. However, the data plotted in Figures 3b.1 & 3b.2 suggest that there may be more participants for whom the thresholds obtained are well outside the range of expected results. To explore the degree to which outliers were more prevalent with at home testing, a mixed-effects ANOVA with the between-subject factor of Experiment (LAB vs HOME) and within-subjects factor Session (ONE vs TWO) was conducted on the variance statistic of all nine assessment and the spatial release metric distributions. Neither Experiment ($F_{(1,16)}$= 1.05, p = .32, η_p^2 = 0.06) nor Session ($f_{(1,16)}$= 0.27, p = .61, η_p^2 = 0.01) showed significant differences in variance, nor was there a statistically significant interaction ($f_{(1,16)}$= 0.42, p = .52, η_p^2 = 0.02).

Similar results were obtained when including outliers in both samples with no statistical significant main effects either of Experiment ($f_{(1,16)}$= 0.84, p = .37, η_p^2 = 0.05) or of Session ($f_{(1,16)}$= 1.64, p = .21, η_p^2 = 0.09), nor was there a

statistically significant interaction ($f_{(1,16)}$= 0.58, p = .45, η_p^2 = 0.03). These data suggest that, at least within the normative range, the at-home sample contained variability that was not statistically different from that obtained in the laboratory tests.

Assessment	*Grand Mean (SD) All Data*	*Grand Mean (SD) No outliers*	*Cases Rejected ± 3 SD Of norms*	*Percentage of outliers consistent across sessions (<1z diff.)*
Gap (lab)	2.51 (3.38)	2.36 (3.16)	2%	0
@ home	4.63 (3.91)	4.63 (3.91)	0	0
DichoticFM (lab)	0.53 (2.41)	0.52 (2.3)	1%	100%
@ home	1.23 (3.04)	0.89 (2.57)	13.8%	28%
DioticFM (lab)	6.38 (1.9)	6.2 (1.76)	3%	40%
@ home	10.15 (2.64)	8.2 (2.11)	13.8%	0
TM (lab)	1.59 (1.10)	1.49 (0.83)	2%	50%
@ home	2.85 (2.002)	1.92 (0.91)	33.3%	58%
SM (lab)	1.71 (1.12)	1.52 (.75)	6%	44%
@ home	2.56 (1.82)	2.13 (1.2)	19.4%	28%
STM (lab)	1.18 (1.04)	0.95 (0.51)	8%	30%
@ home	2.03 (1.95)	1.27(0.62)	19.4%	71%
Notch (lab)	-30.98 (8.37)	-32.06 (3.5)	4%	0
@ home	-32.81 (5.88)	-33.55 (4.45)	2.7%	0
Co-located (lab)	1.51 (2.83)	1.94 (2.03)	4%	70%
@ home	2.82 (1.53)	3.04 (1.78)	0	0
Separated (lab)	-4.34 (3.52)	-4.47 (3.31)	1%	0
@ home	-2.5 (4.34)	-2.61 (4.09)	2.7%	0
Spatial R. (lab)	5.86 (3.58)	6.19 (3.32)	6%	55%
@ home	5.33 (4.22)	5.42 (4.04)	2.7%	0

Table 3b. 3: Outlier exploration. Mean thresholds averaged across session are shown both with and without outlier rejection for all ten assessments. The fourth column shows the percentage of cases rejected from each dataset by the outlier rejection rule of ± *3 SD*. The fifth column shows the percentage of those participants whose thresholds were outside the criterion in both sessions.

To approach this from another angle, the number of outliers (3 *SD* from normative dataset parameters) obtained in Experiment were compared, and more were found in the at-home than the in-lab samples (see Table 3b.3). This was a particular issue for the TM assessment, with 33% of participants outside of the normative range and the STM and SM sub-batteries each showing 20% outliers. To examine the degree to which those performing outside the expected range were consistently doing so, session to session consistency was also evaluated (see Table 3b.3). While the TM outliers were relatively consistent (>50%) across sessions, most other tests had most of their outlying values restricted to a single session.

Equipment Effects

While there was considerable variability in terms of devices and headphones used, the preponderance of participants used in-ear headphones on iOS devices. As a result, there was insufficient statistical power to conduct meaningful significance testing contrasting devices or headphones even for the broad categories we used. Nonetheless, overall effects of these categories were characterized using the composite scores. These data are shown in the scatter plot presented in Figure 3b.3. To evaluate possible differences across equipment categories with a test of significance, a mixed-effects ANOVA was conducted, with the within-subjects factor Session (ONE vs TWO) and the between-subject

factor Device+Headphone (4 levels: iOS+In-ear, iOS+On-ear, Android+In-ear, Android+On-ear), outliers included. No significant main effects were found for either Device+Headpone ($f_{(3,36)}$= 0.18, p = .906, η_p^2 = 0.01) or for Session ($f_{(1,16)}$= 0.14, p = .707, η_p^2 = 0.004), nor was there a statistically significant interaction ($f_{(1,16)}$= 1.81, p = .16, η_p^2 = 0.13).

DISCUSSION

These results suggest that precise and reliable tests of auditory processing ability can be collected remotely using personally-owned equipment, at least for normally-hearing undergraduates. While thresholds were consistently worse with remote testing, the differences were no more than half a standard deviation in comparison to those obtained in-lab with calibrated hardware (Larrea-Mancera et al., 2020). Furthermore, precision as test-retest consistency as well as the variance of the distributions were similar. The consistency of these results suggest that in-lab vs at-home differences are not random and thus noisy measurement impacting the accuracy of estimated thresholds may be addressed via a corrective factor, such as subtracting 2 Hz from a Diotic FM detection threshold obtained at home in order to compare with published data collected in a laboratory. For the full range of corrective factors suggested by these data, see Table 3b.1. We note however, a greater proportion of the at-home sample

obtained thresholds more than three standard deviations worse than the mean of the normative data, suggesting that the less controlled at-home environment may occasionally lead to spurious test results. These do not seem to be directly related to the general device types used as performance within the expected range was found both on iOS and Andriod and with in-ear and on-ear headphones. Future work should explicitly recruit participants using devices and headphones of specific types (or provide such devices) to confirm this preliminary conclusion.

These results replicate and extend the findings of Larrea-Mancera et al. (2020) to show that the normative data may be applied to a wider range of devices and headphone types than were previously tested with a small to medium systematic impact on accuracy. These data also suggest that careful calibration might not be essential for these suprathreshold tests, although calibration cannot be ruled out as factor contributing to the small differences we found. Be that as it may, in the previous study differences in output levels greater than 10 nominal dB between the two headphone types tested did not affect thresholds or variability. Further testing is needed to expand these results beyond the limited sample of devices and headphones tested here.

Ambient noise is another important factor that could explain the small differences in estimated thresholds. Previously, no effect of cafeteria noise played through a loudspeaker was found on assessments carried out in PART.

Here, only self-report information was available regarding ambient noise levels. However, it is likely that incidental background sounds, especially those that are relevant to the participants, could have led to impairments in task-performance. One potential indication of the effects of transient noises is the occasional outlier performances, which were most likely to occur in the FM, TM, SM and STM tasks. It is possible that the targets in competition tasks are particularly immune to ambient noise and thus are best suited to at-home testing. An important next step would be to monitor ambient noise levels directly to clarify their impact on performance and better understand how different types of sounds differentially impact the various tasks.

Similarly, visual distraction, and other and ecologically relevant environmental factors, likely impact performance. Although sessions were monitored via videocalls to ensure participants were performing the task, this provides only a limited glimpse into the relevant events that might be happening in the remote *participant's own* environment. For example, many participants are in close quarters with other people, and animals, that are engaging in activities that can be distracting in ways that don't exist in normal laboratory settings. These factors are hard to control, but future studies could query participants in more detail on distractors in their environment.

This study suggests substantial potential for the clinical applicability of at-home testing with PART. While performance seems to be worse at home, this

effect appears to be relatively consistent and may be addressable through a correction factor. Likewise, while there were more outlying cases in the at-home study, these occurred inconsistently across sessions. One approach to outliers is to simply repeat any tests on which performance falls outside the expected range. If participants consistently show outlying performance, this would be evidence that they would benefit from being seen in a clinical setting, where environmental factors would be controlled. Importantly, the screening value of the at-home testing would still be high as it is unlikely that participants would perform in the normal range at-home but then show impairment in the clinic. Under this framework, remote testing does not represent an alternative to the clinic but rather compliments the needed but less accessible clinical procedures that can be used to follow-up remote screenings with greater accuracy. Next steps will involve validation of these remote testing procedures for people with hearing impairment to test the extent to which remote testing provides clinically relevant information.

The results of this study demonstrate the potential of conducting remote auditory testing using people's own devices in their home environments. Valid and reliable remote testing would increase access to clinical assessment, address both clinical and research needs related to the social distancing with the COVID-19 pandemic, and help generate larger datasets of auditory ability in diverse populations. The use of devices already distributed among the general

public accelerates data collection and facilitates the inclusion of a more diverse set of participants in clinical research studies. However, studies that require more experimental control could easily distribute calibrated devices at a lower costs than in person clinical assessment. While the data presented in this study are only a first start, additional controls, such as ambient sound monitoring and more precise assessment of the at-home environment, will lead to even more precise psychophysical measures of auditory function in home environments, which will advance both auditory research and clinical practice.

REFERENCES

Bernstein, J. G., Mehraei, G., Shamma, S., Gallun, F. J., Theodoroff, S. M. and Leek, M. R. (**2013**). "Spectro-temporal modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners," J Am Acad Audiol **24**(4), pp. 293-306.

Bland, J. M., & Altman, D. G. (**1999**). "Measuring agreement in method comparison studies," Statistical Methods in Medical Research, **8**(2), pp. 135–160.

Bolia, R. S., Nelson, W. T., Ericson, M. A. and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," Journal of the Acoustical Society of America **107**(2), pp. 1065-1066.

Gallun, F. J., Diedesch, A. C., Kampel, S. D., & Jakien, K. M. (**2013**). "Independent impacts of age and hearing loss on spatial release in a complex auditory environment," Frontiers in Neuroscience, 7, pp. 1–11.

Gallun, F. J., Seitz, A., Eddins, D. A., Molis, M. R., Stavropoulos, T., Jakien, K. M., Srinivasan, N. (**2018**). "Development and validation of Portable Automated Rapid Testing (PART) measures for auditory research," **33**(175). Paper 2pPPb15.

Grose, J. H., and Mamo, S. K. (**2010**). "Processing of temporal fine structure as a function of age," Ear and Hearing, **31**, pp. 755-760.

Hoover, E. C., Kinney, B. N., Bell, K. L., Gallun, F. J., & Eddins, D. A. (**2019**). "A Comparison of Behavioral Methods for Indexing the Auditory Processing of Temporal Fine Structure Cues," Journal of Speech, Language, and Hearing Research : JSLHR, **62**(6), pp. 2018–2034.

Hoover, E. C., Eddins, A. C., & Eddins, D. A. (**2018**). "Distribution of spectral modulation transfer functions in a young, normal-hearing population," The Journal of the Acoustical Society of America, **143**(1), pp. 306–309.

Larrea-Mancera, E.S.L., Stavropoulos, T., Hoover, E., Eddins, D., Gallun, F., & Seitz, A. (**2020**). “Portable Automated Rapid Testing (PART) for auditory research: Validation in a normal hearing population,” *Journal of the Acoustical Society of America*, *148*(4), 1831–1851. https://doi.org/10.1101/2020.01.08.899088

Moore, B. C. J. (**1987**). “Distribution of auditory-filter bandwidths at 2 kHz in young normal listeners,” Journal of the Acoustical Society of America, **81**(5), pp. 1633–1635.

Stavropoulos, T. A., Isarangura, S., Hoover, E. C., Eddins, D. A., Seitz, A. R., & Gallun, F. J. (**2021**). “Exponential spectro-temporal modulation generation,” Journal of the Acoustical Society of America, **149**(3), pp. 1434–1443.

Technical Committee on Psychological and Physiological Acoustics (PP) of the Acoustical Society of America (ASA), (2020, November 17). Remote Testing Wiki. Spatial Hearing.org. https://www.spatialhearing.org/remotetesting/

Viemeister, N. F. (**1979**). “Temporal modulation transfer functions based upon modulation thresholds,” Journal of the Acoustical Society of America, **66**(5), pp. 1364–1380.

Witton, C., Green, G. G. R., Rees, A., & Henning, G. B. (**2000**). “Monaural and binaural detection of sinusoidal phase modulation of a 500-Hz tone,” The Journal of the Acoustical Society of America, **108**(4), pp. 1826–1833.

CHAPTER FOUR: Training with an auditory perceptual learning game transfers to speech in competition

This chapter presents an implementation of assessment and training on the mechanical sense of hearing. The auditory training studied integrates several elements of perceptual learning studies that have independently shown to promote generalization of learning and trainees were tested along several dimensions of auditory processing including some of the assessments of the previous chapters (3 & 3b). This work is currently available as a preprint here: https://doi.org/10.1101/2021.01.26.428343 and has been submitted for publication to the Journal of Cognitive Enhancement (April, 2021). It will be presented in the Perceptual Learning Workshop organized by Elizabeth Quinlan and Aaron Seitz in Alaska, 2022.

ABSTRACT

Understanding speech in the presence of acoustical competition is a major complaint of those with hearing difficulties. Here, a novel perceptual learning game was tested for its effectiveness in reducing difficulties with hearing speech in competition. The game was designed to train a mixture of auditory processing skills thought to underlie speech in competition, such as spectral-temporal processing, sound localization and auditory working memory. Training on these

skills occurred both in quiet and in competition with noise. Thirty college-aged participants without any known hearing difficulties were assigned either to this mixed-training condition or an active control consisting of frequency discrimination training within the same gamified setting. To assess training effectiveness, tests of speech in competition (primary outcome), as well as basic supra-threshold auditory processing and cognitive processing abilities (secondary outcomes) were administered before and after training. Results suggest modest improvements on speech in competition tests in the mixed-training compared to the frequency-discrimination control condition (*Cohen's d* = 0.68). While the sample is small, and in normally hearing individuals, these data suggest promise of future study in populations with hearing difficulties.

INTRODUCTION

Despite a vast amount of research conducted across multiple fields, clinicians and researchers still disagree about the best ways to confront the full diversity of hearing difficulties individuals face throughout their lives. Historically, auditory rehabilitation has been focused on the ability to *detect* sounds—audibility. To this end, hearing loss due to elevation of auditory detection thresholds can often be addressed through the use of amplification technologies such as hearing aids (Chisolm et al., 2007). However, although hearing aids can restore at least partial audibility for some listeners, even in the presence of competing sounds (Humes et al., 2009), and are increasingly recommended for those with hearing complaints associated with central auditory processing (CAP) dysfunction (Koerner, Papesh & Gallun, 2020), there is little documented clinical evidence supporting the prescription of hearing aids for those with pure-tone detection thresholds in or near the normative range for young adults. Moreover, amplification technologies may actually present difficulties in noisy environments since both sounds of interest and competing background noises are amplified together with no relative increase in the audibility of the signal.

Similarly, conventional hearing aids may not provide the best solution for those with supra-threshold auditory processing difficulties, which often manifest as a reduced capacity to *discriminate* among competing sounds and hinders

ones ability to separate auditory signals of interest from competing background noises: for example, individuals with supra-threshold auditory processing difficulties may struggle to understand one voice out of a group of many talkers even when sounds are audible (above hearing threshold). The more general case of this difficulty of hearing in multiple talker environments is often referred to as the cocktail party effect (Cherry, 1953; McDermott, 2009). Because currently there are no widely-accepted methods to assess and treat supra-threshold auditory processing difficulties, there is a significant need for novel approaches to evaluate and rehabilitate these common hearing complaints (Gallun et al., 2014; Weihing, Chermak & Musiek, 2015; Hoover, Souza & Gallun, 2017; Gallun et al., 2018; Larrea-Mancera et al., 2020).

Auditory training (AT) has been proposed as a promising rehabilitation approach for individuals experiencing hearing difficulties associated with supra-threshold auditory processing (Chermak & Musiek, 2002; Moore & Amitay, 2007; Weihing, Chermak & Musiek, 2015), including those already using hearing aids for sound amplification (for review see Henshaw & Ferguson, 2013; Stropahl, Besser & Launer, 2020). There is an extensive literature on AT employing a variety of training targets applied to a variety of target populations (see Ferguson & Henshaw, 2015). Training targets have ranged from simple frequency discriminations (Goldsworthy & Shannon, 2014) to phonemes (Ferguson, Henshaw, Clark & Moore, 2014; Kimball et al., 2013; Wade & Holt, 2005),

modified speech (Merzenich et al., 1996; Tallal et al., 1996), speech in noise (Burk, Humes, Amos & Strauser, 2006; Humes et al., 2014; Kuchinsky et al., 2014), active conversation listening (Lavie, Attias & Karni, 2013), and music (Schellenberg, 2016; Zendel, West, Belleville & Peretz, 2017). Target populations have included children with learning difficulties (Merzenich et al., 1996; Tallal et al., 1996), cochlear implant users (Goldsworthy & Shannon, 2014), young adults with normal hearing (Whitton, Hancock & Polley, 2014; Kimball et al., 2013; Wade & Holt, 2005), older adults both with normal hearing (Karawani, Bitan, Attias & Banai, 2016; Zendel et al., 2017), and those with hearing difficulties (Anderson et al., 2013ab; Henshaw & Ferguson, 2013; Whitton, Hancock, Shannon & Polley, 2017; Stropahl, Besser & Launer, 2020). However, the key limitation of many of these training studies is the lack of significant and lasting transfer of learning beyond the trained context (Seitz, 2017).

The goal of the current study is to test a novel approach to auditory training that targets multiple dimensions of hearing with the goal of achieving transfer to supra-threshold processing abilities such as the ability to recognize speech in competition. We adopt a novel “gamified” AT approach that integrates training principles from two main fields of knowledge: perceptual learning (PL; see Seitz, 2017) and video-game play (see Bavelier, Green & Dye, 2009).

In PL, transfer of learning to untrained stimulation or conditions has been shown after repeated training with perceptual stimuli when 1) the task is neither

too hard nor too easy (Ahissar & Hochstein, 1997; Ghose, Yang & Maunsell, 2002; Hung & Seitz, 2014), 2) training includes a diverse stimulus set (Deveau, Lovcik & Seitz, 2014; Xiao et al., 2008; Zhang et al., 2011), 3) exogenous (Donovan, Szpiro, & Carrasco, 2015) or endogenous attention is directed towards trained cues (Donovan & Carrasco, 2018), and 4) more than one sensory modality guides participant interactions with the training stimuli (Shams & Seitz, 2008; Shams, Wozny, Kim & Seitz, 2011).

Gamification is motivated based on findings that some commercial video games lead to broad improvements across a number of visual and cognitive processing skills (Bavelier, Green & Dye 2009; Bediou et al., 2018; Green & Bavelier, 2003). However, careful integration and design of game-elements is essential as game elements can also be distracting and interfere with learning (Katz et al., 2014; Mohammed et al., 2018, see also Seitz et al., 2010). Furthermore, games do not always focus performance on the intended processes. For example, Stewart et al. (2020) showed an advantage for action video-game players in visual but not auditory attention, and there was no difference on measures of speech-in-competition ability. Of note, even when auditory cues are useful in so-called “action video-games”, they rarely are essential for solving the tasks or maximizing outcomes, which may explain why visuo-spatial skills are more likely to be trained than are auditory skills.

Previous work at the University of California, Riverside Brain Game Center for Mental Fitness and Well-being (BGC) has successfully integrated the framework of PL with commercial video-game principles in the visual domain (Deveau & Seitz, 2014; Deveau, Lovcik & Seitz, 2014; Deveau, Ozer & Seitz, 2014). Deveau and colleagues developed a game where the goal was to quickly find oriented line patterns ("Gabor patches") that varied on a number of stimulus dimensions to train vision. The authors found that systematic training across visual primitive features such as the spatial frequencies, orientations, and locations of presentation of classic low-level visual stimuli (Gabor patches), with adaptive difficulty on detectability of the stimuli, resulted in broad transfer of learning across basic tests of vision (Deveau, Lovcik & Seitz, 2014), reading (Deveau & Seitz, 2014) and even to on-field performance in baseball athletes (Deveau, Ozer & Seitz, 2014).

