Portable Automated Rapid Testing (PART) for auditory research: Validation in a normal hearing population

E. Sebastian Lelo de Larrea-Mancera[1], Trevor Stavropoulos[1], Eric C. Hoover[2], David A. Eddins[3], Frederick J. Gallun[4], Aaron R. Seitz[1]

[1]Brain Game Center, University of California Riverside, 900 University Ave, Riverside, California 92507.

[2]University of Maryland, College Park, MD

[3]University of South Florida, Tampa, FL

[4]Oregon Health and Science University, Portland, OR

correspondence should be sent to: elelo001@ucr.edu

***Abstract*— We describe data collected using Portable Automated Rapid Testing (PART), a freely-available application for psychoacoustical testing that harnesses commercially available tablet computer technology to translate current psychophysical knowledge into clinical practice. PART tests included the detection of tones in noise with and without spectral gaps; spectral, temporal, and spectro-temporal modulation; diotic and dichotic frequency modulation; and temporal gaps inserted between brief tone pulses. Listeners also performed a speech-on-speech spatial release from masking test. Data from 150 undergraduate students were collected using both passive and active noise-attenuating headphones in a silent environment and in the presence of recorded cafeteria noise. Across these and other manipulations of equipment and threshold-estimation techniques, performance reliably approximated that reported in the literature. These data serve as validation that accessible auditory hardware can be used to test auditory function with sufficient precision to provide clinical assessments of central auditory function in individual listeners. 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. PART has the potential to supplement the testing currently being done in the clinic to provide a clearer picture of auditory function and health.**

Keywords*— *psychophysics, ambulatory auditory testing, central auditory function, normative data

I. Introduction

There is much that is still not well understood about the diversity of hearing difficulties that people may face throughout their lifespan as they attempt to make sense of different auditory scenes. One of the main reasons we have so much to learn about the disabilities of hearing is the way in which hearing loss has been studied and approached in the clinic. Modern clinical audiology was translated from the laboratory before the 1960s (Carhart & Jerger, 1959, Hughson & Westlake, 1944), and has remained focused on using pure-tone audiograms to assess audibility and speech tests to evaluate the ability to detect particular acoustical cues in speech (see CHABA, 1988). As a result of this limited focus on audibility, there are very few tools and even fewer protocols available to develop a profile of auditory abilities that might be used in the diagnosis and/or treatment of auditory difficulties associated with hearing loss or other diseases. Indeed, many measures of auditory perceptual abilities associated with age or hearing impairment have been identified in the laboratory as independent from, or only weakly predicted by, audibility or performance on clinical speech tests (Moore, 2014; Eddins & Hall, 2010; Gallun et al., 2013). Thus, the fields of audiology and hearing science are deeply divided on the question of how to handle impairments of the non-peripheral auditory system, which are often called (central) auditory processing disorders (see letter by Moore, 2018, and response by Iliadou, et al., 2018).

The motivation for this work is the belief that to make progress on this topic, clinically accessible tests of the many components of functional hearing are needed. Such tests should be applied and validated across individuals with diverse hearing abilities to clearly characterize the measures that are most informative about the variety of hearing needs of different individuals and groups of individuals. While a number of candidate tests have been developed and are relatively well studied in laboratory settings (e.g. Moore et al., 1987; Grose and Mamo, 2012; Bernstein et al., 2013; Gallun et al., 2014; Jakien et al., 2017), to date it has been difficult to translate them to clinical-practice and clinical-research settings mainly because they are costly in time and human resources. Until such testing can be reliably, quickly, and easily accomplished, it will be difficult to gather the datasets necessary to bridge the knowledge gap between laboratory and clinical practice, and to discover relationships between perceptual abilities and remediation. To address this gap, several state-of-the-art psychometric tests that are currently being used to research central auditory processes in

the laboratory have been translated into the application PART (Portable Automatic Rapid Testing). PART was developed by the University of California Brain Game Center (https://braingamecenter.ucr.edu) and is currently freely available on the Apple App Store. PART can be run on mobile devices (e.g. iPad, iPhone) and has been shown to be capable of accurately reproducing highly precise acoustic stimuli (Gallun et al., 2018).

PART was designed to ease the implementation of tests that have been shown to represent a broad range of hearing abilities that differ between young adults with normal hearing adults and one or more groups of people for whom listening in complex environments is difficult, for example, older adult listeners and others with hearing impairment (e.g., Bernstein et al., 2013; Gallun et al., 2013; Füllgrabe, Moore & Stone, 2015) such as after traumatic brain injury (Hoover, Souza & Gallun, 2017). The specific psychophysical tests chosen for the battery used in this study represent a small subset of PART's functionality. The battery was designed to reflect the description of the central auditory system put forth by current research in psychoacoustics and auditory neuroscience (e.g., Stecker & Gallun, 2012; Bernstein et al., 2013; Depireux, Simon, Klein & Shamma, 2000). We synthesized this description into three sub-batteries of tests with supporting evidence of clinical utility, namely the temporal fine structure, spectro-temporal modulation, and targets in competition. These three groups of tests address different stages of auditory processing by the central nervous system.

Acoustic information processing in the periphery (cochlea & auditory nerve) results from differences in the relative amplitude of motion in distinct parts of the cochlear partition (Békésy, 1960; Pfeiffer & Kim, 1975). This gives rise to patterns of temporal information in the timing of the all-or-none spikes carried by the auditory nerve fibers ("temporal fine structure" or TFS) that correspond to the temporal patterns of these movements of the cochlear partition (Békésy, 1960; Pfeiffer & Kim, 1975). This 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), along with other systems that have yet to be fully defined but that result in representations of spectro-temporally modulated (STM) information observed in the inferior colliculus (Versnel, Zwiers & Opstal, 2009) and auditory cortex (Kowalsky, Depireux & Shamma, 1996). Psychoacoustical data suggest that these brainstem and early cortical representations are essential in providing cues for auditory scene analysis (Shinn-Cunningham, 2008) and to select targets in noisy multi-talker environments. Together, all of these processes mediate our ability to understand speech in real world conditions. The tests included in the current study were chosen such that they capture one or more aspects of each of these levels of analysis, from temporal fine structure, to spectro-temporal modulation, to speech signals in competition.

TFS coding is assumed to rely upon the precision of phase-locking in populations of auditory nerve fibers and other brainstem neurons that inherit this sensitivity (Tremblay, Piskosz & Souza, 2003; Schimel et al., 2008; Grose & Mamo, 2012; Hoover et al., 2019). 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 conventional speech tests evaluate the detection of frequency modulation, or use any type of spatialization of auditory signals. However, 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, we included a diotic frequency modulation test to assess monaural TFS sensitivity, and a dichotic frequency modulation test to assess binaural TFS sensitivity. We also included a temporal gap detection test (inter-click delay) which has been classically used to assess the sensitivity of temporal processes (Gallun et al., 2014). These three tests have been previously proposed as measures of TFS with potential clinical utility (Hoover et al., 2019).

Spectro-temporal modulation (STM) has been of increasing interest in laboratory studies as auditory cognitive neuroscience has revealed that cortical neurons are most sensitive to modulation of sound in both time and spectrum (Kowalsky, Depireux & Shamma, 1996; Theunissen, Sen & Doupe, 2000; Shamma, 2001; Schonwiesner & Zatorre, 2009). Due to the nature of sound generation, all natural sounds can be characterized as falling within a particular range 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 precede language understanding but require processing 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, in this study, we included tests for temporal-, spectral- and STM sensitivities, assessments that are largely absent from the clinic.

The identification of an acoustic target in competition is considered fundamental to auditory perception and scene analysis beyond peripheral audibility (Shinn-Cunningham, 2008; Moore, 2014). Thus, we included tests that assess the capacity of the system to select relevant information and suppress test-irrelevant interference. On such test is the notched-noise method (Patterson, 1976; Moore and Glasberg, 1990) that evaluates the detection of a 2kHz tone presented in competition with noise either with or without a spectral notch around the target frequency as has been described in Moore (1987). This test allows us to evaluate not only 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 in the context of speech and binaural listening, we used a spatial release from masking test (SRM; Marrone et al., 2008; Gallun et al., 2013; Jakien et al., 2017)

which evaluates speech understanding in competition as well as the ability of the system to use spatial cues to better perceive the target signal. This test uses the Coordinate Response Measure (CRM) corpus (Bolia et al., 2000) to implement a speech understanding evaluation identical to that described by Jakien et al. (2017). Speech understanding in competition is assessed with speech maskers that are co-located in simulated space with the target speech, or with the maskers separated from the target by 45 degrees in simulated space.

The purpose of the work reported here was to establish a normative dataset for this initial PART battery and to examine the degree to which thresholds obtained with PART approximate those reported in the literature for the same tests. A secondary goal was to have a more complete picture of the accessibility afforded by PART. To this end, data were collected using both passive and active noise-attenuating headphones in a silent environment and in the presence of recorded cafeteria noise. To achieve these goals, we obtained threshold estimates in young normal hearers recruited from the University of California, Riverside campus. The data presented below are interpreted to mean that across these and other manipulations of equipment and threshold-estimation techniques, PART was able to produce reliable estimates. These data serve as validation that accessible auditory hardware can be used to test auditory function with sufficient precision to provide clinical evidence of central auditory function in individual listeners. This data set 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.

II. METHODS

A. Participants

We recruited 150 undergraduate students from the University of California, Riverside (47 male, M age = 19.3 years, SD = 2.36 years), who received class credit for their participation. All participants provided signed informed consent as approved by the University of California, Riverside Human Subject Review Board, reported normal hearing and vision, and no history of psychiatric or neurological disorders. Since the sample being tested were university students participating in exchange for class credit. Some failures to comply with procedures were observed and so cases of outlying performance were likely due to participant lapses rather than auditory dysfunction per se. Consequently, no referrals or follow-ups were provided for these cases. Currently, there is no gold-standard test of central auditory processing, so it was not possible to rule out dysfunction in these cases. While most data were included in the reported dataset, in the case of the notch test, one extreme outlier had to be removed from all further analysis (see Supplemental Materials).

B. Materials

All experiments were conducted on iPad tablets (Apple, Inc., Cupertino, CA) running the PART (Portable Automatic Rapid Testing) application. Stimuli were delivered via internal soundcard and either Sennheiser 280 Pro headphones (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) 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). The SLM app and iBoundary microphone system was 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. This procedure is further detailed in Gallun et al. (2018) and can be done in the field with relatively inexpensive commercially available equipment. Central auditory function assessments were delivered to participants through either Sennheiser 280 Pro headphones 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. The same calibration settings were used for the two headphone types, which resulted in different output levels. The levels described are for the calibrated Sennheiser system. Differences in output and performance for the Bose system are described below.

C. Procedure

In each session, participants sat in a comfortable chair in a sound-treated room and listened through a set of headphones connected to an iPad running PART. Instructions were delivered as on-screen text displayed in PART and responses during psychophysical testing were recorded by touching the iPad screen. Participants started each session with our screening test which presented 10 trials of a 2kHz tonal signal at 45dB SPL in quiet. All participants were able to detect this signal with at least 90% accuracy and continued with the detection of that same 2kHz signal in competition with noise. After this first block of testing, participants would continue with one of the following testing-block possibilities: TFS, which included 3 assessments (described below); STM (3 assessments); or speech-on-speech competition (2 assessments). All assessments were preceded by 5 non-adaptive practice trials at a high point in the staircase (i.e. at an easy point on the test). The delivery of the testing-blocks was counter balanced across sessions and participants. On each trial of the 2-cue, 2-alternative forced choice tests, four intervals were sequentially presented to participants audio-visually (see fig. 1 top left panel) with an inter-stimulus-interval (ISI) of 250ms, the first and last of which were presented as cues. Participants had to find the instructed signal among the two alternatives presented between the cues and respond by touching the corresponding square on the screen. The square would then turn either green (correct) or red (incorrect) as response feedback before proceeding to the next trial (1 sec ITI). All participants completed all test-blocks on each session of the experiment. Each individual assessment took around 5 minutes to complete and it took about 50 minutes to complete the battery with all four testing-blocks.

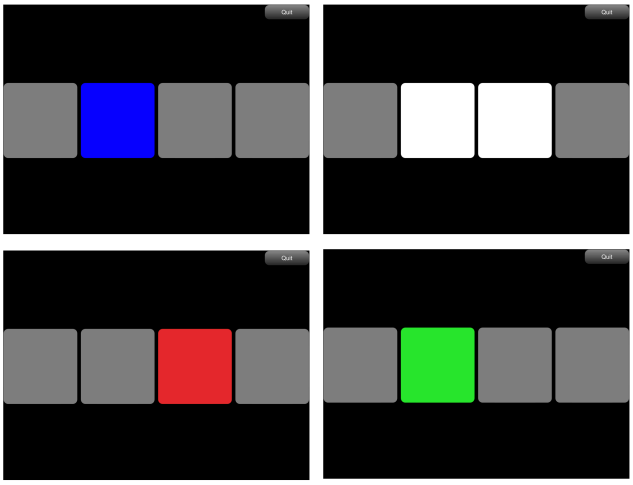

Fig 1. Each panel represents a screen-shot taken from PART while on a 2-cue, 2-alternative forced-choice test. Each box lit up sequentially in blue emitting a sound (top-left). After all intervals were played the 2 alternatives in the middle became available for response (top-right). Feedback is shown by color code (red = wrong; bottom panels).

To adjust difficulty, with the exception of the spatial release tests that progressed in a set sequence, the designated parameter was adapted via a 2-down 1-up staged staircase. It took 2 correct responses to either decrease the signal-to-masker ratio (SMR) or the magnitude of the sound modulation depth, so that the test would become more difficult to perform. Incorrect responses either increased SMR or the magnitude of the sound modulation depth employed, thus making the test easier to perform in general terms. The size of the steps-up were 1.5 times the size of the steps-down (2:1 for the first experiment), and this relative magnitude was maintained across 2 stages in the adaptive staircase. This step ratio is different from the standard ratio of 1:1 used in most adaptive testing for psychoacoustics. This different step size ratio was chosen to encourage learning of the test, as errors resulted in a rapid return to clearly identifiable targets. The consistency of the results and the similarity to published data, despite very brief tracks, indicates that this step size choice had a negligible effect on performance. Further, the number of trials necessary to reach a threshold estimate was reduced slightly for the 1.5:1 ratio (see Supplemental Materials). The first stage of all staircases contained 3 reversals and the magnitude of the steps was five times the size of the second stage. The second stage of the staircase stopped after 6 reversals. Further details about the adaptive procedures are described for each test below. All data were automatically saved on the iPad for later analysis. Participants were encouraged to take small breaks between testing-blocks. The total duration of each experimental session was about 1 hour.

D. Stimuli

1. Temporal Fine Structure

a. Temporal Gap - In the context of a 4-interval, 2-cue, 2-alternative forced-choice test (Gallun et al., 2014), the target signal consisted of a gap or delay placed between two 0.5 kHz tone bursts of 4 ms played at 80dB SPL, similar to the monaural gap detection reported in Gallun et al. (2014) and Hoover et al. (2019) but delivered diotically. Non-targets presented both clicks sequentially with no additional gap between them. Inter-click delay was the designated adaptive parameter with a minimum of 0 ms and a maximum of 100 ms starting at 20 ms. The staircase adapted on an exponential scale with a major factor of 2 divided in 20 steps. Thus, one step corresponds with a factor of 2^(1/20), and 20 steps works out to a total factor of 2. The first three reversals of the staircase adapted on an up: down step-size ratio of 15:10 steps and the next six reversals on a 3:2 ratio. Note that Gallun et al. (2014) referred to this as a gap discrimination test, as the standard also has a gap due to the amplitude ramps imposed on the two tone bursts.

b. Diotic Frequency Modulation - Similar to Grose & Mamo (2012), Whiteford & Oxenham (2015), Whiteford et al. (2017), and Hoover et al. (2019) and in the context of a 4-interval, 2-cue, 2-alternative forced choice test, the target signal consisted of a pure tone carrier frequency randomized between 460 and 550 Hz with a frequency modulation of 2 Hz presented at 75dB SPL for 400 ms. Presentation of the modulation was identical between the ears (diotic frequency modulation). Non-targets were a non-modulated version of the stimulus. Modulation depth was the designated adaptive parameter with a minimum of 0 Hz and a maximum of 10 kHz starting at 6 Hz. The staircase adapted on an exponential scale with a major factor of 2 divided in 20 steps, yielding the same step factor of 2^(1/20) as above. The first three reversals of the staircase adapted on an up: down step-size ratio of 15:10 steps and the next six reversals on a 3:2 ratio. Listeners were instructed to detect the “wobble” in the target interval, which corresponds to a small shift in the carrier frequency over time at the modulation rate of 2 Hz. Randomization of the carrier frequency ensures that the test cannot be successfully conducted by simply listening for a different frequency in one of the intervals. Instead, the listener must detect a within-interval change in frequency over time.

c. Dichotic Frequency Modulation - Following recent versions of this test method (e.g., Grose & Mamo, 2012; Hoover et al., 2019), the dichotic FM stimuli were identical to the diotic FM stimuli described above with the exception that the modulation was inverted or anti-phasic between the ears. This corresponds to the FM/FM condition in Grose & Mamo (2012) and replicated in Hoover et al. (2019). Non-targets were again a non-modulated version of the stimulus. Modulation depth was the designated adaptive parameter with a minimum of 0 Hz and a maximum of 10 kHz starting at 3 Hz. The staircase adapted through an exponential scale with a major factor of 2 divided in 20 steps. The first three reversals of the staircase adapted on an up: down step-size ratio of 15:10 steps and the next six reversals on a 3:2 ratio. This stimulus, first developed by Green et al. (1976), creates a continuously shifting interaural phase difference (IPD) in the target interval. The depth of the FM determines the size of the IPD. For those listeners sensitive to IPD, Green et al. (1976) and later investigators have demonstrated that dichotic FM can be detected when the depth is much smaller than in the diotic case. Hence, FM depth at threshold relative to threshold in the diotic test provides a measure of the degree to which IPD is being used to detect the target interval instead of monaural frequency differences.

2. Spectro-Temporal Sensitivity

a. Spectral Modulation - Similar to Hoover, Eddins & Eddins (2018), but using the same presentation method as in the other tests described above, the target signal consisted of a broad-band noise (400 to 8 kHz) upon which a sinusoidal spectral modulation was imposed at a rate of 2 cycles per octave (c/o) on a logarithmic amplitude scale. Non-targets were a non-modulated version of the same broad-band noise. All noises were generated for each interval using random amplitude and phase values, and the phase of the modulation was randomized. The stimuli were generated in the frequency domain and the number of components were the maximum allowed by a 44.1 kHz sampling rate. Modulation depth (designated as M dB) was measured on a logarithmic scale with reference midpoint-to-peak dB and was adaptively varied with a minimum of 0.2 (M) dB and a maximum of 40 (M) dB and a starting value of 6 (M) dB. For details on the measurement of modulation depth for spectrally-modulated signals, see Isarangura et al. (2019). The staircase adapted through a linear scale divided in steps of .05 (M) dB. The first three reversals of the staircase adapted on an up: down step-size ratio of 15:10 steps and the next six reversals on a 3:2 ratio.

b. Temporal Modulation - The stimulus, based on that used by Viemeister (1979) and many others since, was the same randomly-generated bandpass noise used for the spectral modulation detection test, built with temporal amplitude modulation (AM) at a rate of 4 Hz imposed on a flat-frequency broadband carrier. Non-targets were identical to the standard stimulus used in the spectral modulation test. Modulation depth (M) dB was the designated adaptive parameter with a minimum of 0.2 (M) dB and a maximum of 40 (M) dB starting at 6 (M) dB. The staircase adapted through a linear scale divided in steps of .05 (M) dB. The first three reversals of the staircase adapted on an up: down step-size ratio of 15:10 steps and the next six reversals on a 3:2 ratio.

c. Spectro-Temporal Modulation - The same stimulus used in the spectral and temporal modulation tests was used, with the difference that the target interval contained both 2 c/o spectral modulation and 4 Hz AM, based on the findings of Bernstein et al. (2013) that this is a crucial combination of spectral and temporal modulation for predicting how well hearing-impaired listeners can understand speech in noise. The direction of the resulting spectrotemporal modulation (STM) was randomly assigned to go upward or downward in frequency over time on each trial. A signal-generation technique was developed that allowed real-time generation of stimuli in the frequency domain. Modulation depth (M) dB was the designated adaptive parameter with a minimum of 0.2 (M) dB and a maximum of 40 (M) dB starting at 6 (M) dB. The staircase adapted through a linear scale divided in steps of .05 (M) dB. The first three reversals of the staircase adapted on an up: down step-size ratio of 15:10 steps and the next six reversals on a 3:2 ratio.