Crucial to this approach was the use of stimuli that align with primitive features found to be systematically represented in the early sensory cortices (Hubel & Wiesel, 1962; 1968) and in particular their sufficiency as a basis set (e.g. spanning a set of dimensions that in combination can represent any stimulus in a particular stimulus space). For example, in the case of vision, a set of filters that span dimensions of spatial frequency, orientation and spatial location are mathematically sufficient to represent any image (ignoring color). This represents a core concept in our approach, that training based upon a basis

set of perceptual dimensions that are sufficient to represent the perceptual space of interest would provide a principled approach to obtain transfer of learning to the broad range of stimuli described by that space of features (Seitz, 2018).

This project tests the hypothesis that improvements in supra-threshold auditory processing, including speech in competition, will result from training with the basic perceptual features or processes from which they arise. One challenge in this endeavor is to identify the critical dimensions across tasks and stimuli that are sufficient as a basis function of central auditory processing relevant to speech in competition (Seitz, 2018). Of note, here we are focusing on primitive features that underlie the extraction of speech sounds from competing sources, and are not targeting higher level processes related to the representation of speech itself. One of the most promising sets of candidates for basis functions are spectral and temporal amplitude modulations. There is substantial evidence that these both describe response properties of neurons in the auditory cortex (Kowalsky, Depireux & Shamma, 1996; Shamma, 2001) and can computationally be used to represent any auditory stimulus within a time-spectrum space. Further, spectro-temporal processing ability has also been shown to predict speech intelligibility in individuals who have difficulties both in detecting pure-tones and in understanding speech in quiet and in noise (Bernstein et al., 2013; Mehraei, Gallun, Leek & Bernstein, 2014). Based upon this literature, a first set of

candidate dimensions for training are spectral-temporal modulation (STM) processing at a variety of frequency ranges, direction, and modulation duration.

Another potential dimension that may form an essential basis set, crucial for auditory scene analysis and for speech in competition, is the information underlying the ability to localize sounds in the environment. Spatial hearing can help segregate information coming from sound targets including speech and reduce the interference caused by distractors at different locations (Gallun, Diedesch, Kampel & Jakien, 2013). The ability to benefit from spatial hearing cues declines with increases in age and/or in pure-tone detection thresholds (Füllgrabe, Moore & Stone, 2015; Gallun, 2021) and so it represents another candidate for systematic variation.

Additionally, the ability to process sounds in memory (auditory working memory) is an important mediator of auditory learning (Zhang et al., 2016) that is essential to the recognition of speech in competition (Gallun & Jakien, 2019) as well as the listening effort associated with complex acoustical conditions (Peelle, 2018). Previous work has shown that working memory demands are effective at infusing cognitive challenge into perceptual tasks that may in turn promote learning (Bavelier, Green & Dye, 2009; Green & Bavelier, 2015).

In sum, based on neuroscientific and behavioral grounds, we selected fundamental dimensions of auditory processing that individually and collectively contribute to the ability to listen successfully to speech targets in competition.

These were presented in a gamified setting with adaptive difficulty in tasks that focused training on stimulus duration, STM slope, modulation depth, spatial offset, or auditory memory load. Training included resolving these stimuli from competing noise sources, which further allowed task difficulty to be adapted across an ecologically valid dimension (McDermott, 2009). While there have been previous studies that have used video-game elements with AT (Tallal et al., 1996; Wade & Holt, 2005; Vlahou, Protopapas & Seitz, 2012; Kimball et al. 2013), these typically trained on more limited stimulus sets. A notable exception is Whitton et al. (2017), who used a gamified approach that trained older adults with hearing loss on pitch, level, amplitude modulation and speech sounds and found learning transfer to measures of speech in competition (Whitton, Hancock, Shannon & Polley, 2017). Still, research examining how training with a wide range of psychoacoustical and cognitive tasks may lead to improvements in listening to speech in competition is limited and there is a need to examine how training on a theoretically-motivated basis set of basic auditory features may or may not lead to the broad based learning outcomes that have been seen in the case of vision (Deveau, Lovcik & Seitz, 2014).

In this study, the effectiveness of this gamified mixed-training approach was examined in a population of college-aged adults with no reported hearing difficulties. This AT training program, called *Listen* (https://braingamecenter.ucr.edu/games/listen-an-auditory-training-experience/),

was developed at the BGC, can be run on mobile devices (e.g. iPad, iPhone, Android) or standard desktop computers (MacOS, Windows) and is currently freely available through the Apple App Store, the Google Play Store, and the Microsoft Store. The AT implemented in *Listen* "gamifies" auditory perceptual tasks into an "endless runner" type of video-game in which the player makes judgements based on spectro-temporal modulations, spatialized sound cues, and previously presented sounds stored in working memory to avoid obstacles and progress within the game environment. Correct and incorrect responses have direct and immediate influence on the adaptive parameters of the game, which we hypothesize will have powerful PL consequences.

To evaluate the effectiveness of this gamified AT approach, we examined the primary outcome of transfer to speech in competition, and then secondary outcomes of transfer to measures of central auditory and cognitive processing before, in the middle (in the case of speech in competition) and after training, with one month follow-up (again only speech in competition). For hearing assessments, we used the Portable Automated Rapid Testing (PART) app (https://braingamecenter.ucr.edu/games/p-a-r-t/), which we have previously demonstrated is capable of reproducing precise acoustic stimuli outside of a controlled lab environment (Gallun et al., 2018) and have validated its performance with a group of college-aged participants with no known hearing difficulties in conditions of moderate environmental noise (Larrea-Mancera et al.,

2020). The mixed-training approach was compared to an active control condition comprised of pure-tone frequency discrimination training presented in the same gamified framework but lacking most of the elements that we believe are needed to promote transfer of learning. Results provide initial evidence that the mixed-training AT can generalize to speech in competition outcome measures beyond the active control condition. The results from this early-stage effectiveness study in individuals with no known hearing difficulties sets the ground for future research to determine the possible effectiveness of *Listen* for populations with hearing difficulties, as well as mechanistic studies to determine the extent to which the different ingredients of the mixed training contribute to the AT outcomes and further definition, expansion, and refinement of the hypothesized basis sets tested here.

METHODS

Participants

Fifty-four undergraduate students (47 female, *M* age = 20.8 years, *SD* = 3.24 years) from the University of California, Riverside were recruited for participation in the study. All participants reported having no difficulties with their hearing or vision, and no history of psychiatric or neurological disorders, and provided signed informed consent as approved by the University of California, Riverside

Human Subject Review Board. They received course credit for their participation. Because the data collection took place during the COVID-19 pandemic, testing was administered remotely in participants' homes via video calls and using their own equipment (e.g. computer or tablet and headphones). Because this was a fairly lengthy study—37 sessions—and data collection took place during the summer months when the COVID-19 infection rate was on the rise, it was challenging to recruit participants and there was significant attrition, with 21 participants leaving the study before its completion. An additional three subjects were excluded due to incomplete data sets caused by administration errors. Thus, the data presented represent the 30 remaining participants who completed all test sessions divided in two groups, the mixed training group (13 female, *M* age = 21.26 years, *SD* = 4.25 years) and the frequency discrimination control group (12 female, *M* age = 21.06 years, *SD* = 2.43 years) further described below.

Materials

Participants used the hardware that was available to them (most commonly iPhones) as well as the headphones of their choice (most commonly Apple AirPods) and were asked to use the same combination of device and headphones for the entire study. In this aspect, this is an effectiveness study of auditory training which embraces the diversity in technological systems (e.g. tablet and headphones) and environmental conditions (see Green et al., 2019) as

features that allow us to determine the extent to which the AT will be effective in ecological conditions (e.g. what could be expected from people downloading and using the training program in their individualized ecological conditions).

Minimum Audibility

All participants were able to respond to the stimuli, thus ensuring minimal audibility. Because participants were using their own equipment and testing took place remotely, we were not able to calibrate the devices and so the exact presentation levels are unknown. To address this, signal audibility was assessed in two ways that are common in the practice of audiology —2-kHz pure tone detection and single-talker sentence detection. Gallun et al. (2018) showed that the single-talker sentence detection task in PART correlates well with speech detection thresholds tested clinically. For ease of reference, levels are specified in nominal dB, which refers to the level that would have been obtained in an acoustical system consisting of a calibrated digital-to-analog converter and set of electroacoustical transducers (such as headphones) to which the same digital signal was applied. Our experience is that uncalibrated Apple iOS devices are typically within a few dB of their calibrated equivalents.

While there were several participants with surprisingly high detection thresholds on both the tone and speech audibility tasks (see Supplement Figure SB1), they were still able to perform the training task, suggesting that the high

detection thresholds represent motivational lapses, or distractions in within their testing environments (a topic of relevance for further research and approaches to control), rather than poor audibility that might occur due either to hardware incapable of producing the range of sounds needed, or to listeners incapable of detecting the sounds used in the training. For this reason, none of the participants were excluded from the study on the basis of detection thresholds.

2-kHz pure tone detection in quiet

Participants were asked to indicate if they heard a 100-ms, 2-kHz pure tone presented diotically (to both ears). Presentation level started at a nominal level of 70 dB. Following three consecutive 'yes' responses, indicating the detection of the tone presented, the presentation level of the tone decreased first by a step of 20 dB, then in steps of 10 dB until a presentation level of 10 dB was reached, at which point presentation level decreased in steps of 5 dB until a value of 0 dB was reached or three consecutive 'no' responses were recorded. The level with the last correct response made was registered as threshold. Participants were able to detect the tone at presentation levels under 30 dB on average (M = 21.16, SD = 13.9). This suggests both that the hardware used was capable of presenting soft sounds and that the listeners were generally able to detect those soft sounds.

Single-talker speech identification in quiet

Sentences from the Coordinate Response Measure corpus (CRM, Bolia et al., 2000) produced by a single talker were presented (e.g. *Ready Baron go to blue six now.*) and participants were asked to correctly identify which combinations of four possible color and eight possible number keywords they heard. Responses were made on a 4 X 8 a grid of color-number combinations. Presentation level started at a nominal level of 60 dB. The level was decreased by 5 dB after every three trials until 2 out of three responses at a given presentation level were incorrect which ended the task. Participants were able to perform under 40 dB on average (M = 36.83, SD = 8.5), again suggesting that the hardware and listeners were performing within the expected range.

Procedure

This study is considered a double-blind randomized actively-controlled study, as both research assistants and participants were blind to the fact one condition was designed as a control for the learning hypothesis behind the other condition (see Green et al., 2019). The study began with an initial enrollment in which participants completed their informed consent forms, were informed of the experimental schedule, demographic information was collected, and device and headphones type they were planning to use were noted. Participants were then randomly assigned either to the mixed (experimental) training condition or to the

frequency discrimination (control) training. After attrition, fifteen participants completed the study in the mixed-training condition and another fifteen participants finished the control condition.

Following enrollment, each participant was asked to complete a total of 38 sessions: divided in 30 training sessions and 8 assessment sessions. The assessment conditions consisted of three pre-test, one mid-test, three post-test and one follow-up session that was conducted approximately one month after training (see Figure 4.1). The three pre-test sessions were monitored via video using internet-capable video calling software. The first pre-test session consisted of an audiologic case history, the minimum audibility assessments, and the speech in competition assessments (about 30 minutes). The second pre-test session consisted of the rest of our supra-threshold hearing assessments (about 36 minutes). In the third pre-test session, participants completed the cognitive assessments (about 25 minutes) as well as the first session of training (25 minutes). The assessments will be described in detail below.

After the pre-test sessions, participants completed their first session of training with supervision and were asked short questions to assess initial expectations. After this, they were asked to complete two unsupervised training sessions per day (25 minutes each) on seven days for another 14 sessions of training. There was a lockout that ensured participants did not do more than two sessions every 24 hours and that participants delayed no more than one week so

that this first phase of training would conclude in no longer than two weeks. Then the mid-training assessments were applied and monitored via video (minimum audibility and speech in competition tests; 25 minutes). In this same session and after a short break, participants completed their 16th session of training. Following this, participants trained at their homes the recommended two sessions per day (25 minutes each) for the remaining unsupervised training sessions. After this, participants completed the post-training assessment sessions which were organized identical to the pre-training assessment sessions and were monitored via video. About a month after all the post-tests were completed, a video-monitored follow-up session was carried out; this session was identical to the mid-training assessment session and contained only minimum audibility and speech in competition assessments.

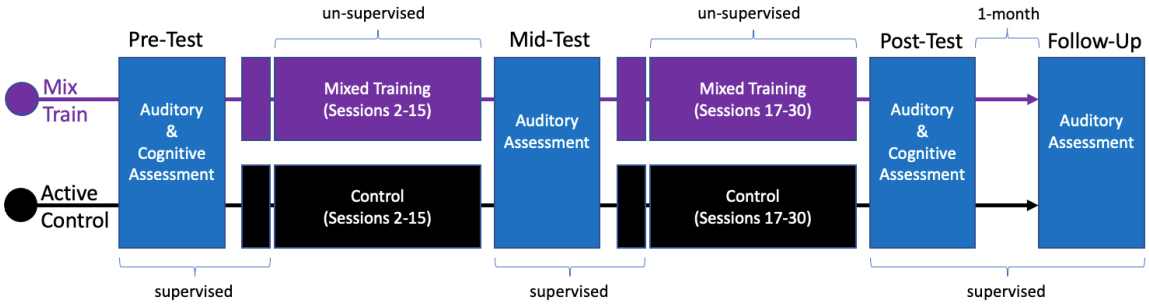

Figure 4. 1: Schematic of the procedures of each training group. Supervised assessment sessions of central auditory or cognitive processing are shown in blue. Training is shown in purple for the mixed-training and black for the active control. First and 16th session of training were also supervised. Follow-up assessments were conducted one month after the last session.

Training

In the game experience, players controlled a game avatar (the “wisp”) that stayed in the center of the screen while the landscape’s optic flow suggested movement towards it, giving the impression that the wisp was flying through the landscape (see Figure 4.2). Players were asked to help the wisp avoid obstacles or choose from among options based on a variety of sound cues. Correct responses made the wisp avoid obstacles and absorb energy from the environment, while incorrect responses made the wisp crash into obstacles and lose energy. Both the positive and negative energy effects were accompanied by auditory feedback that indicated whether participants made a correct or incorrect response. The difficulty of the task adapted along a single parameter depending on these responses. As players made progress through the game, new levels with new sounds were unlocked and difficulty related to sound processing was increased along a number of parameters associated to task types as detailed below.

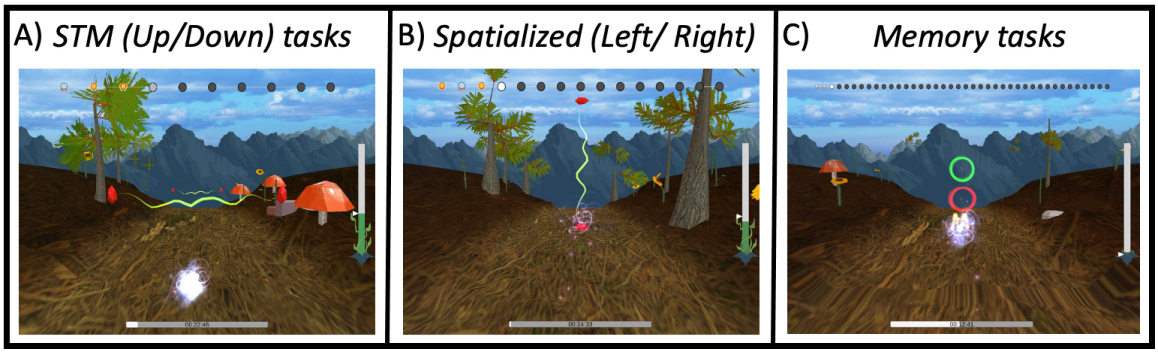

Figure 4. 2: Screenshots of the game *Listen*. The three main task categories are shown in panels left to right: the STM up/down tasks, the spatialized left/right tasks and the memory tasks.

All trials (obstacles) were presented in "*streaks*" of varying size. Within a streak, the adaptive parameter was not changed. The number of streaks was determined for each task type in the beginning of a "*run*" depending on the game's progression logic and was displayed in the upper section of the screen as nodes to be filled up with medal-like or red cross tokens depending on within streak proficiency of performance. The number of trials within each streak was equal to the number of correct responses in the prior streak plus one, with a minimum of one and a maximum of five. After each streak, if every trial within the streak received a correct response, then the adaptive parameter was stepped down by a magnitude specified for each task type below. If fewer than 75% of the trials received a correct response, then the adaptive parameter was stepped up a number of times equal to the number of errors made within the streak, otherwise the adaptive parameter remained unchanged. A status bar on the right side of the screen indicated proficiency of performance within a run (see Figure 4.2).

Once all streaks in a run were finished, the next task was selected randomly from the pool of available unlocked tasks. A timer was displayed at the bottom of the screen that indicated the time remaining in the given session. Training sessions ended once both the timer reached zero and the current run was completed.

Active Control: Frequency Discrimination Training

Frequency discrimination training contained the same visual landscape and positive and negative feedback described above and the task required participants to avoid obstacles by swiping upward or downward on the touchscreen to indicate whether a test frequency associated with the obstacle was higher or lower, respectively, than a target frequency that was presented 500 ms before the test frequency during the game at the beginning of each streak (containing anywhere from 1 to 5 trials with test frequencies). Target frequencies were centered at 250 Hz, 500 Hz, 1 kHz, 2 kHz or 3 kHz with a random rove of 15% around the center frequency to prevent sensory adaptation. The adaptive parameter of this task was the frequency ratio between target and test frequencies. As participants made progress, the frequency ratio decreased from a value of 0.5 (frequency difference is equal to half the target frequency) towards zero (no difference) with a minimum value of 0.001 and a maximum of 1. This control was designed to have all the gamified aspects but contain a stimulation that lacks the diverse approach (including testing signals in competition) that was taken in the Mixed-training condition. No specific temporal,

spectral, spectro-temporal or spatial dimension of auditory processing was targeted for this control. In consequence no specific aspect can be attributed to the differences found between-groups.

Experimental (mixed) Training

Mixed-training differed from the active control frequency discrimination task in that it contained three different task categories: up/down STM discrimination; left/right spatial discrimination, and auditory memory (see Figures 4.2 and 4.3). The task conditions, stimuli, and progression logic are described schematically in Figure 4.3. All these tasks were presented both in quiet and with competing background noise. Competing noise was either broad-band white noise or "Carlile" noise (Carlile and Corkhill, 2015), which is created by vocoding speech into 22 bands spaced on an equivalent rectangular bandwidth (ERB) scale from 50 to 16.5 kHz, and then temporally offsetting each band by rotating randomly in a circular buffer. Carlile noise thus contains the long-term spectrum and within-band amplitude modulations of speech but is unintelligible.

STM tasks

The STM up/down tasks required the participant to swipe upwards or downwards to help the wisp move up or down to avoid a horizontal obstacle, as shown in Figure 4.2. The cue provided was a narrow-band spectro-temporal modulated noise with one octave bandwidth centered around one of five different frequencies: 250 Hz, 500 Hz, 1 kHz, 2 kHz and 3 kHz with a random rove of 15%

around the center frequency. The additional acoustical details of these stimuli are as described below in the STM discrimination assessment. This category of tasks started with the *Intro* task type with a center frequency of 1 kHz. In this *Intro* task an additional frequency-modulated (FM) sweep was presented with the STM narrow-band stimuli to help the listeners learn how to move the wisp up or down in space in response to a stimulus that moved up or down in frequency. The FM sweep adapted on a sweep-to-STM level difference with a step size of 5 dB, from -20 dB where only the FM sweep is presented, to +20 dB where only the STM stimulus is presented. After completing this *Intro* task, new frequencies for the Intro task type were unlocked as well as the *Duration* task type with 1 kHz center frequency as shown in Figure 4.3.

Duration task types adapted on the duration of the STM sound with a step factor of 1.05, an initial and maximum value of 500 ms, and a minimum value of 60 ms. The temporal modulation rate scaled such that one complete temporal cycle was always completed over the duration of the stimulus. When participants reached a duration of 300 ms with their performance, the *Depth*, *Slope*, and *Noise* task types would unlock with a fixed duration of 300 ms. *Duration* tasks remained available in the pool of task types and when a performance value of 60 ms was reached for a given center frequency, a harder version of *Depth, Slope* and *Noise* tasks with 60 ms fixed stimulus duration was unlocked.

Depth task types adapted modulation depth on an exponential scale with a step factor of 1.2, from 40 dB to 0.01 dB. The *Slope* task types adapted on the percentage of a complete cycle that was completed over the duration of the stimulus and adapted using a step factor of 1.1, from 1.0 to 0.01 cycles. Finally, the STM *Noise* task types presented white noise in competition with the STM stimulus and adapted on noise-to-signal ratio with a step size of 2 dB, from -20 dB to +30 dB. At the extrema of the range (-20 dB, +30 dB), only the louder stimulus was presented.

Spatialized tasks

The Spatialized left/right tasks required the participant to swipe leftwards or rightwards on their touchscreens to help the wisp move left or right in visual space to avoid vertical obstacles in response to a stimulus that was presented to the left or right of midline in auditory space. Stimuli were 240-ms long synthetic vowels—/a/ , /ae/, /i/, and /u/—generated with a Klatt speech synthesizer implemented through Praat (Boersma and Weenick, 2016), using a 44.1 kHz sampling rate and were low-pass filtered at 5 kHz. Onset and offset ramps were 20 ms. Different task types adapted on either spatial offset or noise level. This category of tasks started with the *Offset* type where the stimulus is adapted on angular offset from center with a step factor of 1.1, starting from 60 degrees and down to 0.1 degree. Depending on the value reached in this *Offset* task, the *White Noise* and *Carlile Noise* tasks would be unlocked at a fixed offset of 60

degrees. Spatialized *Noise* task types presented noise spatialized forward adapting on the noise-to-signal level with a step size of 2 dB between -20 dB and +20 dB. At the extrema (-20 dB and +20 dB), only the louder stimulus was presented. Achieving a noise-to-signal performance level of 0 dB unlocked the next fixed offset (e.g. 45 degrees) for the *Noise* tasks if that offset had already been unlocked from the *Offset* task.

Memory tasks

The Memory tasks required the player to swipe upwards or downwards to help the wisp choose between the two rings presented instead of obstacles (shown in the far right panel of Figure 4.2). This task did not use the streaking mechanism and each trial was evaluated individually in the staircase. When the rings appeared, the player was required to compare the sound just heard with one stored in memory. In the “1-back” condition, the wisp needed to fly through the top green circle if the last sound matched the one before it, or through the bottom red circle if the sound did not match the one before. In the “2-back” condition, the comparison was to the sound that had been played two before it, representing a greater memory load. If there was not a match, the player was to direct the wisp through the bottom red ring. The sounds to be memorized were distributed in three task types: *Pure Tone* using sinusoidal tones, *Voice Intro* using synthetic vowels in quiet, and *Voice + Noise* which used vowels in competition with white noise. Progression occurred from simpler sounds towards more complex and the

"2-back" conditions were only unlocked for each task type after a 90% accuracy of performance was reached for the "1-back" conditions. Once the *Voice + Noise* task "2-back" condition was achieved, this would be the only memory task available for training.

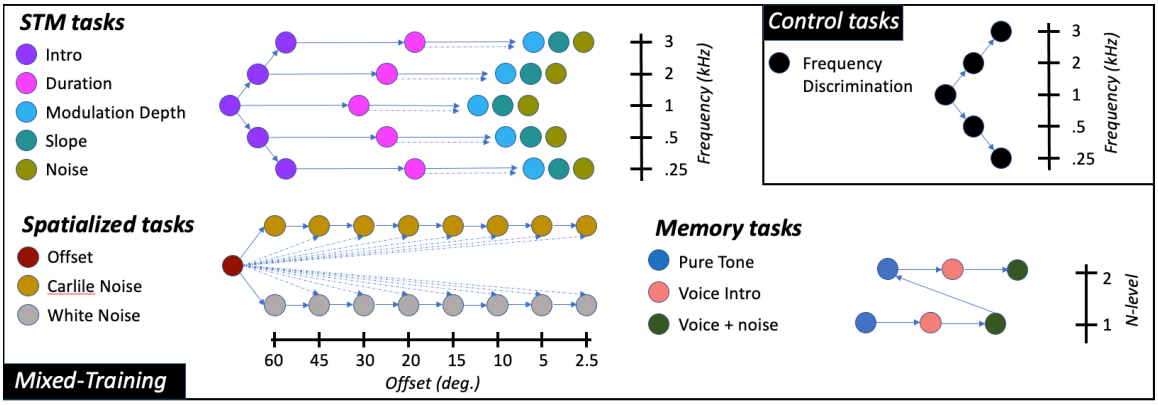

Figure 4. 3: Schematic of the tasks and progression for the mixed-training and active control. Different task types are presented in different colors and are grouped in three categories (e.g. left/right). Solid arrows show progression based on some level of performance. Dotted arrows indicate additional conditional relations. Each of the different task types adapts on a single perceptual parameter (usually name of task). Up/down category tasks are further divided in five target center frequencies (so is the control). Left/right category, noise type tasks are further divided in fixed offset-from-center versions. Memory tasks are further divided depending on memory load. The control condition is shown in the top right panel, this tasks adapts separately on each tone frequency.

Assessments

All participants completed the same assessments before, in the middle, and after training. Assessments were carried out remotely using applications developed at the BGC: PART for the auditory perception tasks and Recollect (https://braingamecenter.ucr.edu/games/recollect/) for the cognitive processing measures, also available online. The assessments were organized into three groups: speech in competition assessments (primary outcome), basic auditory tests of supra-threshold hearing, and cognitive assessments. The speech in competition assessments included tests of spatial release from masking and identification of spoken digits in noise, and were carried out at pre-, mid-, post-training, and follow-up time points. The basic supra-threshold auditory tests included dichotic FM, gaps-in-noise, and spectro-temporal modulation detection and discrimination tests. These basic supra-threshold tests were assessed only at the pre- and post-training time points. The cognitive assessments included spatial working memory, working memory updating, countermanding, and cancellation tests, and were also only applied at pre- and post-training time-points. This design reflects our interest on the speech in competition measures as primary outcome measures with the other measures considered to be secondary/exploratory outcomes.

Assessments of Speech in Competition

Spatial Release from Masking

Identification of speech targets in the presence of two competing speech maskers was measured using a method developed by Marrone, Mason & Kidd (2008) and modified by Gallun et al. (2013). Two conditions were tested: one in which all three talkers are presented with the same interaural differences ("*colocated*") and one in which the target appears to be located in front of the listener and the maskers are located to the left and right of center with an offset of 45 degrees ("*separated*"). All spatial locations were simulated over headphones by convolving the speech stimuli with the appropriate head-related impulse responses for each location as described in Gallun et al. (2013). Target level was fixed at a nominal level of 65 dB and the level of each masker was progressively increased after every two responses, starting at a target-to-masker ratio (TMR) of 10 dB and progressing over 20 trials to a TMR of -8dB. Threshold is estimated based on the total number of correct responses as described in Gallun et al. (2013). Speech stimuli were taken from the same CRM corpus that was used for the speech audibility pre-test, described above. On each trial, as in the pre-test, participants identified color/number combinations uttered along with the call-sign "*Charlie*" by one of three male speakers. In this case, however, the target was presented in competition with two other male speakers uttering different color/number/call-sign combinations from the CRM sentences. The

color/number combination was identified by clicking on a color/number grid presented on screen. The dB difference between TMR thresholds in the colocated and separated conditions is used as a Spatial Release from Masking metric and reflects the ability to benefit from spatial cues.

Digits in Noise Identification

The targets in this task were digit triplets spoken in competition with white noise (Smits, Goverts & Festen, 2013) and presented in a 25-trial 1-up/1-down adaptive staircase where the presentation level of the target decreased by 2 dB following a correct response and increased by 2 dB following an incorrect response. Both target and noise started at a nominal level of 70 dB, and the noise level was held constant for all trials.

Basic Supra-threshold Auditory Assessments

These tasks employed a 4-interval, 2-cue 2-alternative forced-choice format as described in Larrea-Mancera et al. (2020) where four squares were presented on screen and lit up sequentially in coordination with four auditory intervals. The first and last intervals always presented *standard* stimuli (thus referred to as *cue* intervals), in contrast to the two alternatives in the middle intervals, one of which would match the cues and the other would contain the *target* of interest. The target would differ from the standards based on a single parameter, which would be adaptively varied based on performance. Adaptive tracking involved two-stage adaptive staircases with a 2-down/1-up rule, meaning that two correct responses

would make the task harder and one incorrect response would make it become easier. The first stage of the staircases contained three reversals and had step sizes five times larger than in the second stage, which contained six reversals. Thresholds were calculated from the average of the second stage reversals. Step sizes of the staircases were kept at a ratio of 1:1.5, which indicates that the step up was 1.5 the size of the step down. Further details of the staircase parameters are given for each task below.

Dichotic FM Detection

The stimuli were those used by Larrea-Mancera et al. (2020), based on the dichotic FM detection task developed by Green, Heffer & Ross (1976) and modified by Grose & Mamo (2010, 2012) and Hoover et al., (2019). Standard intervals contained pure tones with a frequency drawn at random from the range 460-540 Hz. Each was 400 ms in duration and was presented at a nominal level of 75 dB. Onset and offset ramps were 20 ms. Target intervals contained tones drawn from the same frequency range, the same level and the same duration as the standard intervals but had an anti-phasic 2-Hz frequency modulation across left and right ears (dichotic). The target interval adapted on modulation range (Hz) on an exponential scale starting at 10 Hz and stepping down by $2^{1/2}$ Hz for the first stage and $2^{1/10}$ Hz for the second with a minimum value of 0 and a maximum of 10 kHz.

Gaps-in-Noise Detection

This assessment involved the use of a noise stimulus upon which are imposed brief silent gaps, the detection of which requires temporal processing of envelope and temporal fine structure cues (see Grose, Eddins & Hall, 1989; Florentine, Buus, & Geng, 1999; Hoover, Pasquesi & Souza, 2015; Hoover et al., 2019). Standard intervals were 400-ms long white noise presented at a nominal level of 70 dB. Onset and offset ramps were 20 ms. Targets were the same noise stimuli into which a brief silent gap had been introduced. Across trials, gap duration (ms) was adaptively varied on an exponential scale starting at 20 ms and stepping down towards zero by $2^{1/2}$ ms for the first stage and $2^{1/10}$ ms for the second with a maximum value of 60 ms.

Spectro-Temporal Modulation Detection

The STM stimuli used were from Larrea-Mancera et al. (2020). Standard intervals were 300 ms white noise from 400 Hz to 8 kHz presented at a nominal level of 70 dB. Onset and offset ramps were 20 ms. For the detection task (labeled simply STM), targets contained a spectral modulation of 2 cycles per octave and a temporal modulation rate of 4 Hz. Thresholds were measured in terms of modulation depth (dB) which was adaptively varied using a logarithmic amplitude scale measured from the middle to the peak of the amplitude range as described in Stavropoulos et al. (2021) as M (expressed in dB). Adaptation

started at 6 dB and stepped down by 0.5 dB for the first stage and 0.1 dB for the second with a minimum value of 0 and a maximum of 10 dB.

Spectro-Temporal Modulation Discrimination

For the STM discrimination tasks (labeled STM_250 and STM_3k), STM stimuli were presented in all four intervals, but the direction of modulation for one of the stimuli in the second and third intervals matched the modulation direction (up or down) in the first and fourth intervals (the standard “cues”), while the other did not. To make the task more difficult, a narrowband noise (1 octave wide) was also presented on each interval. In one task, the targets and maskers were centered at 250 Hz and in a second task, all were centered at 3 kHz. Performance was measured by adaptively varying the modulation depth, starting at 10 dB and stepping down by $2^{1/2}$, every three trials until 4 or more errors were made in the last 6 trials.

Assessments of Cognitive Processing

The cognitive assessments were selected to represent measures of general domain cognitive processes thought to be related to perception and include measures of working memory, attention, and inhibition.

Spatial Working Memory (Corsi blocks)

This task, originally developed by Corsi (1972), relies on accurate sequential storage and retrieval of sequences in working memory. An array of squares

(drawn to represent gopher holes) is distributed asymmetrically in space and presented to subjects. In this modified version, for every trial, gophers come out one at a time from holes already present on the screen (traditionally squares are pointed to or change color in computer versions). Gophers are visible for 1.5 seconds (0.25 seconds rising from the hole, 1 second waiting above the hole, 0.25 seconds descending into the hole) in a random sequence with inter-stimulus-intervals (ISIs) of 0.5 seconds. Participants had to identify the holes where the gophers were presented in the order in which they had appeared. Participants had 10 seconds to respond. After every response, the next trial started after an inter-trial-interval (ITI) of 1 second.

Every time a sequence of holes was identified correctly, the number of elements in the sequence increased, starting with two-element sequences and progressing towards a maximum of ten-element sequences. When an incorrect response occurred, the number of elements in the sequence would not change. If a second incorrect response occurred, the number of elements decreased by two but never went below two. The second time two incorrect responses were provided in a row, the test would end. Span scores were computed the longest sequence achieved with at least one correct response.

Working Memory Updating (n-back).

Similar to the memory task used in the mixed-training, we used an n-back task (Kircher, 1958; see Pergher et al., 2020) in which participants were required to

report what they saw (rather than what they heard) n-items back in a continuous, sequential presentation divided in 5 blocks. On each trial participants had 2500 ms to respond if the presented animal cartoon (e.g. sheep) matched (or not) the animal presented “n” (load) trials back with an inter-stimulus-interval (ISI) of 500 ms. We first presented 29+n trials of the 1-back, then we presented a block of 9+n practice trials of the 2-back followed by a block of 29+n trials of the 2-back, then a block of 9+n practice trials of the 3-back followed by a block of 29+n trials of the 3-back. Accuracy was calculated for each of the n-levels from the number of hits divided by the sum of hits, misses and false alarms. Correct rejections did not contribute to accuracy scores.

Countermanding

This task provides a measure of cognition additional to those of working memory related to inhibition. It is based on Wright & Diamond (2014) but uses dogs and monkeys instead of hearts and flowers for congruent and incongruent stimuli. On each trial, two buttons were presented on the sides of the screen. Atop one of them, one of two stimuli was presented. A picture of a dog required the participant to press the button on the side of the screen with the picture. A picture of a monkey required the participant to press the button on the other side. Participants were instructed to respond as fast as they could. The key process is that participants need to inhibit one stimulus-response relation to act on the other. After a short introduction of three trials, a dogs-only condition was tested

for 12 trials. Then monkeys were introduced for three trials and tested for 12 trials. After this, a mixed condition with dogs and monkeys is introduced for three trials and then tested for 48 trials. Reaction times constitute the main outcome measure of this test.

Cancellation

This is a test of selective and sustained attention that resembles the D2 test (Brickenkamp & Zilmer, 1998) where participants are instructed to sequentially search and mark a set of target items in a series of similar items. In our variant called UCancellation, participants were presented sequentially with visual targets in the form of dogs and monkeys that varied in their orientation (facing right or left) and color distribution (same color palette). Participants had to select a target type of dog/monkey among distractors with similar features and colors. Eight pictures were displayed per row, with 3-5 targets per row; every 10 rows had exactly 40 targets. Each row was displayed for a maximum time of 6 seconds (with 1 second blank screen interval between rows). One auditory cue signaled that time had run out for a particular row and a different auditory cue was presented if the participant cancelled all targets in a row with no false alarms. Participants completed a short practice run of about 10 trials and then were tested for 3 minutes and 30 seconds. Scores were computed out of the number of hits minus the number of false alarms.

Data Analysis

Data analyses were organized around two main questions: 1) Was there an improvement in the outcome measures collected within the groups from the Pre-Test to the Post-Test?, 2) Are any improvements found greater in the experimental group compared to the active control? For the first question we conducted related-samples t-tests between pre- and post-test scores within each group. For the second question we conducted independent-samples t-tests on the difference between pre- and post-test scores (Pre – Post) of each group, which is equivalent to the interaction term of a mixed model 2-by-2 ANOVA. Based upon the *a priori* hypothesis that the mixed training would lead to greater positive changes on speech in competition tasks than the active control, one-tailed tests were conducted for the speech outcome measures. Given that there are multiple measures of the same constructs (as recommended by Simons et al 2016), and to minimize multiple comparison issues, we computed composite scores based on the following groupings: *Basic Auditory Composite:* gap in noise, dichotic FM, and the STM; *Speech in Competition Composite:* spatial release from masking tasks (colocated and separated conditions) and the digits in noise task; and the *Cognitive Composite:* spatial working memory, working memory updating, countermanding and cancellation. Composites were calculated from the average of z-scores associated to the relevant groupings of

assessments. Both pre- and post-tests were included in the standarization of each assessment.

RESULTS

The results are presented in three sections: training data (Section A), auditory perceptual outcomes (Section B), and cognitive outcomes (Section C).

Training Data

Because the training was designed to give participants experience across a range of hearing dimensions, it is difficult to compute a simple measure that captures overall performance. However, one way to understand training performance is to examine the extent to which participants progressed across the task matrix presented in Figure 4.3. Further results are described in the Supplemental Materials section SA. All individual runs for all tasks used in both training conditions are shown in figures SA1 to SA10.

All participants in the mixed-training group made substantial progress through the game's different levels. In the case of Spatialized (left/right) tasks, all participants made progress in terms of the offset from center where targets were presented from the highest magnitude of 60 to below 2.5 degrees (see Fig. SA2 in the supplement). Likewise, in the case of the Memory tasks, all participants

progressed to the final 2-back task achieving average SNR thresholds of approximately -4 dB. In the case of the STM discrimination (up/down) tasks, all participants progressed out of the intro layer of tasks and unlocked the duration adaptive layer across all five center frequencies tested. Only two-thirds of the participants unlocked STM discrimination tasks that adapted on either depth, slope or in terms of competition with noise. This implies more aspects of training were potentially still to be leveraged towards training gains in some participants.

In the control training, which involved fewer conditions, participants quickly unlocked all tasks with the five center frequencies tested (see Fig. SA1 in the supplement) and achieved average thresholds that were less than .05 of the center frequencies tested.

Auditory Perceptual Outcomes

Table 4.1 shows mean pre- and post-training performance on assessments for both groups. At baseline, mean performances of the experimental and control groups were similar (within half a standard deviation) to thresholds previously reported for remote testing in a similar sample (Larrea-Mancera et al., 2021) in the dichotic FM assessment (M = 0.82, SD = 2.48), the STM assessment (M = 1.24, SD = 0.61), and the speech-on-speech masking tasks in the colocated (M = 2.89, SD = 1.58) and separated conditions (M = -1.81, SD = 3.68) as well as in the spatial release from masking (SRM) metric (M = 4.43, SD = 3.38). These

data suggest that participants overall performance on auditory tasks was within what would be expected based on previously obtained norms. Table 4.1 also presents the comparisons of training-related change within each group for exploratory purposes only as the main analysis is based on composite scores. Table 4.2 presents the comparisons between the change (difference) scores obtained in each group also for exploratory purposes only.