3. Signals in Competition

a. No-Notch Condition - This abbreviated notch-noise method is adapted from Moore et al. (1987) and includes the no-notch reference condition and the 0.2 fc notch comparison condition, where fc is the center frequency of 2 kHz. The difference in threshold between the two conditions is taken as an index of frequency (spectral) resolution. In the context of the same 4-interval, 2-cue, 2-alternative forced-choice method used in the other tests, the target signal was a 2 kHz tone presented at 45 dB SPL for 500 ms simultaneously with a noise masker centered on the target frequency with a bandwidth of 800 Hz (1.6 to 2.4 kHz). The RMS masker level was the designated adaptive parameter with a minimum of 25 dB SPL and a maximum of 90 dB SPL starting with 35 dB SPL. The staircase adapted through a linear scale divided in steps of 1 dB. The first stage of the staircase adapted on an up: down step-size ratio of 9:6 steps and the second stage on a 3:2 ratio.

b. Notch Condition - The 0.2 fc notch stimulus was identical to the no-notch condition but in this test the masker was divided into two maskers of 400 Hz in width placed below and above the signal frequency with a separation of 400 Hz such that the two maskers covered the frequency regions of 1.2-1.6 kHz and 2.4-2.8 kHz. This is equivalent to a notch width of 0.2 times the center frequency (fc) of 2 kHz as described by Moore et al. (1987). Masker level was the designated adaptive parameter with a minimum of 25 dB SPL and a maximum of 90 dB SPL starting with 35 dB SPL. The staircase adapted through a linear scale divided in steps of 1 dB. The first stage of the staircase adapted on an up: down step-size ratio of 9:6 steps and the second stage on a 3:2 ratio.

c. SRM Co-located - The three-talker speech-on-speech masking method of Marrone et al. (2008) which was 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, who 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 (co-located). Each progressive track included 20 trials in which the maskers both progressed in level from 55 dB SPL to 73 dB SPL in steps of 2 dB as reported in Jakien et al. (2017), resulting in 2 responses at each of the 10 target-to-masker ratios (TMRs).

d. SRM Separated - The stimuli were identical to those in the co-located 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 (2 tracks) from 55 dB SPL to 73 dB SPL in steps of 2 dB as reported in Jakien et al. (2017).

E. Experimental Design

In addition to the method development that is the central theme of this work, we evaluated whether PART could be used in a variety of settings. For PART to serve as a supplemental tool to clinical practice, it would be optimal if it were not only portable, automated, and rapid, but also did not require specialized resources. For example, rather than requiring administration in a sound booth, it would be ideal if testing could be conducted in noisier conditions like a standard procedure room, or even a waiting room. To this end, we tested whether PART would produce reliable threshold estimates in different external noise conditions. This further motivated us to test the effects of noise cancelling headphones across both silent and noisy conditions. Thus, to test the flexibility of PART to be used in different levels of environmental noise and with headphones varying from standard passively attenuating circumaural headphones to consumer-grade noise attenuating headphones, we tested students using PART in the three experiments described below. Each test session could involve up to three participants seated next to each other in a single room, listening and responding independently.

1. Experiment 1 (standard).- The first 51 students were tested in a double-walled experiment room while wearing Sennheiser 280 Pro headphones. These headphones were rated to have up to 30 dB passive attenuation, based on frequency, and were originally used to calibrate the system (Gallun et al., 2018).

2. Experiment 2 (headphones comparison in silence).- The next 51 participants were tested in the same room but were tested with either the Sennheiser 280 Pro headphones or active-noise cancelling Bose Quiet Comfort 35 headphones. Each participant was tested with each headphone type once with the order of sessions being counter-balanced between participants. The calibration was not adjusted when switching the headphones, which provided a further test of the robustness of the system to hardware variations. Testing of the Bose system 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 revealed an overall reduction in the output level by 14 dB, but no distortions in the time or frequency domain. Retaining the same calibration for both headphones allowed the same system to be used and only the headphones changed, which reduced the chance of errors during data collection. It also provided a test of the robustness of the system to variations in overall level.

3. Experiment 3 (headphones comparison in noise).- The next 48 participants were tested in a noisy environment, with methods otherwise identical to Experiment 2. The noise was recorded in a local coffee shop, combined through digital waveform editing to remove silent gaps between recordings and transient recording noise at the beginning and ends of the recordings, and then bandpass filtered to fall entirely in the region between 20 and 20,000 Hz. The coffee shop noise contained a large number of sound sources at all times,

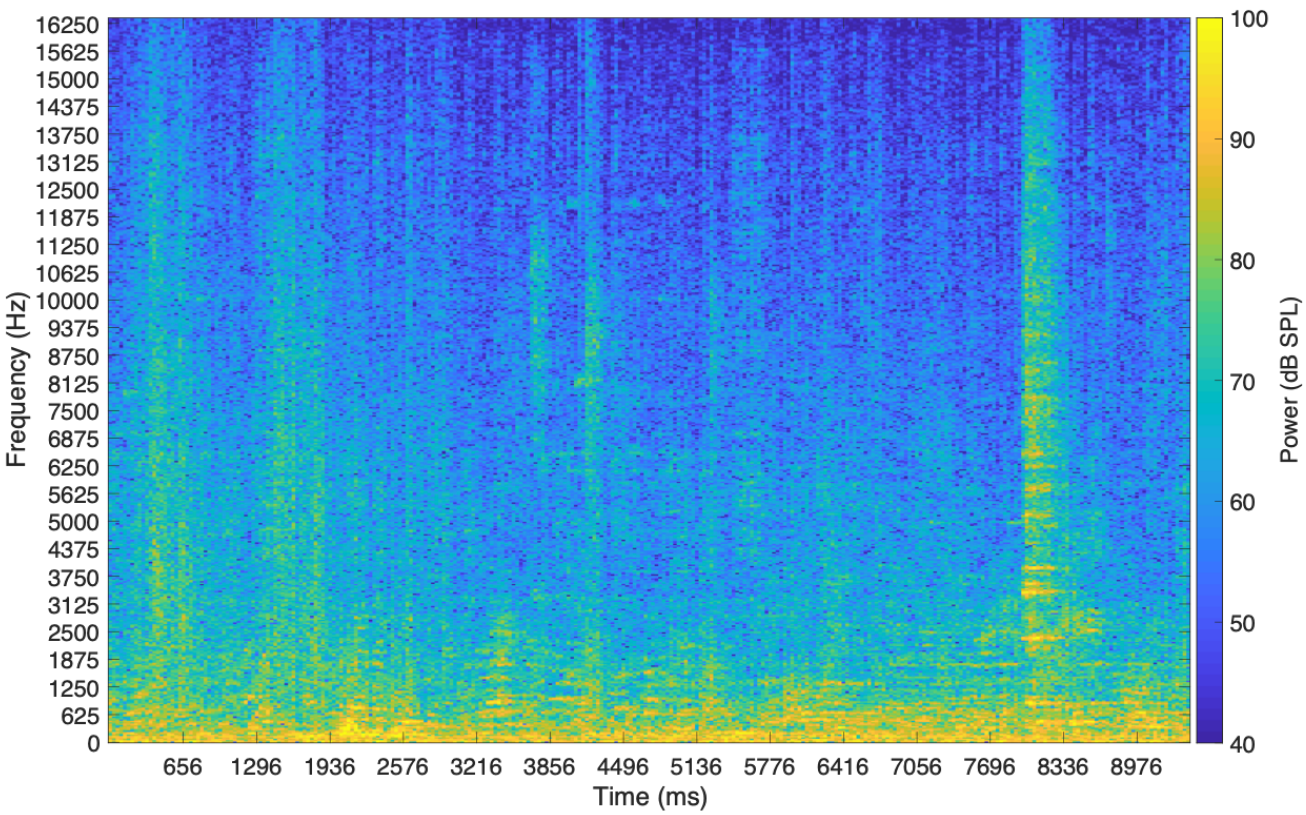

Fig 2. Representative nine-second segment of the cafeteria noise utilized in Experiment Three. The total recording had a duration of 33 minutes and was played in a continuous loop during testing.

including both speech and environmental sounds. A spectrogram of a representative segment is shown in Figure 2. Sound files, after processing, were 33 minutes in duration and were played on a loop through a loudspeakers placed 3 meters from the center of the listening room. Calibration of sound output was implemented by placing an iBoundary microphone in the center of the listening room and adjusting the output level to 70 dB SPL.

F. Data Analysis

Behavioral thresholds were calculated from the mean of the last 6 reversals of each of the adaptive assessments. The exception being the speech-on-speech competition assessments, where thresholds were calculated by subtracting the number of correct responses from the starting Target-to-Masker Ratio (TMR) as described in Jakien et al. (2017). The results are divided in three subsections: 1) Test-retest reliability for the two sessions in each experiment was evaluated using Limits of Agreement tests (LoA, Altman & Bland, 1983) to assess systematic error and bias, as well as by correlations between sessions. Additionally, differences between sessions were analyzed using repeated-measures t-tests and their associated effect sizes. All data except for the rejected outlier mentioned above and detailed in the supplemental materials are presented in this section. In some cases we provide additional analysis using a conservative outlier rejection filter of 3.92 SD (twice the critical value of the z distribution) in order to ensure that only very unrepresentative values were removed. This is appropriate as the goal is not to measure typical performance but rather to measure the degree to which the two sessions produced similar threshold estimates. 2) The second subsection is designed to allow comparison to the extant literature. Here the goal is to determine whether or not PART methods produce results that are consistent with previously published data. To minimize the influence of confounding factors on this analysis and to maximize relatability to previous reports in the laboratory, we used a more rigorous filter of 1.96 SD (critical value of a z distribution) as outlier rejection. 3) The third subsection involves analysis of the effects of headphone types with and without noise-attenuation technology and external noise conditions. To address the effects of our experimental manipulations, composite scores were computed by normalizing each individual assessment relative to its mean and standard deviation (a z-score transform), and averaging z-scores across tests for each participant. LoA plots, Pearson correlations, and t-tests are reported for the composite score estimates divided by headphone noise-attenuating type, for each experiment separately. To test differences across experimental manipulations, a mixed-model Analysis of Variance (ANOVA) was used to compare composite scores across the factors of interest.

III. RESULTS

Because the results are highly similar across all three experiments (see Fig. 3), and our main intention with this dataset is to provide normative threshold estimates of central auditory function in a variety of settings to potentially supplement clinical practice, the results are described a manner that focuses on the combined data set across experiments. This analysis uses composite measures to address the differences between experiments (aggregating across tests). More detailed reports about the tests in each experiment can be found in the supplemental materials.

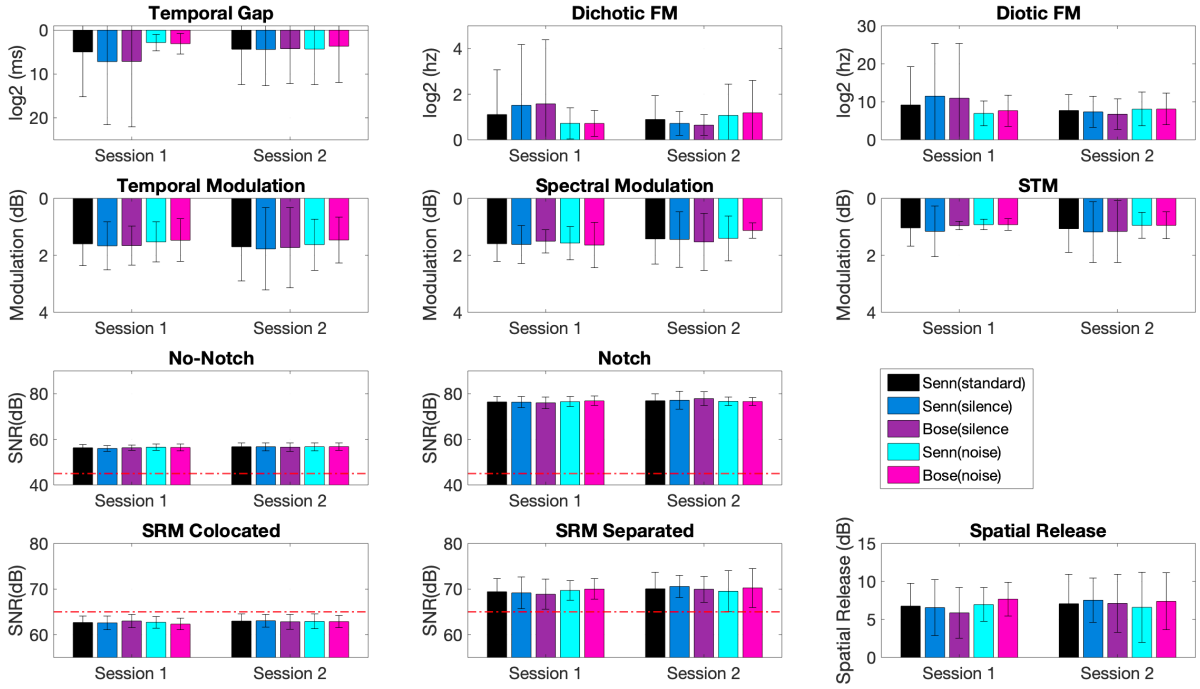

Fig 3. Mean thresholds and standard deviations obtained for each headphone used in each experiment in each test. The direction of the y-axes has been inverted when necessary so that better performance is always towards the top. A red dotted line indicates the level of the target in the target in competition tests.

A. *PART yields reliable threshold estimation in young normal hearing participants*

To address the extent to which tests in PART are reliable, limits of agreement (LoA), as described by Altman and Bland (1983), were used to evaluate test-retest reliability of estimates. An advantage of this technique is that it affords the evaluation of reliability based on the within-subject standard deviation of the measures. This analysis is based on the evaluation of performance across sessions (mean of test and re-test) as a function of their difference. This analysis can help indicate systematic bias (e.g. if either session consistently yields better estimates, such as learning effects), 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). LoA is used as a main analysis instead of the more typical Pearson correlation because we anticipated the between-subject variability to be small –as the sample consisted solely of young normal listeners– and correlations are known to depend heavily on between-subject variability (Altman & Bland, 1983; Bland & Altman, 1990). Nevertheless, numerous studies and research groups have used Pearson correlations to give account of test re-test reliability, and so we do the same and report correlations for the purpose of relatability.

1. Test re-test reliability using limits of agreement

LoA plots for the full dataset are shown in Figure 4. In order to facilitate visual inspection and comparisons across different tests, TFS tests were transformed to log2 units and target-in-competition tests were converted to signal-to-masker ratios. The mean of both sessions is plotted on the x-axis to give a point estimate for each participant relating to the magnitude of the estimated threshold. The difference between sessions is plotted on the y-axis. 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 measurement bias. The 95% limits of agreement (±1.96 SD (difference between sessions)) are plotted as dotted lines and indicates an estimate of the region in which we may expect to observe 95% of the within-subject, between-session differences to be found. Finally, a single bigger circle indicates the mean threshold across participants and sessions along the abscissa and is centered at zero bias on the ordinate. As can be observed in Figure 4., the mean

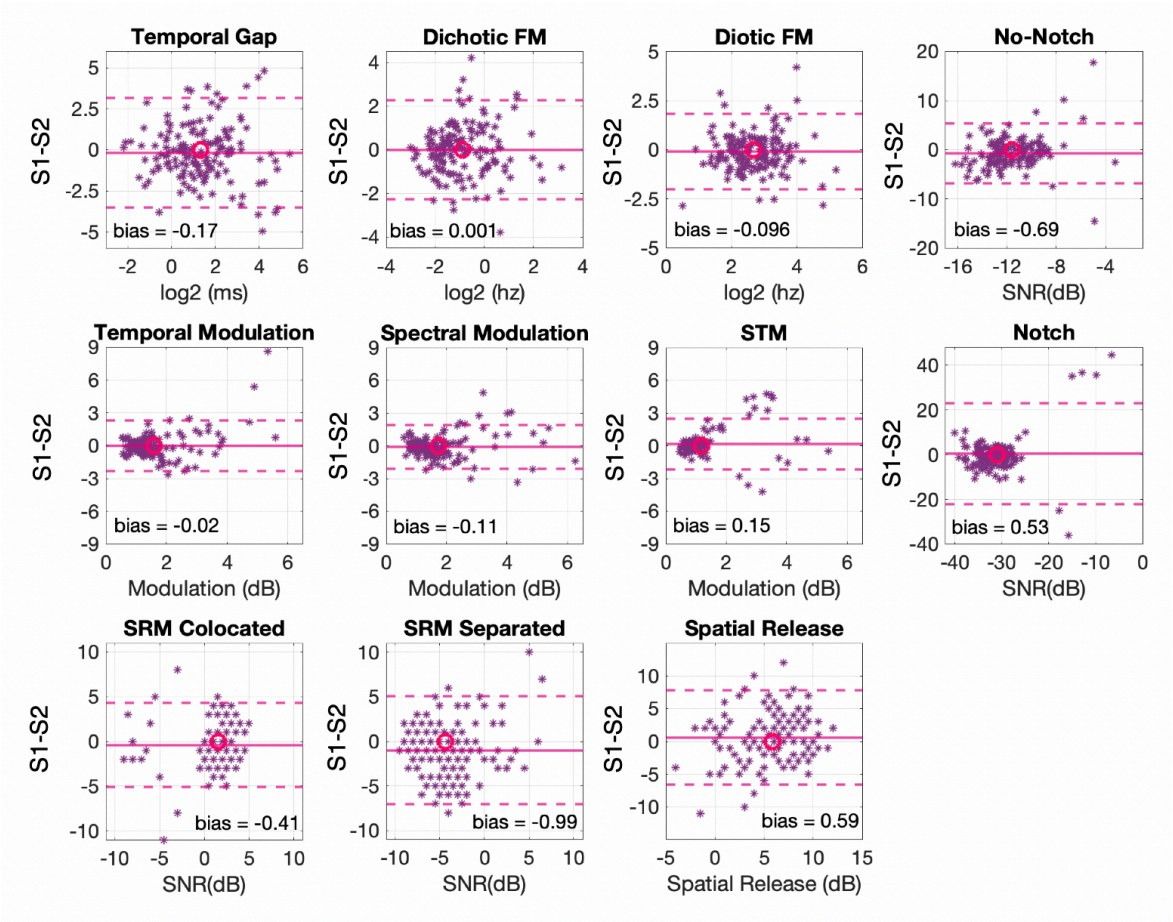

Figure 4. Limits of agreement of the estimated thresholds between sessions for all tests. The solid lines indicate the mean difference between sessions. Dotted lines indicate the 95% limits of agreement. The single bigger circle indicates the mean threshold for each test centered at zero difference between sessions. Any solid lines that fall below zero indicate better performance on session 2 (except the spatial release metric). Masker levels for the targets in competition tests have been converted to signal-to-masker ratios for the purpose of relatability.

difference between sessions is close to zero in all of the psychophysical tests we used, indicating little systematic bias. The measurement error is less than ± 2 log2 (Hz) modulation rates for the frequency modulation tests, ± 3 log2 (ms) for the gap detection, ± 3 dB for the TM, SM and STM tests, and ± 8 dB signal-to-masker ratios for the targets in competition tests. This was not the case for the notch test where a few outliers make the agreement range more than double; an outlier rejection based on ± 3.92 SD eliminates 6 cases and yields limits of agreement = [-9.37, 7.48]. The distribution of the threshold estimates has no salient asymmetries, session differences were similar across different levels of performance (symmetry along the abscissa), and there is little systematic bias between sessions (symmetry along the ordinate) suggesting similar measurement error for both sessions. This analysis demonstrates good test re-test reliabilities, and unbiased estimates at the group level (see Table 2 for relevant statistics).