Assessment	Mean (SD) Pre-Test	Mean (SD) Post-Test	Units	Within-subjects t (df)	P (two-tail)	Cohen D	BF_{10}
GIN (control)	2.53 (0.88)	2.64 (1.37)	ms	-0.49 (14)	0.62	-0.09	0.19
(mixed train)	2.4 (0.51)	3.32 (3.89)		-0.91 (14)	0.37	-0.33	0.15
DichoticFM (control)	0.86 (0.91)	0.51 (0.56)	Hz	1.4 (14)	0.18	0.45	1.07
(mixed train)	0.79 (0.58)	0.68 (0.58)		0.78 (14)	0.44	0.19	0.52
STM detection (control)	1.73 (0.98)	1.44 (0.41)	M	1.39 (14)	0.18	0.38	1.04
(mixed train)	1.23 (0.26)	1.37 (0.67)	(dB)	-0.92 (14)	0.37	-0.28	0.15
STM_250 (control)	6.17 (3.01)	4.47 (3.16)	M	1.79 (14)	0.09	0.55	1.78
(mixed train)	5.24 (2.69)	4.23 (3.16)	(dB)	1.24 (14)	0.23	0.34	0.88
STM_3k (control)	3.5 (3.58)	3.93 (3.44)	M	-0.37 (14)	0.71	-0.12	0.204
(mixed train)	5.53 (3.61)	3.36 (3)	(dB)	2.15 (14)	0.049*	0.65	3.002
Colocated (control)	2.4 (1.76)	2.2 (1.32)	TMR	0.43 (14)	0.76	0.12	0.214
(mixed train)	2.86 (2.44)	1.6 (1.72)	(dB)	2.47 (14)	0.09	0.59	1.78
Separated (control)	-1.86 (3.99)	-2.26 (4.06)	TMR	0.38 (14)	0.29	0.09	0.73
(mixed train)	0.13 (4.03)	-2.73 (3.71)	(dB)	2.86 (14)	0.009*	0.73	12.01
Spatial R. (control)	4.26 (3.59)	4.46 (3.64)	dB	0.18 (14)	0.2	0.05	0.12
(mixed train)	2.73 (3.32)	4.33 (3.88)		1.66 (14)	0.026*	0.44	0.09
Digits (control)	-21.23 (1.86)	-20.88 (1.97)	TMR	-1.06 (14)	0.64	-0.18	0.19
(mixed train)	-20.83 (3.49)	-22.3 (1.52)	(dB)	2.34 (14)	0.14	0.54	1.305

Table 4. 1: Within-group summary statistics. For the auditory assessments addressing within-group training-related change. Related-samples t-tests (frequentist and Bayesian) are also provided.

Assessment	Mean Difference Scores (SD)	Units	Between-subjects t (df)	P (two-tail)	Cohen D	BF_{10}
GIN (control) (mixed train)	-0.1 (0.85) -0.92 (3.89)	ms	0.78 (28)	0.43	0.28	0.43
DichoticFM (control) (mixed train)	0.34 (0.96) 0.11 (0.57)	Hz	0.8 (28)	0.42	0.29	0.44
STM detection (control) (mixed train)	0.28 (0.8) -0.14 (0.6)	M (dB)	1.66 (28)	0.106	0.609	0.96
STM_250 (control) (mixed train)	1.7 (4.47) 1.01 (3.9)	M (dB)	0.55 (28)	0.58	0.201	0.38
STM_3k (control) (mixed train)	-0.43 (3.58) 2.16 (3.61)	M (dB)	-1.69 (28)	0.1007	-0.62	1.001
Colocated (control) (mixed train)	-0.2 (2.59) 1.2 (2.59)	TMR (dB)	-1.47 (28)	0.15	-0.53	0.77
Separated (control) (mixed train)	1.2 (4.26) 4.06 (5.2)	TMR (dB)	-1.65 (28)	0.109	-0.603	0.94
Spatial R. (control) (mixed train)	1.4 (4.03) 2.83 (4.48)	dB	-0.94 (28)	0.35	-0.34	0.48
Digits (control) (mixed train)	-0.26 (2.2) 1.36 (3.39)	TMR (dB)	-1.56 (28)	0.12	-0.18	0.85

Table 4. 2: Between-group summary statistics. For the auditory assessments addressing between-group training-related change using difference scores (pre – post). Independent-samples t-tests (frequentist and Bayesian) are also provided.

To address the primary training outcome, namely the effectiveness of the AT approach to promote transfer to improved performance on measures of speech in competition, we examined changes across time on speech in competition composite score (see Figure 4.4) that consisted of the colocated and separated measures from the spatial release from masking tasks and the digits in noise measure (individual task statistics are shown in Table 4.1). This composite had a strong internal reliability at pre-test across both groups (*Cronbach's alpha* = 0.79) which indicated this composite is suitable to represent the assessments it contains. For this measure we observed a significant improvement for the mixed-training group ($t_{(14)}$ = 2.61, p = 0.01, *Cohen's d* = 1.19) but not for the control group ($t_{(14)}$ = 0.05, p = 0.47, *Cohen's d* = 0.01). Importantly, there was also a

significant difference in the change scores between groups ($t_{(28)}$ = -1.91, p = 0.033, *Cohen's d* = -0.68), showing that the improvement in speech in competition composite was significantly greater than that of the control group. These results provide preliminary evidence that the mixed-training may provide benefits to tasks of speech in competition, however we note the small sample size and that the effect would not pass a two-tailed test, and so it will be important to replicate these results.

To explore whether the AT led to changes in other supra-threshold hearing assessments, we examined a basic auditory processing composite (see Figure 4.4). This composite also had strong internal reliability at pre-test across both groups (*Cronbach's alpha* = 0.73) which indicated this composite is also suitable to represent the assessments it contains. For this measure, we observed no statistically significant changes in either the mixed-training group ($t_{(14)}$ = 0.44, p = 0.66, *Cohen's d* = 0.11), nor the control group ($t_{(14)}$ = 1.57, p = 0.15, *Cohen's d* = 0.39). Further, an independent samples t-tests on these difference scores (mixed-training vs control) revealed no statistically significant differences in the basic auditory composite ($t_{(28)}$ = 0.63, p = 0.54, *Cohen's d* = 0.22), between the training groups.

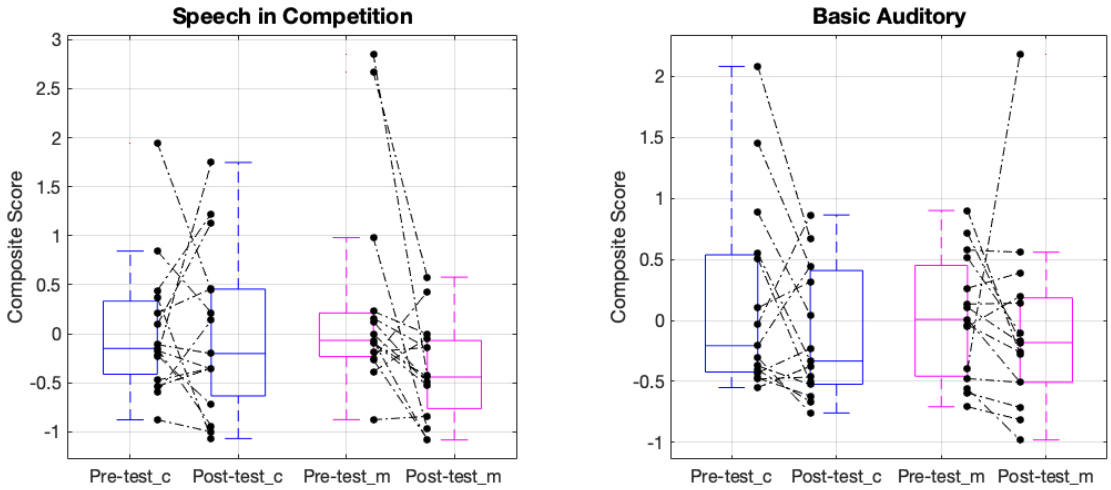

Figure 4. 4: Auditory outcomes. Data from pre- and post- composite measures of hearing. Blue boxes show Control group (_c) data and magenta boxes the mixed-training group (_m). Black dots indicate individual thresholds and dotted lines the individual trajectory of performance change (pre to post).

Dosage and retention effects

To address how much training was required to achieve the observed improvement on the speech in noise tests, we examined data in the mid-test. First, addressing the issue of dosage, we observed an improvement on the speech in competition composite when comparing the pre-test to the mid-test ($t_{(28)}$ = -2.47, p = 0.01, *Cohen's d* = -0.88). Next, we examined whether learning was retained after an interval of one month without training. We did not find statistical evidence in support of a benefit from pre-test to follow-up ($t_{(28)}$ = -0.96, p = 0.17, *Cohen's d* = -0.34). The difference found in thresholds between pre-test and mid-test in the mixed-training group appears to be no different than that of

pre-test to post-test ($t_{(14)}$ = -0.25, p = 0.8, *Cohen's d* = -0.06), suggesting that 15 sessions is a sufficient dose of training, however data from the follow-up fails to show that effects remain the same across time, at least for normally hearing young adults used in the present study (see Figure 4.5).

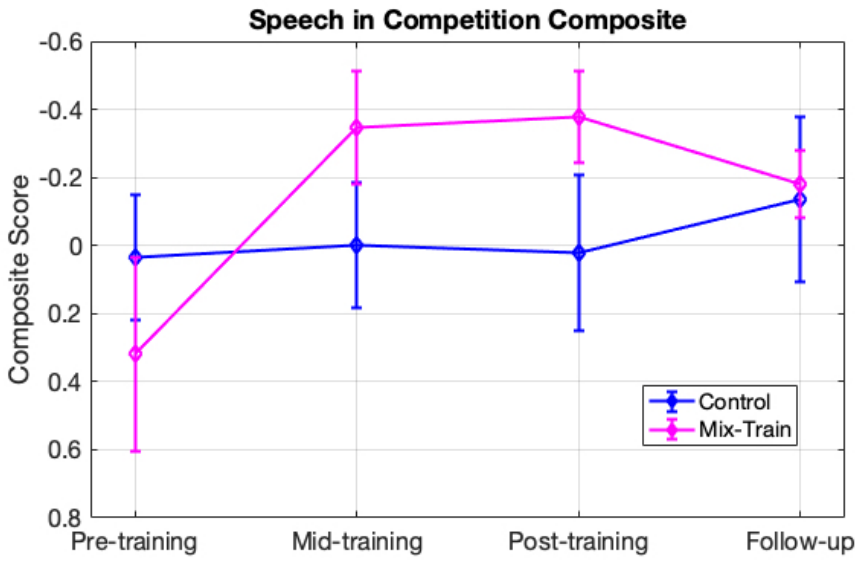

Figure 4. 5: Dosage and retention effects. Shows the average thresholds for the speech in competition composite before, during and after training including a one month follow-up. Error bars represent standard error of the mean.

Cognitive Outcomes

The cognitive composite had only a moderate internal reliability at pre-test across both groups (*Cronbach's alpha* = 0.41) which indicated this composite is probably not the best way to represent the assessments it contains. After this analysis, it was clear that different tasks either explain different aspects of the

variance, or perhaps some of them were unreliable, thus spreading noise through the rest of the measures, and so the cognitive composite was not used to evaluate training outcomes. Instead, each assessment was examined separately (see Figure 4.6). For the countermanding test, a conflict score was computed by subtracting the average reaction time for responding to the dogs from the average reaction time for responding to the monkeys. This metric providence no evidence of change in the mixed-training group ($t_{(14)}$ = 1.21, p = 0.48, *Cohen's d* = -0.306), or in the control group ($t_{(14)}$ = -1.63, p = 0.24, *Cohen's d* = -0.41). For the spatial working memory span, we did not find significant change in either the control ($t_{(14)}$ = -0.79, p = 0.42, *Cohen's d* = -0.201) or the mixed-training ($t_{(14)}$ = -0.89, p = 0.76, *Cohen's d* = -0.22). For working memory updating, performance accuracy on the 1-back was at ceiling and the 3-back at chance performance for most participants, and so we chose to focus on the 2-back. We found accuracy improved significantly for the mixed-training group ($t_{(14)}$ = -3.74, p < 0.01, *Cohen's d* = -0.94) but not for the control group ($t_{(14)}$ = -1.96, p = 0.069, *Cohen's d* = -0.49). However, this change from pre to post-test did not differ significantly between the mixed-training and control groups ($t_{(28)}$ = 1.15, p = 0.25, *Cohen's d* = 0.42). Finally, for the cancellation task we found that neither the mixed-training group showed significant within-group change in scores ($t_{(14)}$ = -1.82, p = 0.089, *Cohen's d* = -0.45), nor the control group ($t_{(14)}$ = -0.55, p = 0.58, *Cohen's d* = -

0.13). Thus overall, there is little evidence of a reliable change in cognitive measures from this training above the control condition.

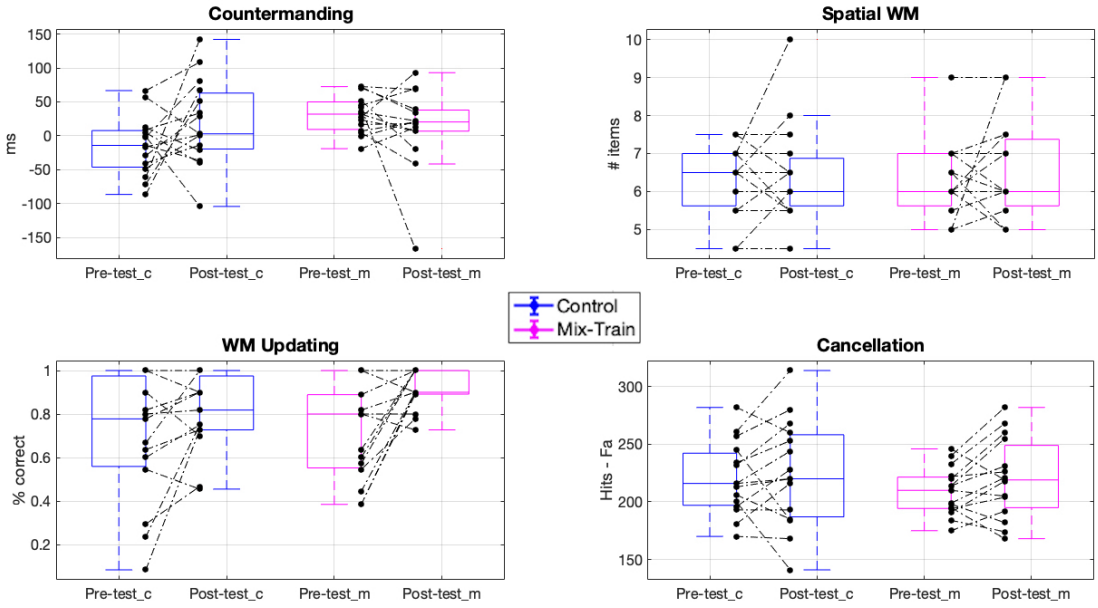

Figure 4. 6: Cognitive outcomes. Data from pre- and post- measures of cognitive processing. Blue boxes show Control group (_c) data and magenta boxes the mixed-training group (_m). Black dots indicate individual thresholds and dotted lines the individual trajectory of performance change (pre to post).

DISCUSSION

This study investigated the effectiveness of a novel gamified approach to Auditory Training (AT) based on neural and cognitive research on speech in competition. Significant improvements were found in the speech in competition tasks relative to those found for an active frequency-discrimination control

training. However, no consistent changes were observed in measures of more basic supra-threshold auditory processes. While at first look this may be surprising, it is worth noting that these tasks primary involve detection, rather than the discrimination tasks used in training, and with detection thresholds being superior to discrimination thresholds, the stimulus values in the tests were largely outside of the range presented during the training task, with the exception of the STM discrimination (250 Hz and 3 kHz) tasks, where training improvements were significant or close to significant. Moreover, we did not find significant differences between mixed-training and control groups in terms of the learning effects of AT on the cognitive measures. These results were found in participants who downloaded the software on their own devices and conducted experimental sessions in their own space, suggesting that the results obtained here are similar to what one would expect from a young normal-hearing individual of similar demographics accessing the training tool on their own outside of controlled laboratory settings. We acknowledge the preliminary nature of these results with a small sample size, and plan replication and extensions to other age groups including those different types of hearing loss once human subject research restrictions related to the COVID-19 pandemic are relaxed.

A key question in the literature has been the extent to which expectations may explain effects of cognitive and perceptual training. To address this, we asked participants to report, after their first experience with the auditory training,

their expectations regarding whether the auditory training would lead to improvements either on the trained conditions or in other tasks of their daily life using a Likert type scale (1 = Not at all; 2 = Not really; 3 = Can't say; 4 = Quite a bit; 5 = Very much). Participants neither exhibited strong expectations of improvements on the trained skills (mixed-training M = 3.8 SD = 0.86; active control M = 3.5 SD = 0.91), or to untrained activities of daily living (mixed-training M = 3.5 SD = 0.74; active control M = 3 SD = 0.75), and there were no statistical differences between the groups (p = 0.41 for near transfer and p = 0.061 for far transfer) although there was a trend for higher expectation for the transfer to tasks of daily life in the mixed training condition. However, there were no significant correlations between expectations and training outcomes on the speech in noise composite (trained skills, r = 0.1, p = 0.6; daily life, r = -0.064, p = 0.73). Thus, we failed to find solid evidence that expectations explained training outcomes of the study, or differences in outcomes between groups.

The effect sizes for some of the speech assessments conducted here are comparable to that of Whitton, Hancock, Shannon & Polley (2017), which has been heralded as a viable type of AT intervention (see Skoe, 2017), with reported benefits of about 1.5 dB signal-to-noise ratio in a group of people with hearing difficulties. After 15 sessions of training our participants, all of whom reported no hearing difficulties, achieved improvements of a similar size, which did not change with the rest of the training. In our training we found mean differences

between pre- and post-training assessments in the speech in competition measures that differ between the mixed training group and the active control by 1.4 dB for the colocated SRM, 2.86 dB for the separated condition and 1.62 dB for the digits in noise test (see Table 4.2).

It is important to note that the effects observed here were not of a size that reached statistical significance when tested one month after training. This lack of retention leads to the question of whether additional training, or maintenance sessions (e.g. top-up sessions that are shorter and less frequent than full training), could have allowed them to retain these observed benefits. Clarifying the extent to which maintenance training will lead to retention will be a target of our future research. Of note, there is also a question of whether retention may depend on age as in previous studies (e.g. Merzenich et al, 1996; Tallal et al., 1996; Moore, Rosenberg & Coleman, 2005) children seemed to retain training for longer periods. Another important future direction will be to test effectiveness of the approach in people of different age groups and with hearing difficulties.

The benefits observed are consistent with Stewart et al. (2020), who suggested that using an action-based video-game that targets auditory cues for its task resolution should yield benefits in the auditory domain. Those authors observed no significant effects after training with an action video-game and suggested this may be due to sensory domain specificity (mainly relying on visuo-spatial cues). Interestingly, we found that the mixed training condition

showed significant improvement after training in the working memory updating task (n-back). This benefit was not statistically significant when compared to the active control condition which also showed a tendency for improvement. These results might reflect expected effects from active gamified tasks on WM processes (Deveau et al., 2015) that are thought to mediate auditory processing (Zhang et al., 2016; Zhang et al., 2017). It is an interesting question of whether WM updating, or attention switching (Dhamani, Leung, Carlile & Sharma, 2013), are particularly susceptible to training and that they then could underpin improved speech in competition (Gallun & Jakien, 2019). While our findings support the idea that perceptual learning as a result of a gamified AT may transfer to speech in competition measures, we note that, given the complexity of the mixed-training approach (e.g. training multiple stimuli, tasks, and with a complex motivational framework), more research will be required to understand which game elements are of importance to this effect, and how training elements may interact, to promote beneficial change throughout the many brain processes that may be involved in this learning (Maniglia and Seitz, 2017). Possible elements of importance include the motivated engagement characteristic of play behavior (Vygotsky, 1967), the direction of exogenous and endogenous attention (Donovan, Szpiro, & Carrasco, 2015; Donovan & Carrasco, 2018), the promotion of cognitively challenging “fast activity” (Green & Bavelier, 2015; Bediou et al., 2018), the use of varied stimulus sets (Deveau, Lovcik & Seitz, 2014; Xiao et al.,

2008; Zhang et al., 2011), adaptive difficulty ensuring a match of skill and challenge (Ahissar & Hochstein, 1997; Hung & Seitz, 2014), multisensory facilitation of learning (Shams and Seitz, 2008), and the sensorimotor nature of tasks that include a diverse exploration of sensory and motor contingencies (O'Reagan & Noë, 2001; Whitton et al., 2014; 2017). While the distinct elements mentioned here may have specific contributions to perceptual learning and transfer, and there is a need to better understand these contributions to gain mechanistic understanding and improve training design, it is likely they all converge in promoting the learning effects observed to some extent (Seitz & Dinse, 2007), although we cannot rule out some interference (Katz et al., 2014).