2. Additional Analysis

Figure 5 shows scatterplots of session 1 vs 2 for each PART assessment. Table 2 shows statistics including the strength of association (Pearson r), differences between sessions (related-samples t-test), and the effect sizes of these differences (mean differences and Cohen's d). Significant correlations were observed for all the assessments except for the notch test. The low correlations found for the notch-noise tests, were mainly due to outlying performance. After a rejection of ± 3.92 SD, 2 cases were eliminated from the no-notch test changing its correlation to r = .41, p < .01; and 6 cases from the notch test changing its correlation to r = .37, p < .01. Overall, the relatively low correlation magnitudes we obtained as an index of reliability are related to performance being distributed across relatively narrow ranges of threshold estimates, as was to be expected for young listeners without hearing problems. 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, which is why the LoA analysis, that is based on within-subject variability, was used as the principal analysis for testing the reliabilities of our measures.

To evaluate whether learning, or other factors, gave rise to systematic changes in performance, thresholds were compared between sessions using tests of significance. There were some small, but significant differences between sessions in the no-notch test (t(149)= 2.7, p < .01), and in the SRM tests (Co-located t(149) = 2.07, p = .04; Separated t(149) = 3.96, p < .01). The magnitude of these differences however is under one fourth of a SD, except in the case of the SRM on the separated condition, where a difference of .32 SD, equivalent to about 1 dB, was found. This difference however is less than the 1.58 dB difference previously

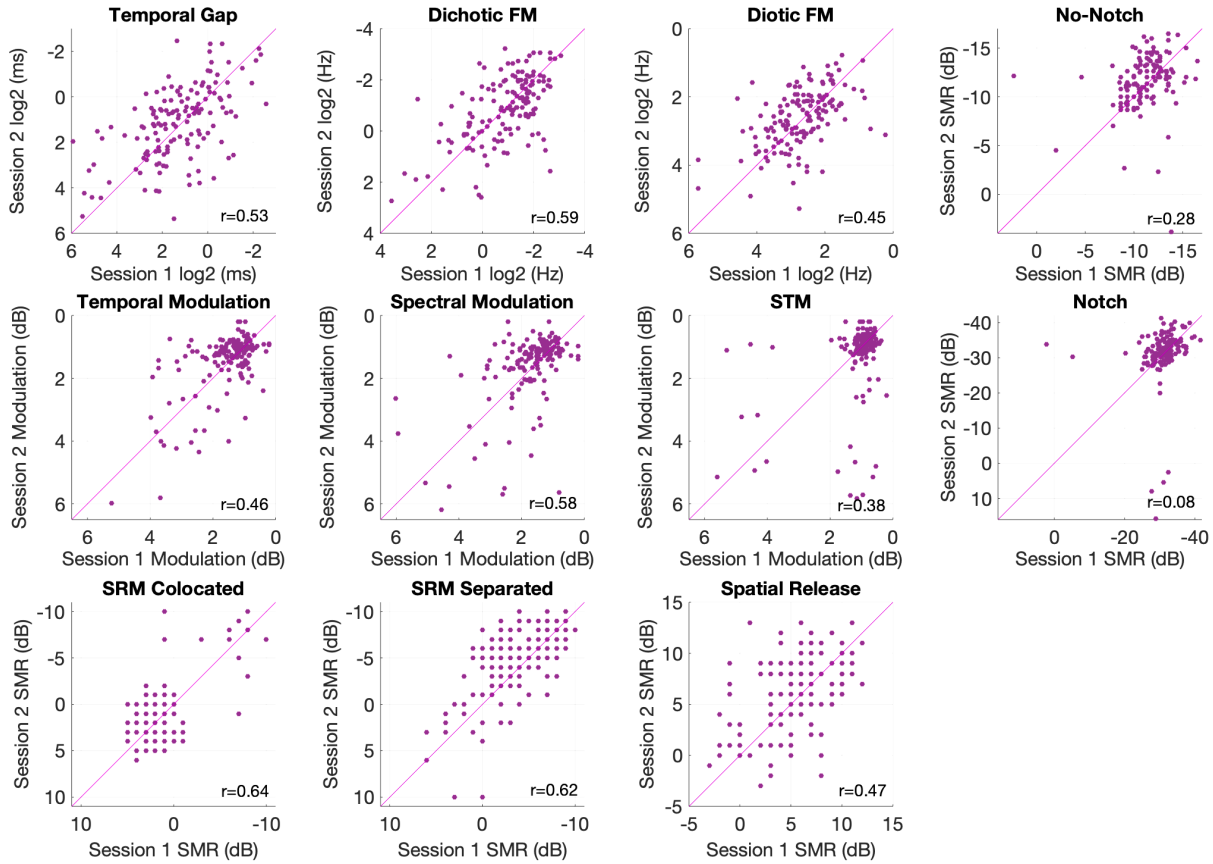

Figure 5. Scatter plots of Session 1 vs Session 2 for the 10 assessments. All axes are oriented to show better performance values away from the origin. Correlations are indicated in the lower right of each panel.

reported by Jakien et al. (2017) as a possible learning effect. A summary of these statistics is shown in Table 2.

B. PART produces threshold estimates similar to those found in the literature

A main aim for this study was to use PART to determine how well thresholds obtained using PART reproduced standard laboratory assessments with the ultimate goal of translating these psychophysical measures in an accessible way to support clinical practice. To this end, we report comparisons of each of our measures to those in the extant literature. To minimize the influence of outlying performance in this section and having already reported the full dataset above, we applied an outlier rejection filter of ± 1.96 SD. Overall, we found threshold estimates that align with previous reports within 1.6 SD (see Table 3).

1. Temporal Fine Structure (TFS)

Sensitivity to TFS was assessed with three different tests; temporal gap detection, dichotic FM, and diotic FM. For temporal gap detection 20 cases were rejected as outliers (6 from Experiment 1; 6 from Experiment 2; 8 from Experiment 3), leaving threshold values that closely resemble those found in the literature (M = 3.57 , SD = 3.42). For example, Schneider et al. (1994) reported thresholds of 3.6 ms on average, using 2kHz tone-bursts similar to the ones we used, however, their stimuli were delivered diotically. Moreover, Hoover et al. (2019) reported thresholds of 1.45 ms using 0.75 kHz tone-bursts, and finally Gallun et al. (2014) used the most similar stimuli (tone-bursts of 2 kHz) and obtained thresholds of 1.2ms on average. The three latter studies in this section used monaural presentation of their stimuli. Despite the mentioned differences, all of these estimates lie within a SD from the PART dataset. The fact that the published data report smaller thresholds and the second run appears 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 worse than those previously reported in the literature. For the dichotic FM test, 11 cases were rejected as outliers (3 from Experiment 1; 3 from Experiment 2; 5 from Experiment 3). Thresholds in Hz (M = .61, SD = .42) are around 1 SD worse than the .2 Hz found by Grose & Mamo (2012), the .15 Hz reported by Whiteford & Oxenham (2015), and the .19 Hz reported by Hoover et al. (2019).

For diotic FM, 13 cases were rejected as outliers (2 from Experiment 1; 7 from Experiment 2; 4 from Experiment 3). Thresholds in Hz (M = 7.07, SD = 3.27), are about 2 SD worse than reports of Grose & Mamo (2012) of 1.9 Hz, Whiteford and 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. 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 2000ms (Whiteford & Oxenham, 2015), but are only 400ms in PART. We used shorter durations than those previously used in the literature following Palandrani et al. (2019), who showed that FM detection thresholds decrease with stimulus duration in cycles of the modulator consistent with other modulation detection tests (e.g. Viemeister 1979). This work would predict thresholds of 3.6 Hz for our stimuli to be comparable to Grose & Mamo (2012), however, our estimates are about 1 SD worse on average. Of note, these different studies report threshold values on different scales including f which is more adequate for a stimulus roving in fundamental frequency. As the measures reported here involved adapting on modulation depth in Hz regardless of the carrier frequency presented in each trial (roving between 460 & 540 Hz), we present our modulation depth values in Hz. These values can easily be converted to f by applying the procedures detailed in Witton et al. (2000). As with the temporal gap, it would not be surprising if repeated testing resulted in reduced thresholds, more similar to those reported in the literature. Nonetheless, the values measured are close (+1 SD) to the range of those anticipated based on previous reports.

2. Spectro-Temporal Modulation (STM)

Sensitivity to STM 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 because reported modulation depth depends on the measurement scale (log or linear), the reference points for the measurement (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, although PART generated stimuli in the modulation scale of M (log midpoint-to-peak) we use equation 1 (below) to convert to 20log(m) dB units as detailed in Isarangura et al. (2019). This provides a single metric that can be used to compare to temporal, spectral and spectrotemporal modulation (after conversion where appropriate).

For STM at 4 Hz and 2 c/o, 22 cases were rejected as outliers (4 from Experiment 1; 6 from Experiment 2; 12 from Experiment 3). STM thresholds obtained (-19.97 dB, SD = 2.08) closely resemble those previously reported in the literature. They are 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).

For SM at 2 c/o, 16 cases were rejected as outliers (1 from Experiment 1; 7 from Experiment 2; 8 from Experiment 3). SM modulation depth thresholds (M = -15.91 dB, SD = 2.94) were better by almost 2 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. These methodological differences are likely to have influenced performance given that in the previous conditions,

performance was slightly better in the literature, which almost always involves more practice for the participants.

For TM at 4 Hz, 14 cases were rejected as outliers (4 from Experiment 1; 2 from Experiment 2; 8 from Experiment 3). TM thresholds (M = -16.17 dB, SD = 3.28) were within 1 SD of those reported by Viemeister (1979) of -18.5 dB for four observers.

3. 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 signal (no-notch condition) or with a 400 Hz spectral notch or protective region without noise (notch condition). 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 to each condition are typically not reported. However, Patterson (1976) reported an average distance between the equivalent of our no-notch and notch conditions of about 24 dB for four participants, which is comparable to the mean distance we obtained here of 20.22 dB (SD = 2.9) where some of our participants performed in the same range. This similarity and the high test-retest reliability of this test suggests that learning plays a small role in the ability to perform this test and that reliable estimates can be obtained with very few trials.

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 co-located in virtual space with the target speech (co-located) 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 co-located condition, 12 cases were rejected as outliers (0 from Experiment 1; 1 from Experiment 2; 11 from Experiment 3). Co-located thresholds (M = 2.18 dB, SD = 1.1) closely resemble those reported by Gallun et al. (2018) across two testing sites (1.85 & 1.96 dB), and those reported by Jakien et al. (2017) (2.8 dB) within half a SD, despite the greater range of ages and hearing abilities in that study.

In the Separated condition, 10 cases were rejected as outliers (3 from Experiment 1; 1 from Experiment 2; 6 from Experiment 3). Separated thresholds (M = -4.91 dB, SD = 2.34) closely resemble those reported by Gallun et al. (2018) across two testing sites (-4.33 & -4.62 dB); they are approximately 1 SD better than those reported by Jakien et al. (2017) for a group of listeners varying in age and hearing ability (-2.5 dB). Further, the difference between the separated and the co-located conditions, a metric indicating the spatial release from masking effects showed spatial release values (M = 6.19 dB, SD = 2.61) that closely resemble the ones reported by Gallun et al. (2018) across two testing sites (6.19 & 6.57 dB), and are within half a SD from those reported by Jakien et al. (2017) (5.3 dB).

C. *Composite scores to evaluate our experimental manipulations.*

To address the robustness of these estimates to variations of procedure, PART was tested in a variety of settings of external noise, equipment, and threshold estimation technique. To address the effects of these manipulations, we constructed a composite score that included all of the measures and compared this composite across experimental manipulations. The composite score was constructed by z-scoring each assessment separately and then averaging the z-scores across the 10 assessments for each participant. We tested the internal reliability of the aggregated composite and found a Cronbach's = .85 which indicates that the composite may be used as a summary score for our central auditory processing test battery. Figure 6 shows the 95% limits of agreement for the composite scores of the whole sample across three experiments (panel on the left). This analysis shows almost zero bias (< 0.001), and limits of agreement [-0.87, 0.87] that indicate that 95% of the time, the composite measures obtained with young normal listeners for each session agree within ± 1 SD. In addition, Figure 6 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 = .69 p<.001, [95% CI = .605, .771]) and represents an estimate of the general reliability of PART.

1. Robust threshold estimates across different experimental manipulations

To address how composite scores changed as a function of listening condition, we also plot the composite score separately for each experiment (Figure 7). Experiment 1 consisted of 51 participants who were tested in a quiet room and received the Sennheiser headphones with 30dB passive attenuation that the system was calibrated on. Experiment 2 was similar to Experiment 1 except participants received either the Sennheiser 280 Pro

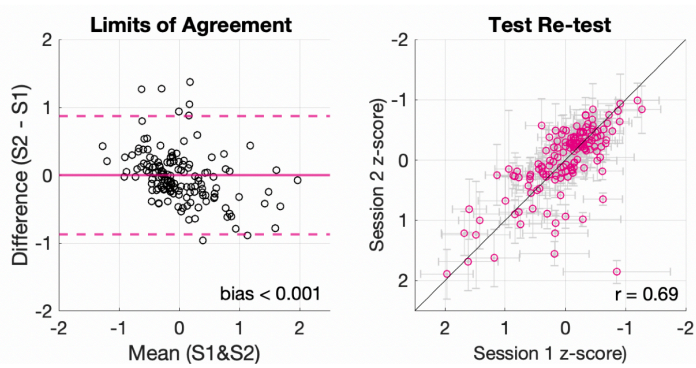

Fig 6. Composite Scores across all three experiments. Panel on the left shows the limits of agreement (see Altman & Bland, 1983) for the composite scores. Panel on the right shows scatterplot of composite scores. The horizontal error bars indicate SEM for session 1 while the vertical bars reflect session 2 SEM.

headphones or the active-noise cancelling Bose Quiet Comfort 35 headphones. Finally, Experiment 3 added environmental noise playing in the background at 70dB on average and is otherwise identical to the methods reported in Experiment 2.

In all three Experiments, composite scores showed minimal bias (Experiment 1 = -.005; Experiment 2 = -.05; Experiment 3 = .06), limits of agreement that resemble the aggregate sample's composite around 1 SD (Experiment 1 [-0.64, 0.63]; Experiment 2 [-1.15, 1.03]; Experiment 3 [-0.74, 0.89]), and similar strength of association between scores of session 1 and 2 with ($r = .691$, $p < .001$) for Experiment 1 (standard); ($r = .694$, $p < .001$) for Experiment 2 (silence); and ($r = .83$ $p < .001$) for Experiment 3 (noise). These correlations are within the 95% confidence intervals of the general aggregate composite r-value. Of note, the correlation between sessions was highest for Experiment 3 conducted in environmental noise, suggesting that environmental noise does not have a negative impact on the reliability of the PART test battery.

To evaluate possible differences in threshold between experimental sessions for different listening conditions, t-tests compared between the two sessions of each Experiment. These tests failed to find significant differences in any of the Experiments (Experiment 1, $t(50) = 0.22$, $p = .82$, Cohen's $d = 0.03$; Experiment 2, $t(50) = 0.79$, $p = .49$, Cohen's $d = 0.11$); Experiment 3, $t(47) = -1.27$, $p = .21$, Cohen's $d = -0.18$). Finally, an ANOVA contrasting the mean thresholds obtained on average between sessions for each experiment failed to find any statistically significant differences ($F(2,147) = 0.43$, $p = .64$, $\eta 2 = .006$).

2. No effect of headphone type on estimated thresholds

To examine the effects of headphone type and the presence of environmental noise, data are presented from Experiment 2 (both headphone types in silence) and Experiment 3 (two headphone types in noise). Figure 8 shows the limits of agreement between measures as well as the scatter plots for the silent and noise listening conditions. The agreement analysis between the estimated thresholds using either set of headphones again shows unbiased estimates (Experiment 2 (silence) .004; Experiment 3 (noise) -.008) and similar limits of agreement near 1 SD (Experiment 2 (silence) [-0.9, 0.89]; Experiment 3 (noise) [-0.64, 0.66]) as reported in the general aggregate.