There are a number of indications in the literature that our training approach can be improved to boost learning. For example, Whitton et al. (2014; 2017) identified the sensori-motor co-generating element in their "foraging" task, which involved searching for targets with manual movements, as being crucial to promote the effects they have found. Likewise, other studies examining music to promote learning have emphasized this synchronous co-generation of motor behavior and perceptual information (see Zatorre, Chen & Penhune, 2007). As our training is an interactive video-game thus already including a series of sensorimotor relations, there is still an opportunity to couple our trained sounds into a co-generative relationship with some of the motor responses they evoke.

This co-generative relationship between sensory and motor processes is typical for example in musical instruments.

Moreover, some have suggested that having a rich multi-sensory training approach might be beneficial to promote learning (Shams & Seitz, 2008) even when the target is unisensory (Shams, Wozny, Kim & Seitz, 2011), as it may benefit from interactions with other sense modalities with different proficiencies (Barakat, Seitz & Shams, 2015). Although our training is audio-visual and thus already addresses some of the possible multi-sensory benefit, extensions can be made to integrate additional multisensory cues with visual stimuli that are congruent with the auditory stimuli and can facilitate the auditory stimuli (Seitz, Kim and Seitz 2006; Shams and Seitz, 2008). Future research should explore additional correspondences between visual and auditory cues (see Yehia, Kuratate & Vatikiotis-Bateson, 2002) and even other senses (see Rosenblum, Dias & Dorsi, 2017). Providing multi-sensory simultaneous co-variation as it typically ocurrs with perceptual objects in the world is a way to exploit the above mentioned correspondances.

A third aspect which could be explored to boost learning is the use of implicit rather than, or in addition to, explicit training. Prior research suggests that implicitly training phonemic categories using temporal synchrony with task-relevant aspects in video-game play may lead to benefits to speech processing (Wade & Holt, 2005; Vlahou, Protopapas & Seitz, 2012; Kimball et al., 2013).

Exploring training both under the explicit focus of attention and implicit temporal coupling to task relevant elements outside the focus of attention (e.g. Seitz & Watanabe, 2003; Seitz, Kim & Watanabe, 2009) may afford more diverse training benefits as different learning mechanisms might be recruited (Seitz & Dinse, 2007; Seitz & Watanabe, 2009).

Notably, given that supra-threshold hearing difficulties differ across individuals, it is likely that more attributes of the training intervention could be personalized to the individual. Our training is designed in such a way that tasks that are difficult for a given individual will remain in the training rotation until the processing precision required by the game to progress to different tasks or difficulties is achieved. In that sense the training is, to some extent, tailored to individual needs, but could still be individualized further. For example, while the frequency-discrimination task is a reasonable control condition for young normally hearing adults, in the case of cochlear implant patients frequency discrimination training directly targets their hearing needs (e.g. Goldsworthy & Shannon, 2014). Thus, for this population there would be important dimensions of hearing to consider (e.g. pure tone discrimination) that might be different than for a population with age-related changes in hearing or for those suffering the effects of traumatic brain injury. Future research with hearing diverse groups of people and across the lifespan is required to further understand what elements of our AT approach may be more important to promote supra-threshold hearing

benefits including improvements understanding speech in competition and how this may differ as a function of different individuals' hearing and listening needs.

Beyond exploring the effectiveness of our AT approach, which represents the main motivation of this study, another matter of interest is of a methodological nature: the extent to which the performance for the different aspects of supra-threshold hearing present in the gamified training match the validated assessments obtained with PART. However, the thresholds obtained during training with similar stimuli to that used for the STM discrimination assessments were of higher magnitude on average (8.23 dB for the 250 Hz and 10.19 dB for the 3 kHz) than the assessment thresholds (see Table 4.1). Further, there was no relation between the assessment thresholds and the training thresholds for either the 3 kHz center frequency (r = -0.001, p = 0.9) or the 250 Hz center frequency (r = -0.63, p = 0.09). Of note, only 9 out of 15 participants in the mixed-training group reached the training task that was equivalent to assessment, making the apparent distance in thresholds even greater. Future work will be required to account for differences in performance between the gamified and non-gamified settings and also which setting may better predict hearing in ecological settings. While the non-gamified testing environment provides a nicely controlled testing environment, the game represents some of the variability of tasks and sounds that are found in ecological settings.

In summary, this study presents a proof of concept that an integral approach to AT that focuses on a basis set of spectral-temporal modulations, sound localization with and without competition, and working memory components can transfer to untrained tasks of speech in competition. Our study demonstrates the feasibility of this dynamical and entertaining game environment to train hearing to be used in participants' homes and on uncalibrated devices, greatly improving the accessibility and thus potential impact of the approach. Moreover, this study and intervention presents a starting point from which to improve development of auditory training in search for a more optimal learning paradigm. However, a small sample was collected and participants in this study had no reported hearing difficulties. Thus, future research both for replication and extensions to address the extent to which this intervention may provide benefits to people with diverse hearing abilities, and across different age groups that better represent those seeking improvements in their hearing abilities.

REFERENCES

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. https://doi.org/10.1038/387401a0

Anderson, S., White-Schwoch, T., Parbery-Clark, A., et al. (2013a). Reversal of age-related neural timing delays with training. *Proc Natl Acad Sci*, *110*, 4357–4362. https://doi.org/10.1073/pnas.1213555110

Anderson, S., White-Schwoch, T., Choi, H. J., et al. (2013b). Training changes processing of speech cues in older adults with hearing loss. *Front Syst Neurosci*, *7*, 97. https://doi.org/10.3389/fnsys.2013.00097

Barakat, B., Seitz, A. R., & Shams, L. (2015). Visual rhythm perception improves through auditory but not visual training. *Current Biology*, *25*(2), R60–R61. http://dx.doi.org/10.1016/j.cub.2014.12.011

Bavelier, D., Green, C. S., & Dye, M. W. G. (2009). Exercising your brain: Training-related brain plasticity. In M. S. Gazzaniga, E. Bizzi, L. M. Chalupa, S. T. Grafton, T. F. Heatherton, C. Koch, J. E. LeDoux, S. J. Luck, G. R. Mangan, J. A. Movshon, H. Neville, E. A. Phelps, P. Rakic, D. L. Schacter, M. Sur, & B. A. Wandell (Eds.), *The cognitive neurosciences* (p. 153–164). Massachusetts Institute of Technology, USA.

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*(1), 77–110. https://doi.org/10.1037/bul0000130

Bernstein, J. G., G. Mehraei, S. Shamma, F. J. Gallun, S. M. Theodoroff & M. R. Leek (2013). Spectrotemporal modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners. *J Am Acad Audiol, 24*(4), 293-306. https://doi.org/10.3766/jaaa.24.4.5

Boersma, P., & Weenink, D. (2014). Praat: Doing Phonetics by Computer. Version 5.3.84. http://www.praat.org/

Bolia, R. S., Nelson, W. T., Ericson, M. A. and Simpson, B. D. (2000). “A speech corpus for multitalker communications research,” *Journal of the Acoustical Society of America, 107*(2), 1065-1066. https://doi.org/10.1121/1.428288

Brickenkamp, R. & Zillmer, E. (1998). *The d2 Test of Attention*. Seattle, Washington: Hogrefe & Huber Publishers, USA.

Burk, M.H., Humes, L.E., Amos, N.E., & Strauser, L.E. (2006). Effect of training on word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners. *Ear Hear*, *27*, 263–278. https://doi.org/10.1097/01.aud.0000215980.21158.a2

Carlile, S., & Corkhill, C. (2015). Selective spatial attention modulates bottom-up informational masking of speech. *Scientific Reports*, *5*, 1–7. https://doi.org/10.1038/srep08662

Chermak, G. D., & Musiek, F. E. (2002). Auditory training: Principles and approaches for remediating and managing auditory processing disorders. *Seminars in Hearing*, *23*(4), 297–308. https://doi.org/10.1055/s-2002-35878

Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979.

Chisolm T H, Johnson C E, Danhauer J L. et al. (2007). A systematic review of health-related quality of life and hearing aids: final report of the American Academy of Audiology Task Force on the Health-Related Quality of Life Benefits of Amplification in Adults. *J Am Acad Audiol*, *18*(2), 151–183.

Corsi, P.M. (1972). *Human memory and the medial temporal region of the brain*. Doctoral Thesis at McGill University (Canada).

Deveau, J., Lovcik, G., & Seitz, A. R. (2014). Broad-based visual benefits from training with an integrated perceptual-learning video game. *Vision Research*, *99*, 134–140. https://doi.org/10.1016/j.visres.2013.12.015

Deveau, J., & Seitz, A. R. (2014). Applying perceptual learning to achieve practical changes in vision. *Frontiers in Psychology*, *5*(October), 1–6. https://doi.org/10.3389/fpsyg.2014.01166

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology*, *24*(4), R146–R147. https://doi.org/10.1016/j.cub.2014.01.004

Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2015). How to build better memory training games. *Frontiers in Systems Neuroscience*, *8*(January), 1–7. https://doi.org/10.3389/fnsys.2014.00243

Dhamani, I., Leung, J., Carlile, S., & Sharma, M. (2013). Switch attention to listen. *Scientific Reports*, *3*, 1–8. https://doi.org/10.1038/srep01297

Donovan, I., Szpiro, S., & Carrasco, M. (2015). Exogenous attention facilitates location transfer of perceptual learning. *Journal of Vision*, *15*(10), 11. https://doi.org/10.1167/15.10.11

Donovan, I., & Carrasco, M. (2018). Endogenous spatial attention during perceptual learning facilitates location transfer. *Journal of Vision*, *18*(11), 7. https://doi.org/10.1167/18.11.7

Ferguson, M., & Henshaw, H. (2015). How Does Auditory Training Work? Joined-Up Thinking and Listening. *Seminars in Hearing*, *36*(4), 237–249. https://doi.org/10.1055/s-0035-1564456

Ferguson M A, Henshaw H, Clark D P, Moore D R. (2014). Benefits of phoneme discrimination training in a randomized controlled trial of 50- to 74-year-olds with mild hearing loss. *Ear Hear*. *35*(4), 110-121. https://doi.org/10.1097/AUD.0000000000000020

Florentine, M., Buus, S., & Geng, W. (1999). Psychometric functions for gap detection in a yes-no procedure*. J. Acoust. Soc. Am*. *106*, 3512–3520. https://doi.org/10.1121/1.428204

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, *7*(JAN), 1–25. https://doi.org/10.3389/fnagi.2014.00347

Gallun, F. J., A. C. Diedesch, S. D. Kampel & K. M. Jakien (2013). "Independent impacts of age and hearing loss on spatial release in a complex

auditory environment." *Front Neurosci* 7: 252. https://doi.org/10.3389/fnins.2013.00252

Gallun, F. J., McMillan, G. P., Molis, M. R., Kampel, S. D., Dann, S. M., & Konrad-Martin, D. L. (2014). Relating age and hearing loss to monaural, bilateral, and binaural temporal sensitivity. *Frontiers in Neuroscience*, *8*(8 JUN), 1–14. https://doi.org/10.3389/fnins.2014.00172

Gallun, F. J., Seitz, A., Eddins, D. A., Molis, M. R., Stavropoulos, T., Jakien, K. M., Kampel, S. D., Diedesch, A. C., Hoover, E. C., Bell, K., Souza, P. E., Sherman, M., Calandruccio, L., Xue, G., Taleb, N., Sebena, R., & Srinivasan, N. (2018). "Development and validation of Portable Automated Rapid Testing (PART) measures for auditory research," *Proc. Mtgs. Acoust. 33*(175), 050002. https://doi.org/10.1121/2.0000878

Gallun, F.J., Jakien, K.M. (2019) "The ability to allocate attentional resources to a memory task predicts speech-on-speech masking for older listeners." *Proceedings of the 23rd International Congress on Acoustics : Integrating 4th EAA Euroregio 2019* : 9-13 September 2019 in Aachen, Germany / Proceedings editors: Martin Ochmann, Michael Vorländer, Janina Fels

Gallun, F. J. (2021). Impaired Binaural Hearing in Adults: A Selected Review of the Literature. *Frontiers in Neuroscience*, *15*(March), 1–22. https://doi.org/10.3389/fnins.2021.610957

Ghose, G. M., Yang, T., & Maunsell, J. H. R. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *Journal of Neurophysiology*, *87*(4), 1867–1888. https://doi.org/10.1152/jn.00690.2001

Goldsworthy, R. L., & Shannon, R. V. (2014). Training improves cochlear implant rate discrimination on a psychophysical task. *The Journal of the Acoustical Society of America*, *135*(1), 334–341. https://doi.org/10.1121/1.4835735

Green, G. G., Heffer, J. S., and Ross, D. A. (1976). "The detectability of apparent source movement effected by interaural phase modulation [proceedings]," J. Physiol, *260*(2), 49P–50P.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–537. https://doi.org/10.1038/nature01647

Green, C. S., & Bavelier, D. (2015). Action video game training for cognitive enhancement. *Current Opinion in Behavioral Sciences*, *4*, 103–108. https://doi.org/10.1016/j.cobeha.2015.04.012

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., Bingel, U., Chein, J. M., Colzato, L. S., Edwards, J. D., Facoetti, A., Gazzaley, A., Gathercole, S. E., Ghisletta, P., Gori, S., Granic, I., Hillman, C. H., Hommel, B., Jaeggi, S. M., et al. (2019). Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement. *Journal of Cognitive Enhancement*, *3*(1), 2–29. https://doi.org/10.1007/s41465-018-0115-y

Grose J.H., Eddins D.A., Hall J.W. (1989). Gap detection as a function of stimulus bandwidth with fixed high- frequency cutoff in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*; 86:1747– 1755. [PubMed: 2808923]

Grose, J. H., & Mamo, S. K. (2010). "Processing of temporal fine structure as a function of age," *Ear and Hearing, 31*, 755-760. https://doi.org/10.1097/AUD.0b013e3181e627e7

Grose, J. H., and Mamo, S. K. (2012). Frequency modulation detection as a measure of temporal processing: Age-related monaural and binaural effects. *Hear. Res, 294(*1-2), 49–54. https://doi.org/10.1016/j.heares.2012.09.007

Henshaw, H., & Ferguson, M. A. (2013). Efficacy of Individual Computer-Based Auditory Training for People with Hearing Loss: A Systematic Review of the Evidence. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0062836

Hoover, E. C., Pasquesi, L., & Souza, P. (2015). Comparison of Clinical and Traditional Gap Detection Tests. *Journal of the American Academy of Audiology*, *26*(6), 540–546. https://doi.org/10.3766/jaaa.14088

Hoover, E. C., Souza, P. E., & Gallun, F. J. (2017). Auditory and cognitive factors associated with speech-in-noise complaints following mild traumatic brain injury. *Journal of the American Academy of Audiology*, *28*(4), 325–339. https://doi.org/10.3766/jaaa.16051

Hoover, E. C., Kinney, B. N., Bell, K. L., Gallun, F. J., and Eddins, D. A. (2019). “A comparison of behavioral methods for indexing the auditory

processing of temporal fine structure cues," J. Speech, Lang. Hear. Res. 62(6), 2018–2034. https://doi.org/10.1044/2019_JSLHR-H-18-0217

Hubel, D. H. & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154.

Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243.

Humes, L. E., Ahlstrom, J. B., Bratt, G. W., et al. (2009). Studies of hearing-aid outcome measures in older adults: A comparison of technologies and an examination of individual differences. *Semin Hear*, 30, 112–128. https://doi.org/10.1055/s-0029-1215439

Humes, L. E., Kinney, D. L., Brown, S. E., et al. (2014). The effects of dosage and duration of auditory training for older adults with hearing impairment. *J Acoust Soc Am*, 136, EL224. https://doi.org/10.1121/1.4890663

Hung, S. C., & Seitz, A. R. (2014). Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *Journal of Neuroscience*, *34*(25), 8423–8431. https://doi.org/10.1523/JNEUROSCI.0745-14.2014

Karawani, H., Bitan, T., Attias, J., & Banai, K. (2016). Auditory Perceptual Learning in Adults with and without Age-Related Hearing Loss. *Frontiers in Psychology*, *6*(February), 1–14. https://doi.org/10.3389/fpsyg.2015.02066

Katz, B., S. Jaeggi, M. Buschkuehl, A. Stegman, & P. Shah. (2014). Differential effect of motivational features on training improvements in school-based cognitive training. *Front Hum Neurosci*, *8*, 242. https://doi.org/10.3389/fnhum.2014.00242

Kimball, G., Cano, R., Feng, J., Hampson, E., Li, E., Christel, M. G., Holt, L. L., Lim, S. J., Liu, R., & Lehet, M. (2013). Supporting research into sound and speech learning through a configurable computer game. *IEEE Consumer Electronics Society's International Games Innovations Conference, IGIC*, 110–113. https://doi.org/10.1109/IGIC.2013.6659172

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Koerner, T.K., Papesh, M.A., & Gallun, F.J. (2020) A Questionnaire Survey of Current Rehabilitation Practices for Adults with Normal Hearing Sensitivity Who Experience Auditory Difficulties, *American Journal of Audiology*, *29*(4), 738-761. https://doi.org/10.1044/2020_AJA-20-00027

Kowalski, N., D.A. Depireux, & Shamma, S.A. (1996), Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *Journal of Neurophysiology*, *76*(5), 3503-3523. https://doi.org/10.1152/jn.1996.76.5.3503

Kuchinsky S E, Ahlstrom J B, Cute S L, Humes L E, Dubno J R, Eckert M A. (2014) Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology*, *51*(10), 1046–1057. https://doi.org/10.1111/psyp.12242

Lavie, L., Attias, J., Karni, A. (2013). Semi-structured listening experience (listening training) in hearing aid fitting: Influence on dichotic listening. *Am J Audiol*, *22*, 347–350. https://doi.org/10.1044/1059-0889(2013/12-0083)

Larrea-Mancera, E. S. L., Stavropoulos, T., Hoover, E., Eddins, D., Gallun, F., & Seitz, A. (2020). Portable Automated Rapid Testing (PART) for auditory research: Validation in a normal hearing population. *Journal of the Acoustical Society of America*, *148*(4), 1831–1851. https://doi.org/10.1121/10.0002108

Larrea-Mancera, E. S. L., Stavropoulos, T., Carrillo, A. A., Cheung, S., Eddins, D. A., Molis, M. R., Gallun, F., Seitz, A. R. (2021, March 25). Portable Automated Rapid Testing (PART) of auditory processing abilities in young normally- hearing listeners: A remotely administered replication with participant-owned devices. *PsyArXiv.* https://psyarxiv.com/9u68p/

Maniglia, M., & Seitz, A. R. (2018). Towards a whole brain model of Perceptual Learning. *Current Opinion in Behavioral Sciences*, *20*, 47–55. https://doi.org/10.1016/j.cobeha.2017.10.004

Marrone, N., Mason, C. R., & Kidd, G. (2008). “The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in

reverberant rooms," *J. Acoust. Soc. Am*, *124*, 3064–3075. https://doi.org/10.1121/1.2980441

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, *19*(22), 1024–1027. https://doi.org/10.1016/j.cub.2009.09.005

Mehraei, G., F.J. Gallun, M.R. Leek, & Bernstein, J.G. (2014). Spectrotemporal modulation sensitivity for hearing-impaired listeners: Dependence on carrier center frequency and the relationship to speech intelligibility. *J Acoust Soc Am*, *136*(1): p. 301. https://doi.org/10.1121/1.4881918

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal Processing Deficits of Language-Learning Impaired Children Ameliorated by Training. *Science*, *271*(5245), 77–81. https://doi.org/10.1126/science.271.5245.77

Mohammed, S., Flores, L., Deveau, J., Cohen Hoffing, R., Phung, C., M. Parlett, C., Sheehan, E., Lee, D., Au, J., Buschkuehl, M., Zordan, V., Jaeggi, S. M., & R. Seitz, A. (2017). The Benefits and Challenges of Implementing Motivational Features to Boost Cognitive Training Outcome. *Journal of Cognitive Enhancement*, *1*(4), 491–507. https://doi.org/10.1007/s41465-017-0047-y

Moore, D. R., Rosenberg, J. F., & Coleman, J. S. (2005). Discrimination training of phonemic contrasts enhances phonological processing in mainstream school children. *Brain and Language*, *94*(1), 72–85. https://doi.org/10.1016/j.bandl.2004.11.009

Moore, D., & Amitay, S. (2007). Auditory Training: Rules and Applications. *Seminars in Hearing*, *28*(2), 099–109. https://doi.org/10.1055/s-2007-973436

O'Regan, J., & Noë, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, *129*(1), 79–103. https://doi.org/10.1023/A:1012699224677

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2020). Divergent Research Methods Limit Understanding of

Working Memory Training. *Journal of Cognitive Enhancement*, *4*(1), 100–120. https://doi.org/10.1007/s41465-019-00134-7

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2017). The supramodal brain: implications for auditory perception. *Journal of Cognitive Psychology*, *29*(1), 65–87. https://doi.org/10.1080/20445911.2016.1181691

Schellenberg, E. G. (2016). Music Training and Nonmusical Abilities. *The Oxford Handbook of Music Psychology*, August 2017, 415–428.