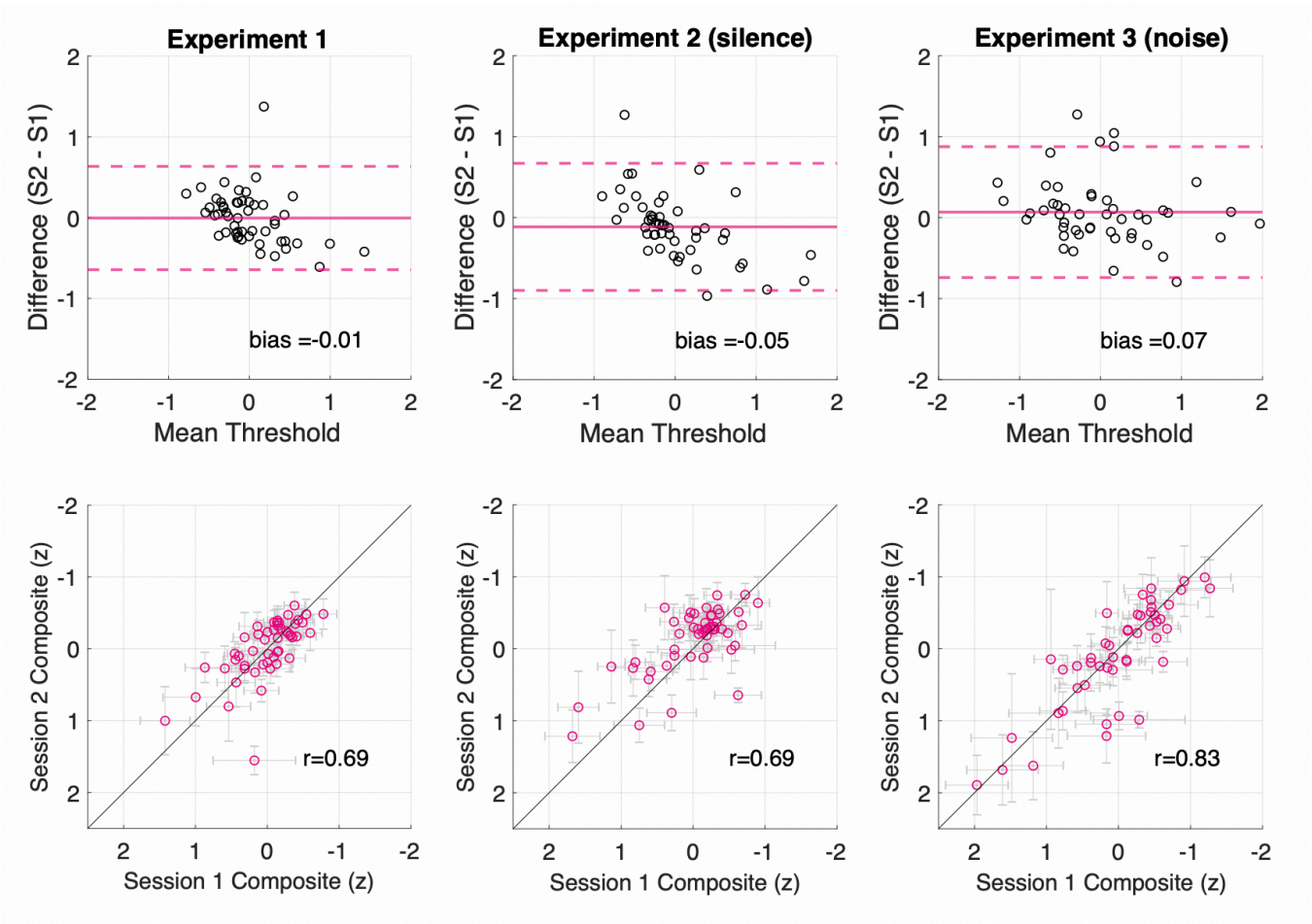

Fig 7. Composite scores for each Experiment. Top panels show the limits of agreement plots. Bottom panels show the composite scatterplots for each experiment. The horizontal error bars indicate SEM for session 1 while the vertical bars reflect session 2 SEM.

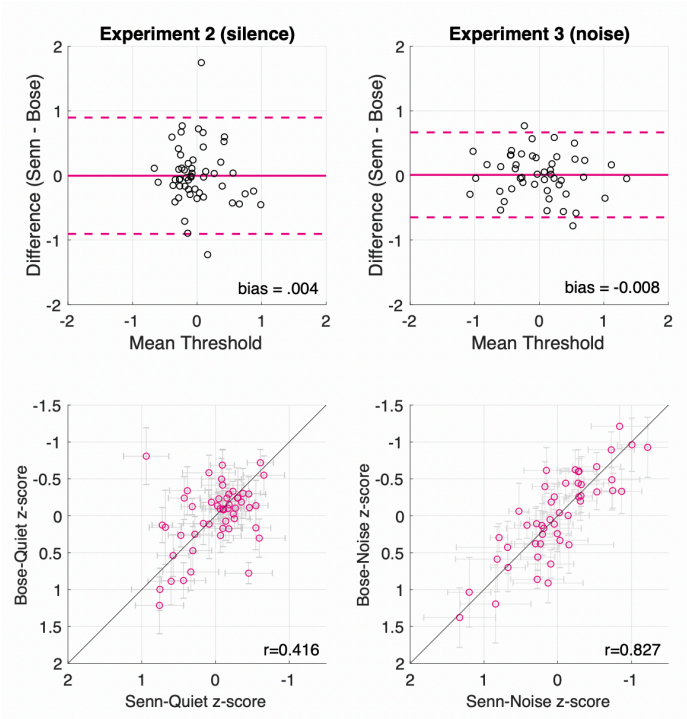

Fig 8. Headphone comparison. Top panels show the limits of agreement plots. Bottom panels show the composite scatterplots relating headphone type used. The horizontal error bars indicate Sennheiser 280 pro headphones.

Composite correlations were also similar to what is reported above, with the correlation for Experiment 2 (silence) ($r = .41$ $p < .001$) suffering due to a reduced between-subject variability and an outlying performance of a single subject (same as above) whose removal yields a correlation of $r = .55$ $p < .001$. A stronger association between measures was found in Experiment 3 (noise) where performance between-subjects is increased in relation to the within-subject variation ($r = .82$, $p < .001$). The LoA and scatterplots for each individual test are reported in the supplemental materials.

This is a notable result, as not only were the headphones different, but also, they shared the same calibration profile, as the tablets were calibrated using the Sennheiser 280 Pro headphones 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 experiments, the settings that produced an 80 dB SPL output for the Sennheiser's were used for the Bose. This meant that even in a silent environment, all of the stimuli were attenuated by 14 dB when Bose headphones were used.

To evaluate possible differences in threshold as a function of headphone type, t-tests were used to compare between the headphone types in each Experiment. Of note, since headphone type was counterbalanced across sessions, these analyses are collapsed by session order. No statistically significant effects were observed in either Experiment (Experiment 2 (silence), $t(50) = -0.07$, $p = .94$, Cohen's $d = -0.01$; Experiment 3 (noise) $t(47) = -0.17$, $p = .86$, Cohen's $d = -0.02$). As an additional test of significance, a 2X2 repeated measures ANOVA with the within subject factor Headphone and the between subjects factor Experiment was conducted to assess headphone effects. Again, no statistically significant effects were found for either Headphone ($F(1,97)= 0.06$, $p= .795$ $\eta2 <.001$) nor for Experiment ($F(2,97)= 0.37$, $p= .54$, $\eta2= .004$), and with no significant interaction ($F(1,97)= 0.79$, $p= .37$, $\eta2= .008$). In summary, the data failed to show any systematic effect of headphone when participants are 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 in output level did not produce noticeable differences in performances for these undergraduate students with hearing in the normal range.

IV. Discussion

This study examined the validity and reliability of a battery of 10 assessments applied to normal-hearing university undergraduate students with the Portable Automatic Rapid Testing (PART) application that aims to evaluate different aspects of central auditory function. Overall, as is revealed by the data in Table 3, the results show that there are few major deviations from what has been reported from laboratory settings in the extant literature and thus portray valid measurements, with the caveat that it is probably advisable to use more than a single track to measure performance for use in research settings. For clinical screening, however, there is good reason to believe that PART could be used with a single adaptive track. This is especially true if the normative values reported here are used rather than the data from the literature, which likely represent well-trained listeners in most cases, which will not be the case in the clinic. Furthermore, our results from Experiment 1 were replicated in Experiments 2 and 3, demonstrating that PART is able to produce consistent threshold estimates across a variety of settings and equipment. In general, the data showed an un-biased or non-systematic variation of the estimates across sessions that held across all levels of performance registered (see LoA plots). The reliability of PART was also assessed with Pearson r, although this measure of reliability was reduced with decreasing between-subject variability, as was to be expected in this sample. When looking at smaller r values in compliment to the LoA plots, it can be seen that these assessments are not less reliable, and the limits of agreement closely resemble other tests with higher r values. Significance testing revealed statistical differences between sessions in some of our tests however, the effect sizes we found are small and can now be considered as test re-test effects in future work.

These results held constant across the three experiments that were set in different external noise conditions (Experiments 1&2 vs 3) and/or used different types of headphones (Experiment 1 vs 2&3) (see figs. 7 & 8). Of note, the correlations were higher for the condition with external cafeteria noise without an impact on the limits of agreement. This is an important result because the advantages of an instrument that is portable, automatic, and rapid in its testing, can only be exploited if accurate measurement can be collected in a variety of settings that deviate from the optimal quiet sound-booth (e.g. while the patient waits for the sound-booth to be available). Here we have shown PART is able to produce threshold estimates of central auditory function assessments that resemble what has been found in the laboratory with untrained listeners in settings as noisy as a university cafeteria. PART can thus be considered as a supplemental tool in the clinic, to collect important information about a person's hearing capabilities while the clinician or sound-booth is available (e.g. in the waiting room). 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.

Another central manipulation in this study was the introduction of the different type of headphone with an active noise-cancelling technology. We were interested in testing this technology because it is now widely available, and little is known about the advantages and disadvantages it could represent for auditory testing. Since these headphones use an active algorithm that senses external noise and computes a cancelling signal, that could potentially aid or distort the perception of the sounds used in our tests. We were interested in evaluating performance both in silent and noisy conditions. We failed to find a statistically significant effect in which differences between headphones used impacted threshold estimation. Estimated thresholds were similar for the Sennheiser 280 Pro in silence (Experiment 2), and this lack of difference manifested similarly in noisy conditions (Experiment 3). 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 stimulation in any way that provokes detectable reductions in performance with the current analysis. Perhaps 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. 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.

We conclude that PART is a reliable platform for testing central auditory processing that 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, as a next step, clinical research is needed to determine how thresholds vary as a function of different types of hearing loss, and to verify that threshold estimates from PART are reliable in hearing-impaired populations. This next step is feasible considering that the PART platform is highly accessible in terms of its monetary cost (it only requires a computer tablet and headphones), time cost (our whole battery of 10 assessments in under 1 hr), human resources cost (it runs the assessments automatically, one after another, including instructions and breaks) and 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 size and variety of psychophysical measures of auditory function necessary to inform clinical practice.

V. TABLES

Table1. Mean thresholds and standard deviations for the 10 assessments utilized plus the derived spatial release metric in PART's native measurement units.