Seitz, A. R., & Watanabe, T. (2003). Is subliminal learning really passive. *Nature*, *422*(6927), 36. https://doi.org/10.1038/422036a

Seitz, A. R., & Dinse, H. R. (2007). A common framework for perceptual learning. *Current Opinion in Neurobiology*, *17*(2), 148–153. https://doi.org/10.1016/j.conb.2007.02.004

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards Evoke Learning of Unconsciously Processed Visual Stimuli in Adult Humans. *Neuron*, *61*(5), 700–707. https://doi.org/10.1016/j.neuron.2009.01.016

Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research*, *49*(21), 2604–2610. https://doi.org/10.1016/j.visres.2009.08.003

Seitz, A.R., Protopapas, A., Tsushima, Y., Vlahou, E.L., Gori, S., Grossberg, S. & Watanabe, T. (2010) Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition, 115*, 435-443. https://doi.org/10.1016/j.cognition.2010.03.004

Seitz, A. R. (2017). Perceptual learning. *Current Biology*, *27*(13), R631–R636. https://doi.org/10.1016/j.cub.2017.05.053

Seitz, A. R. (2018). A New Framework of Design and Continuous Evaluation to Improve Brain Training. *Journal of Cognitive Enhancement*, *2*(1), 78–87. https://doi.org/10.1007/s41465-017-0058-8

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends Cogn Sci*, *5*(8): p. 340-348. https://doi.org/10.1016/S1364-6613(00)01704-6

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Shams, L., Wozny, D. R., Kim, R., & Seitz, A. R. (2011). Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology*, *2*(October), 264. https://doi.org/10.3389/fpsyg.2011.00264

Simons DJ, Boot WR, Charness N, et al. (2016). Do "Brain-Training" Programs Work? *Psychological Science in the Public Interest*, 17(3):103-186. https://doi.org/10.1177/1529100616661983

Skoe, E. (2017). Hearing: The Future of Sensory Rehabilitation? *Current Biology*, *27*(21), R1163–R1165. https://doi.org/10.1016/j.cub.2017.09.053

Smits C., Goverts T., & Festen J.M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *Journal of the Acoustical Society of America*, *133*(3), 1693–1706. https://doi.org/10.1121/1.4789933

Stewart, H. J., Martinez, J. L., Perdew, A., Green, C. S., & Moore, D. R. (2020). Auditory cognition and perception of action video game players. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-71235-z.

Stavropoulos, T. A., Isarangura, S., Hoover, E. C., Eddins, D. A., Seitz, A. R., & Gallun, F. J. (2021). Exponential spectro-temporal modulation generation. *The Journal of the Acoustical Society of America*, *149*(3), 1434–1443. https://doi.org/10.1121/10.0003604

Stropahl, M., Besser, J., & Launer, S. (2020). Auditory Training Supports Auditory Rehabilitation: A State-of-the-Art Review. *Ear & Hearing*, *41*(4), 697–704. https://doi.org/10.1097/AUD.0000000000000806

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M., & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, *271*(5245), 81–84. https://doi.org/10.1126/science.271.5245.81

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*(2), 363–381. https://doi.org/10.1037/a0025014

Vygotsky, Lev S. (1967). Play and Its Role in the Mental Development of the Child. *Soviet Psychology*, *5*, 6–18. https://doi.org/10.2753/RPO1061-040505036

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, *118*(4), 2618. https://doi.org/10.1121/1.2011156

Weihing, J., Chermak, G. D., & Musiek, F. E. (2015). Auditory Training for Central Auditory Processing Disorder. *Seminars in Hearing*, *36*(4), 199–215. https://doi.org/10.1055/s-0035-1564458

Whitton, J. P., Hancock, K. E., & Polley, D. B. (2014). Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(25), E2606-15. https://doi.org/10.1073/pnas.1322184111

Whitton, J. P., Hancock, K. E., Shannon, J. M., & Polley, D. B. (2017). Audiomotor Perceptual Training Enhances Speech Intelligibility in Background Noise. *Current Biology*, *27*(21), 3237-3247. https://doi.org/10.1016/j.cub.2017.09.014

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, *5*(MAR), 1–9. https://doi.org/10.3389/fpsyg.2014.00213

Xiao,L. Q., Zhang,J.Y.,Wang,R.,Klein,S.A.,Levi, D.M.,&Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology*, *18*(24), 1922–1926. https://doi.org/10.1016/j.cub.2008.10.030

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*(3), 555–568. https://doi.org/10.1006/jpho.2002.0165

Zendel, B. R., West, G., Belleville, S., & Peretz, I. (2017). Music training improves the ability to understand speech-in-noise in older adults. *Neurobiology of Aging,* 81, 102–115. https://doi.org/10.1016/j.neurobiolaging.2019.05.015

Zhang, J.-Y., Wang, R., Klein, S., Levi, D., & Yu, C. (2011). Perceptual learning transfers to untrained retinal locations after double training: A piggyback effect. *Journal of Vision*, *11*(11), 1026–1026. https://doi.org/10.1167/11.11.1026

Zhang, Y. X., Moore, D. R., Guiraud, J., Molloy, K., Yan, T. T., & Amitay, S. (2016). Auditory discrimination learning: Role of working memory. *PLoS ONE*, *11*(1), 1–18. https://doi.org/10.1371/journal.pone.0147320

Zhang, Y.-X., Tang, D.-L., Moore, D. R. & Amitay, S. (2017). Supramodal enhancement of auditory perceptual and cognitive learning by video game playing. *Front. Psychol*. 8, 1086. https://doi.org/10.3389/fpsyg.2017.01086

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nature Reviews. Neuroscience*, *8*(7), 547–558. https://doi.org/10.1038/nrn2152

APPENDIX II: Chapter 4 Supplemental Materials

Section A. Adaptive tracking data during training

The following figures show in color individual performance across either trial or training day for the different tasks used for training. In general, it can be observed that participants are progressing towards harder adaptive parameters as training advances. We report first the Control condition (frequency discrimination training; Fig. SA1) followed by the Experimental condition (Mixed training) including its Spatialized (Figs SA2-SA4), Spectrotemporal discrimination (Figs SA5-SA9), and memory tasks (Fig SA10). The training thresholds reported in the main manuscript were extracted from the last 25 trials of the adaptive tracks shown here.

Frequency Discrimination Control

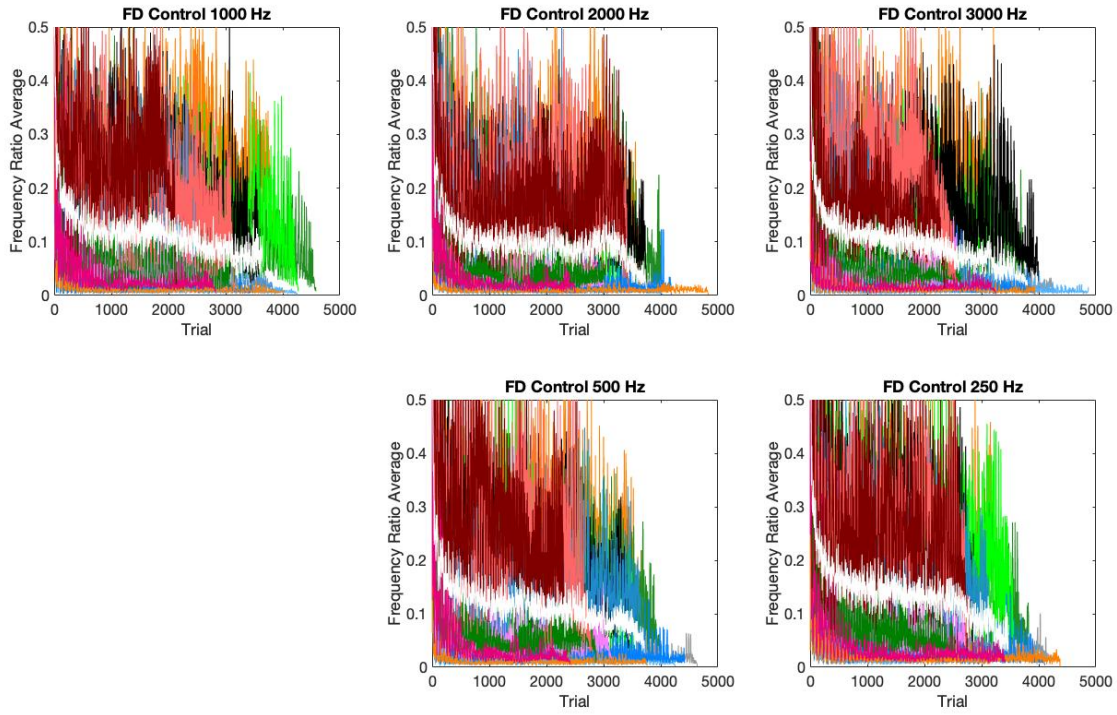

Figure SA 1: Training progression. Shows individual progression across a specified adaptive task parameter in different colors and mean performance is shown in white.

Mixed Training

Spatialized tasks

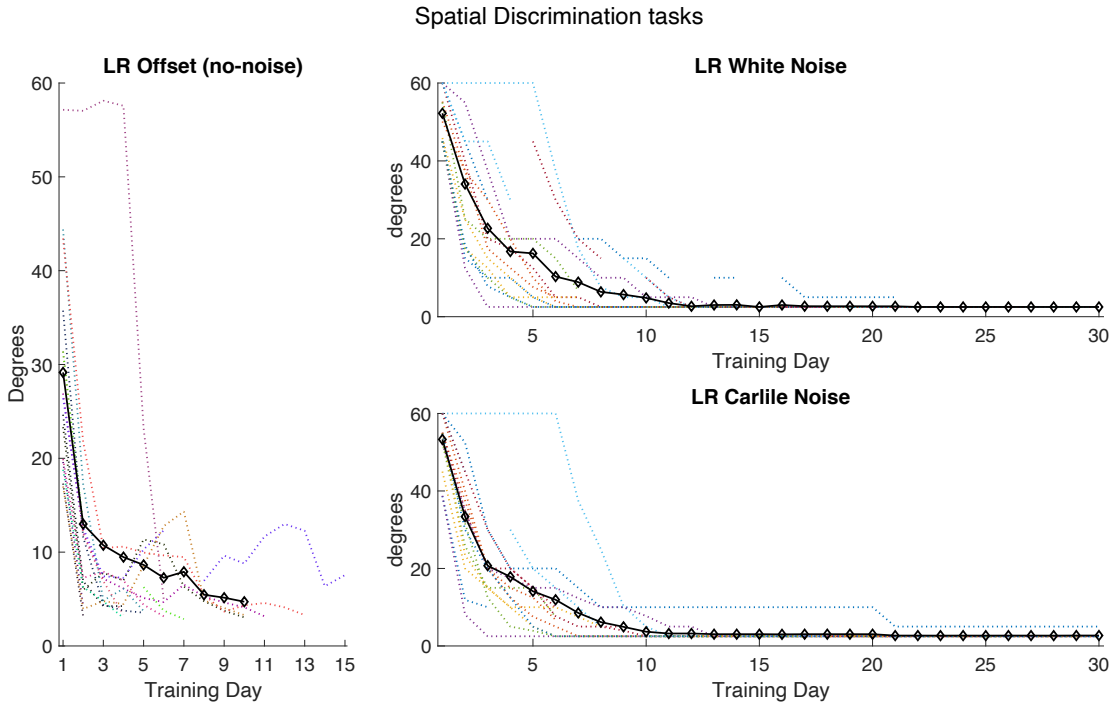

Figure SA 2: Spatialized task progression. Shows all three tasks used for left/right (LR) discrimination. Participants would initiate in a condition without noise (left panel) until they were able to perform the task at each separation magnitude (in degrees) between left and right spatialized sound. This would unlock that specific separation magnitude in the noise tasks (right panels) where noise became the adaptive parameter. Once participants were able to perform a given separation magnitude at the highest noise level, it would be considered complete, and locked out from training until only the smallest separation condition (2.5 degrees) was left. Colored dotted lines indicate individual performance and the bold black line the mean performance.

Spatialized Carlile Noise Tasks

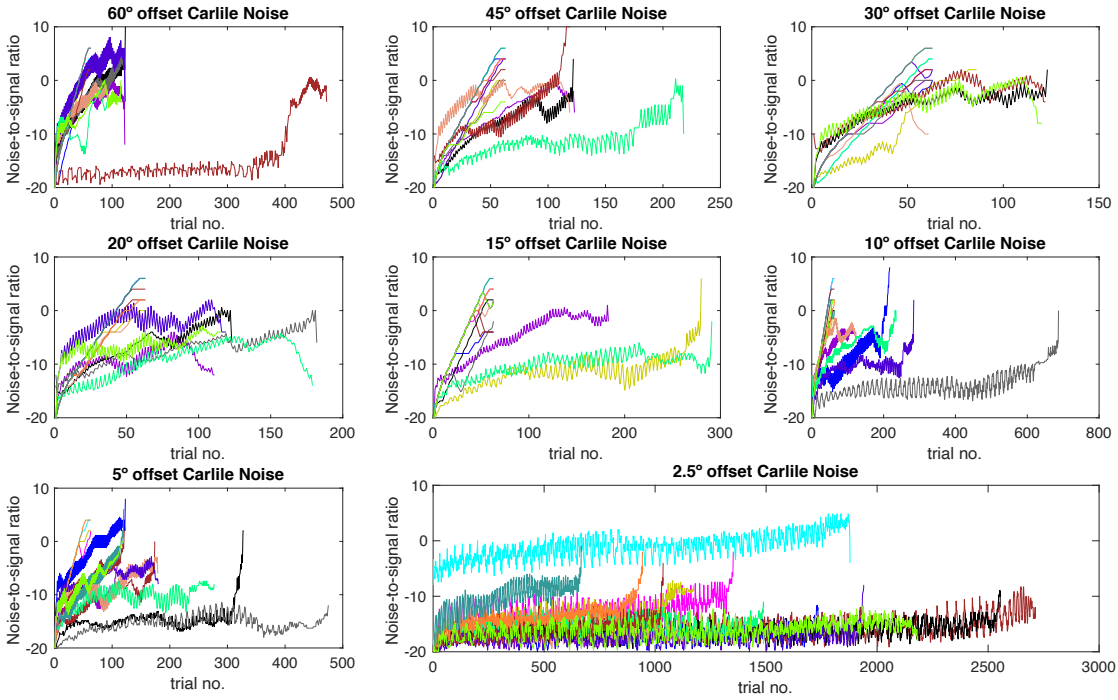

Figure SA 3: Spatialized Carlile noise task progression. Shows individual progression across noise levels relative to target in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

Spatialized White Noise Tasks

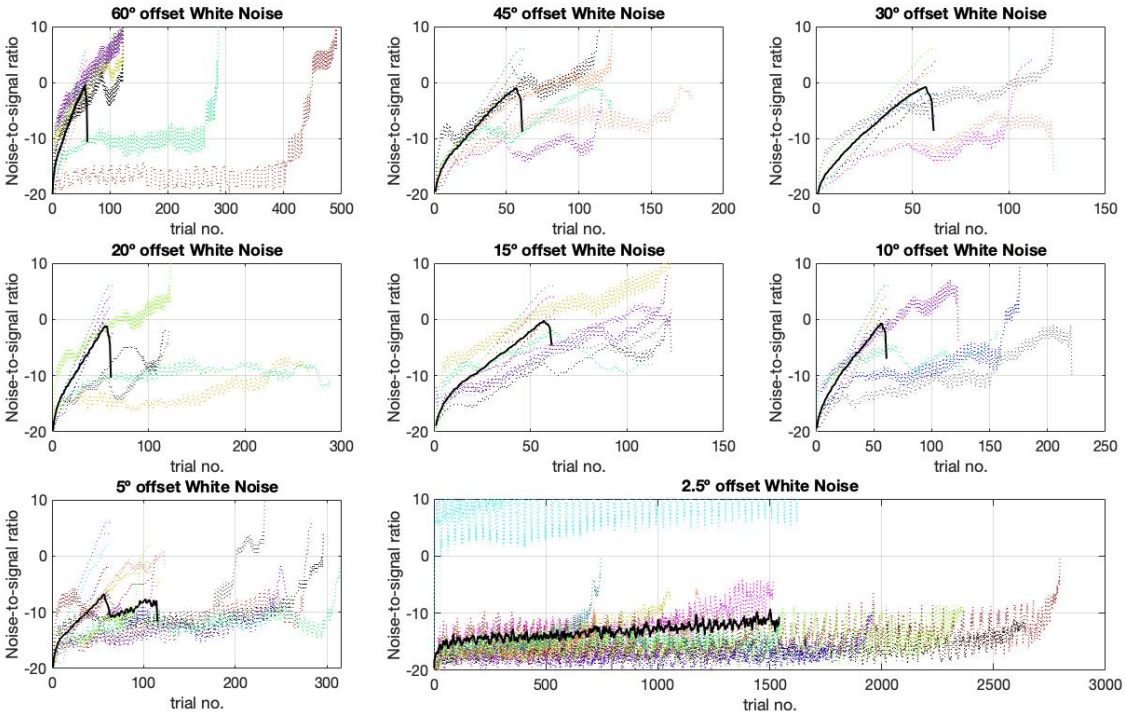

Figure SA 4: Spatialized white noise task progression. Shows individual progression across noise levels relative to target in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

STM Tasks

STM Intro task

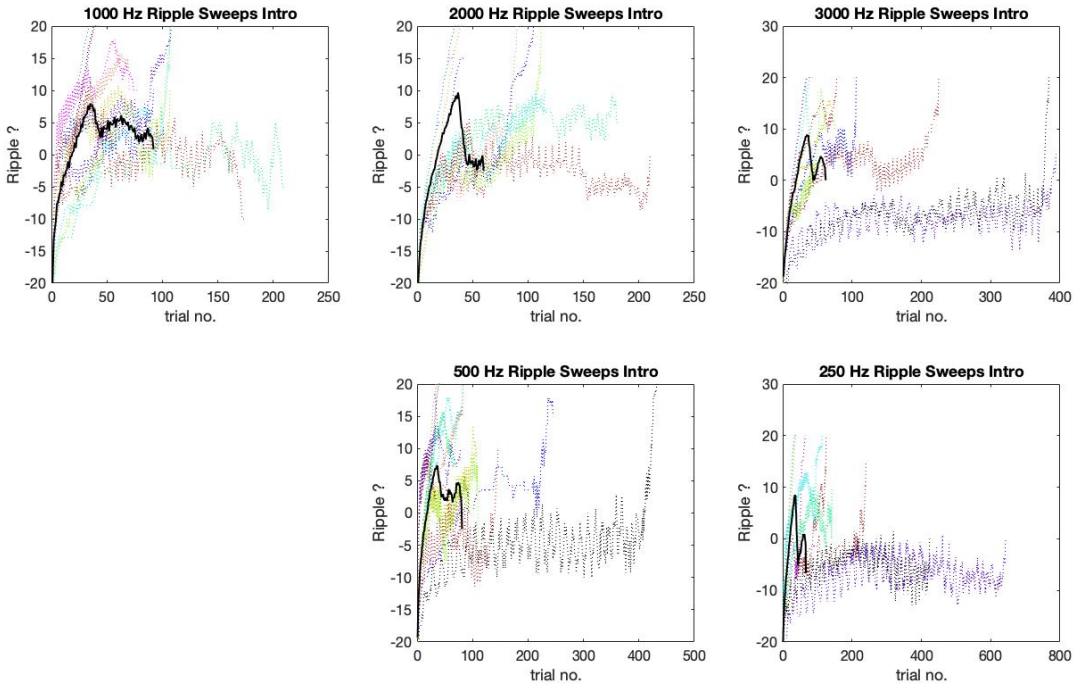

Figure SA 5: STM Intro task progression. Shows individual progression across target ripple levels in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

STM Duration task

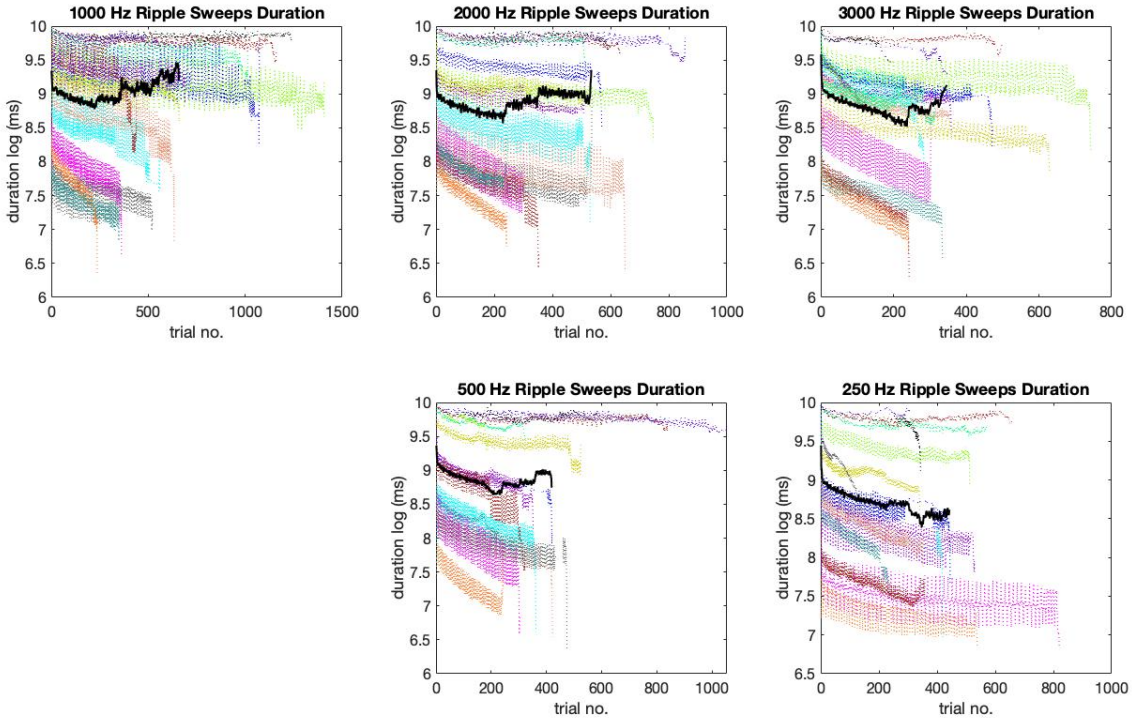

Figure SA 6: STM Duration task progression. Shows individual progression across different target durations (log transformed) in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

STM Slope task

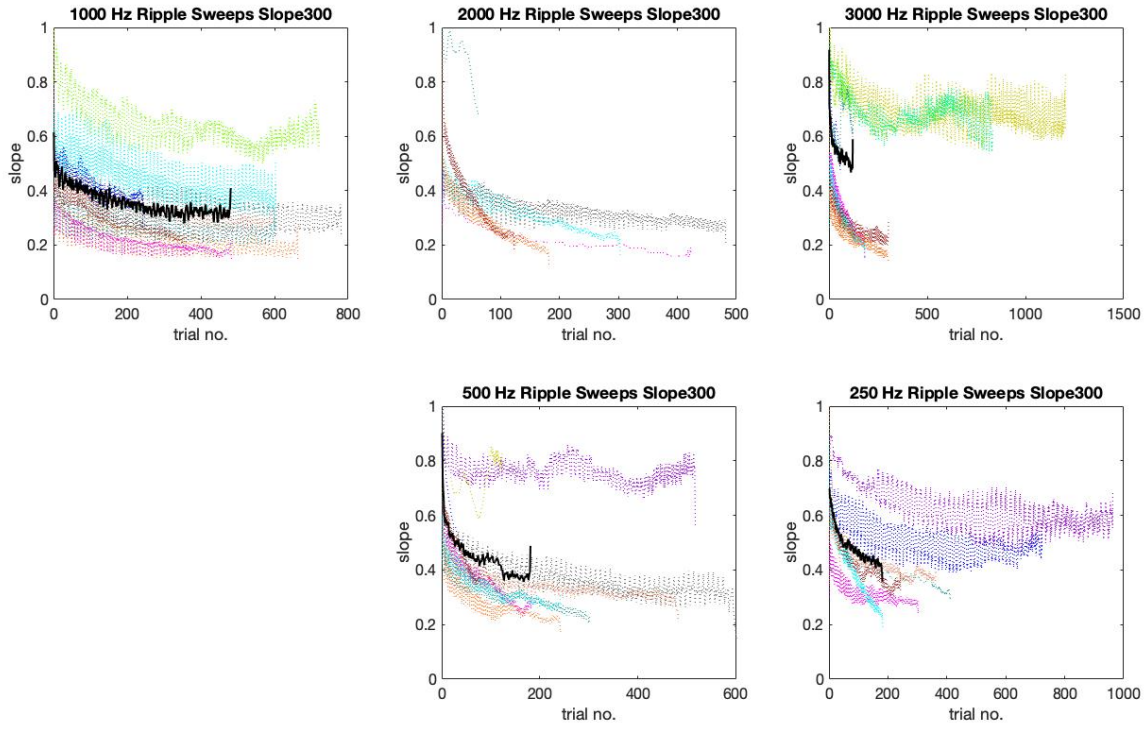

Figure SA 7: STM Slope task progression. Shows individual progression across different ascending or descending target slopes in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

STM Noise task

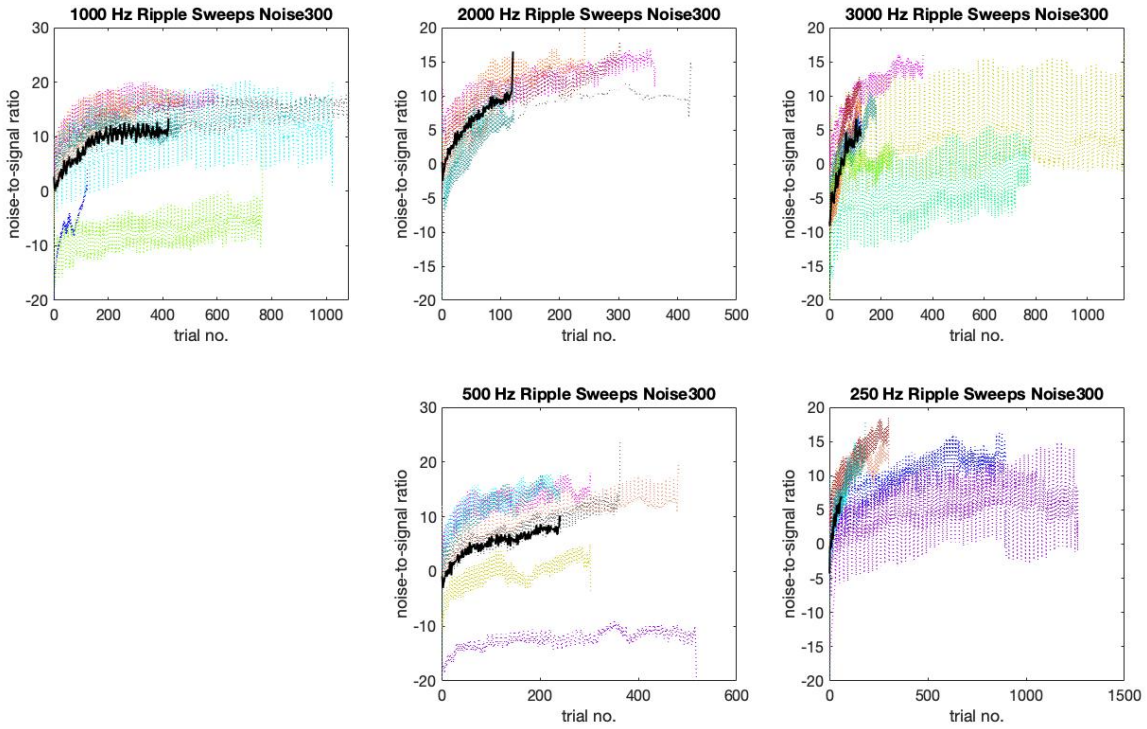

Figure SA 8. STM Noise task progression. Shows individual progression across different levels of noise in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

STM Depth tasks

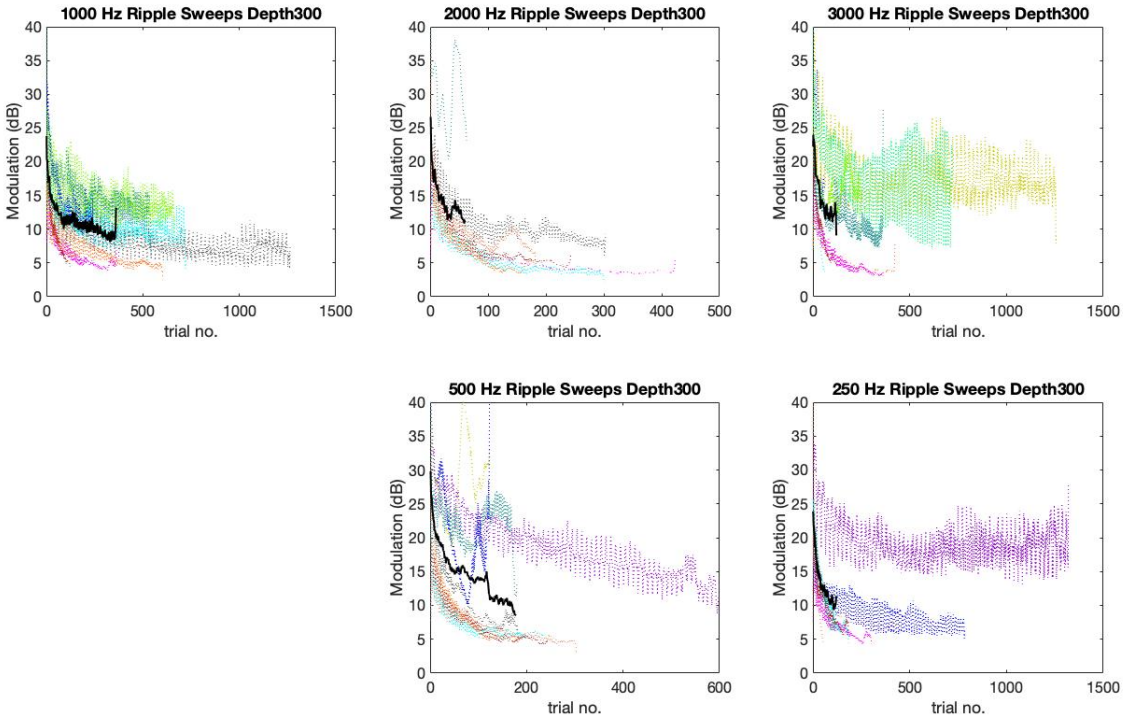

Figure SA 9: STM Depth task progression. Shows individual progression across different levels of modulation depth (dB) in dotted lines of different colors and mean performance is shown in black. Mean performance seems to drop by the end because the better performers have dropped out of the task. All data shown is smoothed with a window of 7 trials, and the mean line shows a minimum of 5 participants.

Memory Tasks

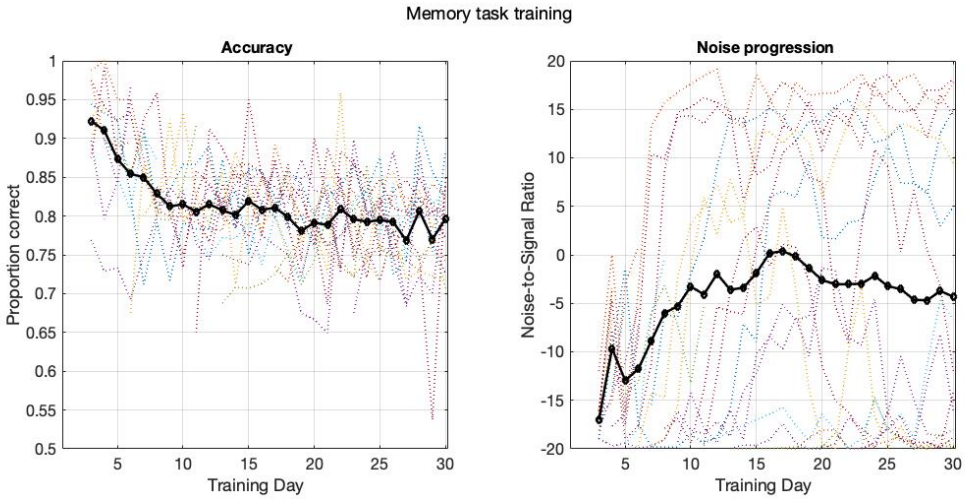

Figure SA 10: Memory tasks progression. Shows performance on the working memory n-back tasks. Accuracy drops at first as participants transition from a 1-back to a 2-back condition and then is kept around 80% (panel on the right). At the same time the noise level increases with training day. Individual performance is depicted dotted lines of different colors and mean performance is shown in black.

Section B. Minimum audibility assessment exploration

In this section, we provide individual's information on each of the assessments tested including the minimum audibility tests. Also, group analysis on mid and follow up tests is shown.

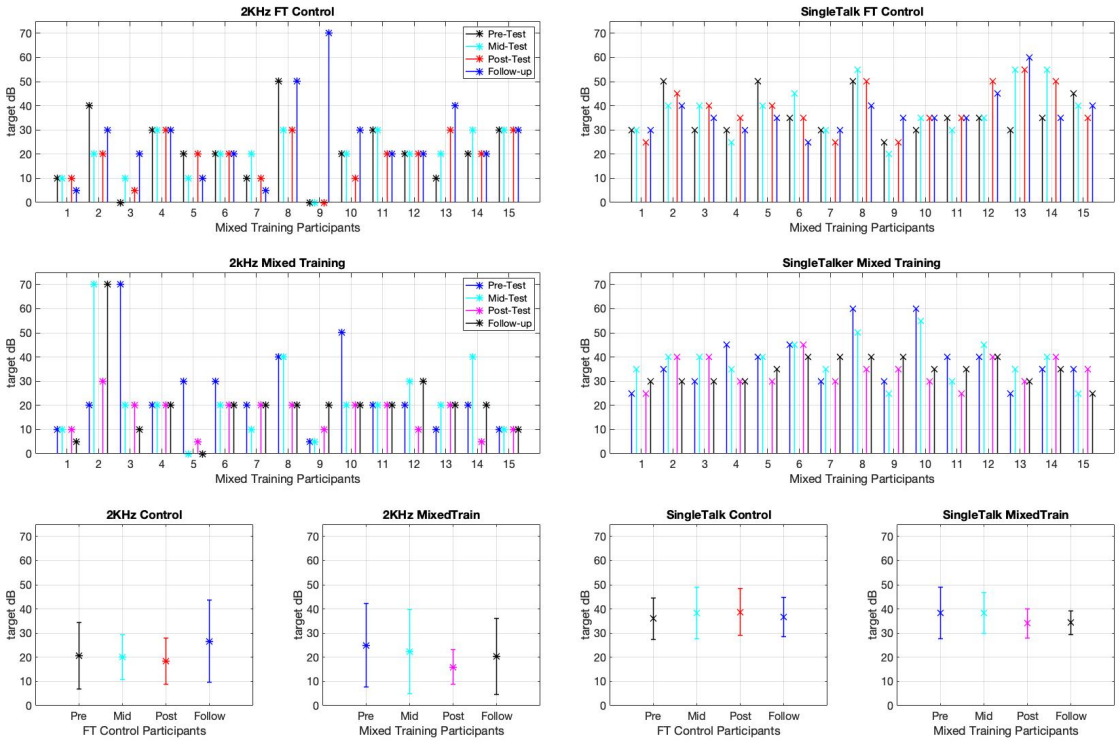

Figure SB 1: Performance on minimum audibility tests. Panels on the left show the 2 kHz pure tone detection task in quiet and panels in the right performance on the CRM single talker condition. Top panels show control group performance, mid panels show the mixed group, and bottom panels show summary data (mean and standard error).

CHAPTER FIVE: GENERAL DISCUSSION AND FUTURE DIRECTIONS

The work presented in this dissertation gives a somewhat broad depiction of the critical elements involved in the investigation of perceptual learning (PL) in the case of the mechanical senses. This overview is achieved by two training studies in different modalities (tactile and audiovisual) and an assessment validation that delves into the evaluation of auditory processes. After the introductory chapter 1, we provide in chapter 2, a specific instantiation of a training study with vibrotactile information (Larrea-Mancera et al., 2019). Chapter 2 makes explicit that assessment selection is crucial for proper evaluation of PL. So in the following chapters 3 & 3b we show an assessment validation study where a testing platform to evaluate different aspects of auditory function is detailed (chapter 3 & 3b; Larrea-Mancera et al., 2020 & Larrea-Mancera et al., 2021a). Finally, before this conclusion, chapter 4 depicts a second training study with *Listen*, an auditory video-game that uses the assessments validated in the previous chapters to evaluate its impact on audition across multiple dimensions of interest, importantly, the ability to understand speech in noisy conditions (Larrea-Mancera et al., 2021b). Chapter 4 is similar to chapter 2 in the sense that they are both training studies, but the complexities both of assessment and training are considerably increased. A wider diversity of auditory and cognitive assessments affords exploration of the neurocognitive mechanisms involved. At the same time,

this complexity is hard to track from a mechanistic point of view as many things are occurring during training that could explain the observed effects in the outcome measures. Nevertheless, the inability to track mechanism beyond an exploratory stage is not an issue for the efficacy scope prespective whithin which the results show promise for future intervention. Strategic selection of perceptual assessments regarding the relevant dimensions in which learning needs to be evaluated and the interaction with the different scopes of the studies can be observed throughout the different chapters and could be informative for decision making in the design of future studies that follow similar methodologies. Finally, the instruments used in chapters 3, 3b and 4, namely PART and *Listen* (see https://braingamecenter.ucr.edu/games/) represent research tools readily available to further research in the domains of hearing assessment and training.