Test	*M*	*SD*	Units	*Mean Difference (S2-S1)*
Gap	2.51	2.9	Gap length (ms)	-1.73
Dichotic FM	0.87	1.25	Modulation Depth (hz)	-0.02
Diotic FM	8.09	7.96	Modulation Depth (hz)	-0.63
TM	1.59	1.08	Modulation depth (dB)	-0.01
SM	1.71	1.12	Modulation depth (dB)	-0.11
STM	1.18	1.03	Modulation depth (dB)	0.15
No-Notch	56.63	2.57	Masker level (dB) (target @ 45dB)	0.69
Notch	75.98	7.88	Masker level (dB) (target @ 45dB)	-0.52
SR Co-located	63.48	2.83	Masker level (dB) (target @ 65dB)	0.4
SR Separated	69.34	3.49	Masker level (dB) (target @ 65dB)	0.99
Spatial Release	5.86	3.57	SR (Sep - Co) (dB)	0.58

Table 2. Significance testing for the 10 assessments utilized at two time points. * indicate significance at = .05

Test	*Bias*	*Limits of Agreement*	Units	*r (p)*	*t (p)*	*df*	*Cohen's d*
Gap	-0.17	[-3.5, 3.16]	Log2 (ms)	.53 (<.01)*	1.22 (.22)	149	0.1
DichoticFM	-0.001	[-2.27, 2.27]	Log2 (hz)	.59 (<.01)*	-0.01 (.98)	149	-0.001
DioticFM	-0.09	[-2, 1.81]	Log2 (hz)	.45 (<.01)*	1.2 (.23)	149	0.09
TM	-0.01	[-2.33, 2.3]	M (dB)	.46(<.01)*	1.57 (.87)	149	0.01
SM	-0.11	[-2.12, 1.9]	M (dB)	.58 (<.01)*	1.33 (.18)	149	0.1
STM	0.15	[-2.16, 2.46]	M (dB)	.38 (<.01)*	-1.56 (.11)	149	-0.12
No-Notch	-0.69	[-6.8, 5.4]	SMR (dB)	.28 (<.01)*	2.7 (<.01) *	149	0.22
Notch	0.52	[-22.02, 23.07]	SMR (dB)	.08 (.34)	-0.55 (.57)	149	-0.04
Co-located	-0.4	[-5.12, 4.3]	SMR (dB)	.64 (<.01)*	2.07 (.04) *	149	0.16
Separated	-0.99	[-7, 5.02]	SMR (dB)	.62 (<.01)*	3.96 (<.01) *	149	0.32
SpatialR	0.58	[-6.63, 7.81]	SMR (dB)	.47 (<.01)*	-1.94 (.05)	149	-0.15

Table 3. Shows a summary of the similarities of the thresholds estimated in the present study using PART and matched psychophysical tests from previous research.

Test	*Closest laboratory test*	*Distance between means in SD units*
Gap	Gallun et al., 2014	0.69
Dichotic FM	Grose & Mamo, 2012; Hoover et al., 2019	0.97; 1
Diotic FM	Grose & Mamo, 2012; Hoover et al., 2019	1.58; 1.59
TM	Viemeister, 1979	0.71
SM	Hoover, Eddins & Eddins, 2005	1.64
STM	Gallun et al., 2018	0.85 (max) 0.75 (min)
No-Notch	Patterson, 1976	1.3
Notch	Patterson, 1976	1.3
SR Co-located	Jakien et al., 2017; Gallun et al., 2018	0.56; 0.3 (max)
SR Separated	Jakien et al., 2017; Gallun et al., 2018	1.02; 0.24 (max)
Spatial Release	Jakien et al., 2017; Gallun et al., 2018	0.34; 0.14 (max)

Acknowledgment

We are grateful to Kasey Jakien and Sittiprapa Isarangura for their kind revisions on the manuscript. This work was funded by NIH NIDCD R01 DC 015051. Equipment and engineering support provided by the VA RR&D NCRAR, the UCR Brain Game Center, and Samuel Gordon (NCRAR). The first author is currently funded by CONACYT and UC Mexus. The views expressed are those of the authors and do not represent the views of the NIH, CONACYT, UC Mexus, or the Department of Veterans Affairs.

References

1. Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. The Statistician, 32(3), 307. https://doi.org/10.2307/2987937
2. 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), 15–22. https://doi.org/10.3109/03005369209077867
3. Bernstein, J. G., G. Mehraei, S. Shamma, F. J. Gallun, S. M. Theodoroff and 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.

4. Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. Statistical Methods in Medical Research, 8(2), 135–160. https://doi.org/10.1191/096228099673819272

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

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

7. Carhart, R., & Jerger, J. F. (1959). Preferred Method For Clinical Determination Of Pure-Tone Thresholds. Journal of Speech and Hearing Disorders, 24(4), 330–345. https://doi.org/10.1044/jshd.2404.330

8. CHABA. (1988). Speech understanding and aging. J. Acoust. Soc. Am. 83, 859–895. doi: 10.1121/1.395965

9. 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, 106(2719), 2719–2732. https://doi.org/10.1121/1.428100

10. 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), 492–503. https://doi.org/10.1121/1.4922700

11. 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), 1220–1234. https://doi.org/10.1152/jn.2001.85.3.1220

12. Eddins, D. A. and J. W. Hall III (2010). Binaural processing and auditory asymmetries. The Aging Auditory System, Springer: 135-165.

13. Füllgrabe, C., B. C. Moore and M. A. Stone (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

14. Gallun, F. and P. Souza (2008). "Exploring the role of the modulation spectrum in phoneme recognition." Ear Hear 29(5): 800-813.

15. Gallun, F. J., A. C. Diedesch, S. D. Kampel and K. M. Jakien (2013). "Independent impacts of age and hearing loss on spatial release in a complex auditory environment." Front Neurosci 7: 252.

16. Gallun, F. J., G. P. McMillan, M. R. Molis, S. D. Kampel, S. M. Dann and D. L. Konrad-Martin (2014). "Relating age and hearing loss to monaural, bilateral, and binaural temporal sensitivity." Front Neurosci 8: 172.

17. Gallun, F. J., Seitz, A., Eddins, D., Molis, M., Hoover, E., Souza, P., … Western, C. (2018). Development and validation of Portable Automated Rapid Testing (PART) measures for auditory research, 33(May).

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

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

20. 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

21. 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

22. 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), 306–309. https://doi.org/10.1121/1.5020787

23. 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), 2018–2034. https://doi.org/10.1044/2019_JSLHR-H-18-0217

24. 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), 1-15.

25. Iliadou, V. (Vivian), 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 Re: Moore, D. R. (2018) Editorial: Auditory Processing Disorder. Ear and Hearing, 39(6), 1236–1242.

26. Isarangura, S., Palandrani, K., Stavropoulos, T., Seitz, A., Hoover, E. C., Gallun, F. J., & 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), 1722-1722. doi:10.1121/1.5101325

27. Jakien, K. M., Kampel, S. D., Stansell, M. M., & Gallun, F. J. (2017). Validating a rapid, automated test of spatial release from masking. American Journal of Audiology, 26(4), 507–518. https://doi.org/10.1044/2017_AJA-17-0013

28. 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), 3503–3523. https://doi.org/10.1152/jn.1996.76.5.3503

29. Marrone N, Mason CR, 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 124:3064–3075. doi: 10.1121/1.2980441

30. Mehraei, G., F. J. Gallun, M. R. Leek and J. G. Bernstein (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): 301.

31. Moore, B. C. J. (1987). Distribution of Auditory-Filter Bandwidths at 2 K{Hz} in Young Normal Listeners. \JournalOfThe\ \Acoust\ \Soc\ of \Ama, 81(5), 1633–1635. https://doi.org/10.1121/1.394518

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

33. 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): 907-922.

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

35. Moore, D. R., Edmondson-Jones, M., Dawes, P., Fortnum, H., McCormack, A., Pierzycki, R. H., et al. (2014). Relation between speech-in-noise threshold, hearing loss and cognition from 40-69 years of age. PLoS ONE 9:e107720. doi: 10.1371/journal.pone.0107720

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

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

38. Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. The Journal of the Acoustical Society of America, 59(3), 640–654. https://doi.org/10.1121/1.380914

39. Pfeiffer, R. R., & Kim, D. O. (2005). Cochlear nerve fiber responses: Distribution along the cochlear partition. The Journal of the Acoustical Society of America, 60(4), 966–966. https://doi.org/10.1121/1.381150

40. Schimmel, O., S. van de Par, J. Breebaart and A. Kohlrausch (2008). "Sound segregation based on temporal envelope structure and binaural cues." Journal of the Acoustical Society of America 124(2): 1130.

41. Schonwiesner, M. and R. J. Zatorre (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): 14611-14616.

42. Shamma, S. (2001). "On the role of space and time in auditory processing." Trends Cogn Sci 5(8): 340-348.

43. Shinn-Cunningham, B. G. (2008). "Object-based auditory and visual attention." Trends Cogn Sci 12(5): 182- 186.

44. Snell, K. B., F. M. Mapes, E. D. Hickman and D. R. Frisina (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): 720-727.

45. Stecker, G. C., & Gallun, F. J. (2012). Binaural Hearing, Sound Localization, and Spatial Hearing. Translational Perspectives in Auditory Neuroscience: Normal Aspects of Hearing, 383–433.

46. 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), 329–334. https://doi.org/10.3109/03005369209076655

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

48. Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. Nature Reviews Neuroscience, 15(6), 355–366. https://doi.org/10.1038/nrn3731

49. 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, 1332-1343.

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

51. 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), 1051–1067. https://doi.org/10.1044/2018_JSLHR-H-18-0045

52. Versnel, H., Zwiers, M. P., & van Opstal, A. J. (2009). Spectrotemporal Response Properties of Inferior Colliculus Neurons in Alert Monkey. Journal of Neuroscience, 29(31), 9725–9739. https://doi.org/10.1523/jneurosci.5459-08.2009

53. Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholdsa). Journal of the Acoustical Society of America, 66(5), 1364–1380. https://doi.org/10.1121/1.383531

54. Von Békésy, G. (2005). Hearing Theories and Complex Sounds. The Journal of the Acoustical Society of America, 35(4), 588–601. https://doi.org/10.1121/1.1918543

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

56. 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), 3093–3104. https://doi.org/10.1121/1.4935018

57. 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), 619–633. https://doi.org/10.1007/s10162-017-0624-x

58. 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), 1826–1833. https://doi.org/10.1121/1.1310195

59. Winer, J. A. (2006). Decoding the auditory corticofugal systems. Hearing Research, 212(1–2), 1–8. https://doi.org/10.1016/j.heares.2005.06.014

60. 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, NY