There are still several challenges to be addressed and lines of inquiry that can be advanced in the case of every chapter compiled here. Starting with the second chapter (Larrea-Mancera et al., 2019) which invloves the research on the critical stimulation aspects that can promote generalization of PL in the tactile domain, there are future directions to pursue. We were able to show that the experimental manipulation of bandwidth –or amount and complexity of the perceptual information– promotes different patterns of transfer across the dimensions tested. However, we were unable to tell the simple story we initially hypothesized, that more bandwidth would correspond to more transfer. This was

found only for simple stimulus features in the broad-band group, and the group of people that trained with the narrow-band simple frequency vibrations showed more generalization of PL across untrained fingers. More research could be done to better understand the training elements that lead to transfer of PL and whether transfer across stimulus dimensions and transfer across digits follow different principles. For example, the differences between the stimulation used in the broad-band case and the narrow-band case were not limited to bandwidth. The broad-band case was based on music, with different streams of complex vibrations including several frequencies modulated over time and spectrum. The narrow-band case was constituted by randomized successions of a limited pool of simple vibrations and fixed durations. Systematic and orthogonal manipulation of the elements of spectral pattern, rhythmic or temporal pattern and bandwidth will be a way to explore this in future studies. Of note, these dimensions have been shown to be processed differentially since very early in sensory processing at least in the auditory case (see Larrea-Mancera, Rodríguez-Agudelo & Solís-Vivanco, 2017).

Furthermore, given the main aim of application of that perceptual work –in robotics and prosthetics– motor activity should perhaps be involved in a much more intimate and reverberated interaction with the sensory information than what the paradigm used affords. Participants received vibratory patterns at their fingertips without the need for interaction with the stimulation. Motor responses

dentoing a perceptual decision was made were only given after two vibratory patterns were delivered and the participants had to decide whether they were the same or not. A more interactive paradigm that allows for touch and motricity to interact reciprocally rather than sequentially would better align with the demands of controlling a prosthetic device (see Gibson, 1966; 1979; O'Reagan & Nöe, 2001). Following this idea and following a notion already present in the first study regarding the integration of information across the hands, we are looking to conduct a follow-up study where participants will have to balance between the fingers an object with variable weights in each trial. Behavioral measurement will be supported by an accelerometer in the balancing object recording its position. We started piloting this study when the COVID-19 pandemic hit the world in the beginning of 2020, and the study has been pending return to in-person research.

It is important to note that this line of research exploring the touch domain should not to be taken independently from the auditory research described in the next chapters (3, 3b & 4) as there has been a lot of studies noting the correspondence of auditory, touch and visual information in tasks like understanding speech (see Yehia, Rubin & Vatikiotis-Bateson, 1997; Rosenblum, Dias & Dorsi, 2017). Interactions with touch stimulation and auditory speech perception have been reported (Gick & Derrick, 2009; Ito, 2009, Treille et al., 2014), and some have even used tactile information to train audition and improve on auditory tasks including speech intelligibility (Fowler & Dekle, 1991; Ciesla et

al., 2021). Also, tactile stimulation can be delivered through most mobile devices including the ones targeted by the Brain Game Center (some of which are reported in this work) and could be included in training approaches like the one reported in chapter 4 (*Listen*: An auditory experience; Larrea-Mancera et al., 2021b). Moreover, tactile influence in the auditory domain can be assessed with tools like the one reported in chapters 3 & 3b (PART; also see Peng et al., 2020) as in addition to multi-sensory interaction, we should also expect unisensory gain from multi-sensory training (Shams et al., 2011).

The auditory assessment tool (PART: Portable Automatic Rapid Testing) presented in chapters 3 and 3b represents the most potent asset gained in the current series of studies. PART presents a tangible possibility to expand the reach of auditory research beyond the confines of the laboratory out to the world through consumer grade devices (see Gallun, 2020) that are able to generate laboratory grade stimulation (see Gallun et al., 2018). Moreover, the type of assessments that are being exported out of laboratory confines have great potential to compliment clinical practice. This is because all of the assessments tested, as well as other psychophysical tests that are possible to generate with PART are mostly absent from the clinic, which has remained focused on assessing pure-tone thresholds (Füllgrabe, Moore & Stone, 2015; Gallun et al., 2013; Gallun et al., 2014; Hoover et al., 2019; Mehraei et al., 2014). The assessment of central auditory processes that could be informative for building a

patient's auditory profile in the clinic has typically been overlooked. PART can thus be instrumental in generating large datasets across a variety of groups of people so that psychophysical testing can be properly translated into clinical practice.

Not only can PART expand the toolset available for the researcher and the clinician, but it may also expand their reach in terms of the groups that can be addressed. This point is very important in research settings were most of the information comes from very specific samples of undergraduate students attending research institutions. This is exactly the type of samples used for the science reported in this manuscript, but PART be used to expand the reach of auditory assessment. We are looking to conduct a series of studies where we take this portable tool and venture into the world at large searching for neglected populations to expand our observations and clarify the extent to which our results are representative. Furthermore, the way this future direction plays out in the clinical domain is perhaps the single-most promising future direction of the present work. PART can be used for clinical screening so that both the classical testing based on pure tone thresholds as well as more complex measures of auditory processes and even cognitive assessments such as those reported in chapter 4 (Larrea-Mancera et al., 2021b) could be gathered. Problematic cases identified by such a screening procedure could be then further addressed in clinical settings using more traditional clinical practices. Work of this sort could

help describe the cognitive and hearing health of underserved populations that have been left out of healthcare systems because of accessibility issues, or simply because their hearing processing loss often goes unnoticed (see Saunders et al., 2019). Describing a somewhat complete picture of auditory processes in the population is of utmost importance as early interventions lead to better hearing outcomes (Pronk et al., 2011), and hearing loss has been reported as an important modifiable risk factor for cognitive decline later in life (Livingston et al., 2017; 2020).

Once more information about the cognitive and auditory health of different groups of people is collected, different types of interventions can be suggested ranging from the traditional amplification-based techniques such as hearing aids to more novel approaches to auditory and cognitive training such the auditory training (AT) using the video-game *Listen* portrayed in chapter 4 (see Pronk et al., 2011; Stropahl et al., 2020). The accurate and reliable identification of auditory processing ability in terms of research and clinical screening presents a first step from which adequate training approaches for different hearing profiles and needs can be developed.

The AT intervention we propose in chapter 4 (*Listen*; Larrea-Mancera et al., 2021b) adapts on a vast number of parameters associated to different dimensions of auditory processing in such a way that progress can be made differentially, and the training adjusted to individual needs. However, the training

needs of different groups are yet to be addressed and it is very likely adjustments will need to be made to the current version of the training to address them. In other words, the variability of personalization afforded by current intervention might not be enough to grant an adequate interaction for different groups of people with different needs. Feasibility data will need to be collected moving forward with different groups of people with diverse needs to ensure intervention adequacy, however the current preliminary results with a young normal hearing population seem promising. As we are able to conduct such feasibility studies with different populations, we will be able to increment the possibilities of *Listen's* adaptability. In the future, Listen may prove to be a useful intervention accessible to many with the potential of improving people's lives through the improvement of hearing in difficult conditions.

An important nuance to note here is that, as is shown in chapter 3b (Larrea-Mancera et al., 2021a) there are small but systematic differences to be expected when testing outside of the lab in the variability of environments, devices and headphones used immanent in remote testing with participant owned devices. We are currently collecting data on another couple of conditions where participants are either: tested remotely or in the laboratory (between-subject factor) both with their own equipment, and with devices and headphones calibrated in the laboratory (within-subject factor), with the aim of clarifying the source of these differences. Whether or not these sources of additional variation

will complicate the remote testing and AT intervention on different groups of people remains to be tested. Efforts in securing quiet testing conditions and environmental sound monitoring can further help control and identify sources of variation of remote testing, however we also note the repeatability of measures was equivalent to laboratory conditions. This suggests the feasibility of correcting for this systematic offset in performance, modifying the expected values for clinical screening. In the case of training studies, this type of systematic deviation from laboratory measurement should express both at the level of pre- and post-assessments and so the use of PART remotely to assess the effect of interventions such as AT circumvent the issue.

There are also ways in which both assessment and training can be improved as well. In the case of assessment, work can be done in developing more efficient algorithms that allow to estimate perceptual thresholds even more rapidly. This in turn would afford the inclusion of more assessments in the same amount of time avoiding fatigue and taking the most advantage of experimental or screening time. Another way to improve assessment would be to find ways to embed it into training so that it more naturally reflects performance and its fluctuations as suggested by Seitz (2018). Further development and optimization of both assessment and training as well as the collection of more data that allow for correlative studies to be performed between validated assessment and training aspects is needed for this end. However, I believe the current approach

using video-games is a potent one as it allows to embed the perceptual elements intended for training in a behaviorally meaningful interaction that can range from simple to complex, from uni-sensory to multi-sensory, from purely receptive to actively generative relationships between stimuli and responses. This complexity potentially present in video-games is the reason why a number of findings in PL regarding generalization of learning can be put together into a single behavioral paradigm as done in Larrea-Mancera et al. (2021b; see also Deveau, Lovcik & Seitz, 2014; Deveau, Ozer & Seitz, 2014). Not only is this characteristic of video-games compelling for conducting research but also allows for more ecologically valid settings of testing where complex aspects of reality can be emulated to different degrees (see MacIver, 2011).

Finally, I want to note that even when in this work there are examples of novel approaches to PL that embrace complexity in the stimulation paradigm used in training (e.g. tactile music and auditory video-game) and greatly expand on usually tested domains (e.g. several measures of central auditory processes), there are a number of ways in which they embrace reductionistic bias that may overlook relevant aspects of PL. Our paradigms reduce the complexity as well as the contingencies of the lawful information in the environment, crucial to ecological behavior, to yield highly artificial conditions. This can be observed in the simple vibro-tactile equipment used in chapter 2 with a single piston stimulating a discrete region in a participant's finger. In the case of the AT video-

game, we may observe the senses of vision and audition are not fused into a single perceptual object. While the visual stimuli are tightly bound to the actions that can be carried out in the game so that every action has a contingent visual aspect, sounds simply dictate which response should be given at a later time. The sounds have no visual contingent aspect and no simultaneous action that covaries with its presentation. Auditory stimuli and their associated responses are artificially segregated from visuo-motor information and into corresponding stimulus and response phases so that the reciprocity and intimacy of the multi-sensory and sensori-motor relationships are broken or not present. Further work using this same paradigm but modifying the existent relationships between audiovisual and audiomotor elements may be informative to determine the relevance of multi-sensory and sensori-motor aspects inherent in naturalistic stimulation and ethological paradigms (see MacIver, 2009). Additionally, the inclusion of haptic feedback into this multi-sensory and sensori-motor logic will be an interesting and potentially fruitful avenue of exploration towards the development of a more optimal AT paradigm.

In conclusion, the work compiled here portrays a somewhat broad picture of conducting PL research across the mechanical senses (touch and audition), where the complexities of assessment and training choices and the different possible scopes of perceptual training were layed out.This dissertation represents a methodological tool at the same time that it provides examples on

the use of PART (auditory assessments) and *Listen* (auditory training), research assessment and training tools developed by the Brain Game Center that have plenty of potential beyond the confines of this dissertation. Lastly, several lines of future work are planned and detailed to some extent above to address limitations and further explore ideas presented in this dissertation.

REFERENCES

Ciesla, K., Lorens, A., Skarżyński, H., Wolak, T., & Amedi, A., (2021). Speech-to-touch sensory substitution : a 10-decibel improvement in speech-in-noise understanding after a short training. 1–17. *Preprint available in Research Square.* https://doi.org/10.21203/rs.3.rs-429202/v1

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology*, *24*(4), R146–R147. https://doi.org/10.1016/j.cub.2014.01.004

Deveau, J., Lovcik, G., & Seitz, A. R. (2014). Broad-based visual benefits from training with an integrated perceptual-learning video game. *Vision Research*, *99*, 134–140. https://doi.org/10.1016/j.visres.2013.12.015

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. Journal of Experimental Psychology: Human Perception and Performance, 17(3), 816–828. https://doi.org/doi:10.1037/0096- 1523.17.3.816

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, *7*(JAN), 1–25. https://doi.org/10.3389/fnagi.2014.00347

Gallun, F. J., Diedesch, A. C., Kampel, S. D., & Jakien, K. M. (2013). Independent impacts of age and hearing loss on spatial release in a complex auditory environment. *Frontiers in Neuroscience*, *7*(7 DEC), 1–11. https://doi.org/10.3389/fnins.2013.00252

Gallun, F. J., McMillan, G. P., Molis, M. R., Kampel, S. D., Dann, S. M., & Konrad-Martin, D. L. (2014). Relating age and hearing loss to monaural, bilateral, and binaural temporal sensitivity. *Frontiers in Neuroscience*, *8*(8 JUN), 1–14. https://doi.org/10.3389/fnins.2014.00172

Gallun, F. J., Seitz, A., Eddins, D. A., Molis, M. R., Stavropoulos, T., Jakien, K. M., Kampel, S. D., Diedesch, A. C., Hoover, E. C., Bell, K., Souza, P. E., Sherman, M., Calandruccio, L., Xue, G., Taleb, N., Sebena, R., & Srinivasan, N. (2018). Development and validation of Portable Automated Rapid Testing (PART) measures for auditory research. *175th Meeting of*

the Acoustical Society of America, *33*(May), 050002. https://doi.org/10.1121/2.0000878

Gallun, F. J. (2020). Flipping the lab: Using consumer electronics for high-quality data collection. *179th Meeting of the Acoustical Society of America*, *032002*(2020), 032002. https://doi.org/10.1121/2.0001418

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston, MA: Houghton Mifflin.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. London: Routledge.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, 462, 502–504. doi:10.1038/ nature08572

Hoover, E. C., Pasquesi, L., & Souza, P. (2015). Comparison of Clinical and Traditional Gap Detection Tests. *Journal of the American Academy of Audiology*, *26*(6), 540–546. https://doi.org/10.3766/jaaa.14088

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. Proceedings of the National Academy of Sciences of the United States of America, 106(4), 1245–1248. https://doi.org/10.1073/pnas.0810063106

Larrea-Mancera, E. S. L., Rodríguez-Agudelo, Y., & Solís-Vivanco, R. (2017). Musical rhythm and pitch: A differential effect on auditory dynamics as revealed by the N1/MMN/P3a complex. *Neuropsychologia*, *100*(January), 44–50. https://doi.org/10.1016/j.neuropsychologia.2017.04.001

Larrea-Mancera, E. S. L., Dempsey-Jones, H., Makin, T., & Seitz, A. R. (2019). Does Training on Broad Band Tactile Stimulation Promote the Generalization of Learning? *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 1–6. https://doi.org/10.1109/DEVLRN.2019.8850704

Larrea-Mancera, E. S. L., Stavropoulos, T., Hoover, E., Eddins, D., Gallun, F., & Seitz, A. (2020). Portable Automated Rapid Testing (PART) for auditory research: Validation in a normal hearing population. *Journal of the Acoustical Society of America*, *148*(4), 1831–1851. https://doi.org/10.1101/2020.01.08.899088

Larrea-Mancera, E. S. L., Stavropoulos, T., Carrillo, A. A., Eddins, D. A., Molis, M. R., Gallun, F. J., & Seitz, A. R. (2021a). Portable Automated Rapid Testing (PART) of auditory processing abilities in young normally-hearing listeners : A remotely administered replication with participant-owned devices . *PsyArXiv*. https://psyarxiv.com/9u68p/

Larrea-Mancera, E. S. L., Philipp, M. A., Stavropoulos, T., Anna, A., Cheung, S., Koerner, T., Molis, M. R., Gallun, F. J., & Aaron, R. (2021b). Training with an auditory perceptual learning game transfers to speech in competition. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2021.01.26.428343

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., Cooper, C., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Larson, E. B., Ritchie, K., Rockwood, K., Sampson, E. L., … Mukadam, N. (2017). Dementia prevention, intervention, and care. *The Lancet*, *390*(10113), 2673–2734. https://doi.org/10.1016/S0140-6736(17)31363-6

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., … Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*, *396*(10248), 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

MacIver, M. A. (2009). Neuroethology: From Morphological Computation to Planning. The Cambridge Handbook of Situated Cognition. P. Robbins and M. Aydede. New York, NY, Cambridge University Press: 480-504.

Mehraei, G., Gallun, F. J., Leek, M. R., & Bernstein, J. G. W. (2014). Spectrotemporal modulation sensitivity for hearing-impaired listeners: Dependence on carrier center frequency and the relationship to speech intelligibility. *The Journal of the Acoustical Society of America*, *136*(1), 301–316. https://doi.org/10.1121/1.4881918

Murray, C. A., Larrea-Mancera, E. S. L. De, & Glicksohn, A. (2020). Revealing multisensory benefit with diffusion modeling. *Journal of Mathematical Psychology*, *99*, 102449. https://doi.org/10.1016/j.jmp.2020.102449

O'Regan, J., & Noë, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, 79–103. http://www.springerlink.com/index/nl361h8276502488.pdf

Peng, Z. E., Buss, E., Shen, Y., Bharadwaj, H., Stecker, G. C., Beim, J. A., Bosen, A. K., Braza, M., Diedesch, A. C., Dorey, C. M., et al. (2020). Remote testing for psychological and physiological acoustics: Initial report of the p&p task force on remote testing. In Proceedings of Meetings on Acoustics 179ASA, volume 42(1), page 050009. Acoustical Society of America. https://doi.org/10.1121/2.0001409

Pronk, M., Kramer, S. E., Davis, A. C., Stephens, D., Smith, P. A., Thodi, C., Anteunis, L. J. C., Parazzini, M., & Grandori, F. (2011). Interventions following hearing screening in adults: A systematic descriptive review. *International Journal of Audiology*, *50*(9), 594–609. https://doi.org/10.3109/14992027.2011.582165

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2017). The supramodal brain: implications for auditory perception. *Journal of Cognitive Psychology*, *29*(1), 65–87. https://doi.org/10.1080/20445911.2016.1181691

Saunders, G. H., Frederick, M. T., Silverman, S. P. C., Penman, T., Gardner, A., Chisolm, T. H., Escabi, C. D., Oree, P. H., Westermann, L. C., Sanchez, V. A., & Arnold, M. L. (2019). Hearing screening in the community. *Journal of the American Academy of Audiology*, *30*(2), 145–152. https://doi.org/10.3766/jaaa.17103

Shams, L., Wozny, D. R., Kim, R., & Seitz, A. R. (2011). Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology*, *2*(October), 264. https://doi.org/10.3389/fpsyg.2011.00264

Seitz, A. R. (2018). A New Framework of Design and Continuous Evaluation to Improve Brain Training. *Journal of Cognitive Enhancement*, *2*(1), 78–87. https://doi.org/10.1007/s41465-017-0058-8

Stropahl, M., Besser, J., & Launer, S. (2020). Auditory Training Supports Auditory Rehabilitation: A State-of-the-Art Review. *Ear & Hearing*, *41*(4), 697–704. https://doi.org/10.1097/aud.0000000000000806

Treille, A., Cordeboeuf, C., Vilain, C., & Sato, M. (2014). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. *Neuropsychologia*, *57*(1), 71–77. https://doi.org/10.1016/j.neuropsychologia.2014.02.004

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1–2), 23–43. https://doi.org/10.1016/S0167-6393(98)00048-X