The Journal of the Acoustical Society of America

AUGUST 09 2023

Web-based psychoacoustics of binaural hearing: Two validation experiments

Z. Ellen Peng ; Emily A. Burg; Tanvi Thakkar; Shelly P. Godar ; Sean R. Anderson; Ruth Y. Litovsky

Check for updates

J. Acoust. Soc. Am. 154, 751–762 (2023)
https://doi.org/10.1121/10.0020567

View Online
Export Citation
CrossMark

21 January 2024 17:37:28

ASA
LEARN MORE
Advance your science and career as a member of the
Acoustical Society of America

ASA

JASA ARTICLE

Web-based psychoacoustics of binaural hearing: Two validation experiments[a)]

Z. Ellen Peng,[b)] Emily A. Burg, Tanvi Thakkar,[c)] Shelly P. Godar, Sean R. Anderson, and Ruth Y. Litovsky
Waisman Center, University of Wisconsin-Madison, Madison, Wisconsin 53715, USA

ABSTRACT:
Web-based testing is an appealing option for expanding psychoacoustics research outside laboratory environments due to its simple logistics. For example, research participants partake in listening tasks using their own computer and audio hardware and can participate in a comfortable environment of their choice at their own pace. However, it is unknown how deviations from conventional in-lab testing affect data quality, particularly in binaural hearing tasks that traditionally require highly precise audio presentation. Here, we used an online platform to replicate two published in-lab experiments: lateralization to interaural time and level differences (ITD and ILD, experiment I) and dichotic and contralateral unmasking of speech (experiment II) in normal-hearing (NH) young adults. Lateralization data collected online were strikingly similar to in-lab results. Likewise, the amount of unmasking measured online and in-lab differed by less than 1 dB, although online participants demonstrated higher speech reception thresholds overall than those tested in-lab by up to ∼7 dB. Results from online participants who completed a hearing screening versus those who self-reported NH did not differ significantly. We conclude that web-based psychoacoustics testing is a viable option for assessing binaural hearing abilities among young NH adults and discuss important considerations for online study design. © *2023 Acoustical Society of America*. https://doi.org/10.1121/10.0020567

(Received 19 December 2022; revised 15 July 2023; accepted 19 July 2023; published online 9 August 2023)

[Editor: James F. Lynch]

Pages: 751–762

21 January 2024 17:37:28

I. INTRODUCTION

Remote testing outside the laboratory has been gaining popularity in recent years as a means to increase study sample size and increase participant diversity by including individuals beyond local communities (Hartshorne *et al.*, 2019; Larrouy-Maestri *et al.*, 2019; Mehr *et al.*, 2019; Müllensiefen *et al.*, 2014; Peretz and Vuvan, 2017). Many behavioral studies have found success in recruiting and replicating studies through web-based experiments, which enable participants to use personal hardware and complete tasks in their chosen environments (Mehr *et al.*, 2017; Mehr *et al.*, 2018; Mehr *et al.*, 2019; Peretz *et al.*, 2008; Swanepoel *et al.*, 2019; Vuvan *et al.*, 2018). However, few research studies in psychoacoustics have been adapted for web-based data collection due to the assumption that high-quality audio equipment is needed to effectively deliver complex auditory stimuli and testing must take place in acoustically controlled environments like sound-proof booths. This results in difficulty recruiting large sample sizes and individuals with unique clinical characteristics (e.g., hearing device users), who may need to travel great distances to participate in research. With the rising need to collect data safely during the COVID-19 pandemic and increase participant diversity, scientists have begun to establish best practices for conducting psychoacoustics studies on web-based platforms including validation studies to replicate experimental effects that were measured in-lab (Peng *et al.*, 2022).

Most psychoacoustic experiments rely on high precision of intensity, temporal, and spectral characteristics in the audio signals to assess phenomena of interest. For binaural hearing tasks conducted over headphones, there are additional requirements on outputs presented to the two ears because they need to align in time and equal in level for faithful representation of interaural timing and level differences (ITD and ILD), which are on the order of ∼10 μs and 1–2 decibels, respectively. Thus, there are three main issues surrounding the quality of data collected via web-based testing, particularly for psychoacoustics of binaural hearing. First, web-based testing using commercial-grade audio hardware raises the concern on variable audio quality and inconsistent delivery of auditory stimuli during experiments. Second, the remote testing locations chosen by individual participants may not be ideal for auditory tasks that require focused attention over a prolonged period of time without supervision or feedback (Peng *et al.*, 2022). A third potential issue arising from online recruitment for web-based experiments is the lack of traditional screen for normal hearing using calibrated signals, such as pure tones at known intensities, which is a common practice for in-lab recruitments.

While these challenges from web-based remote testing may be critical for psychoacoustics, the extent to which they may impact data quality, particularly in binaural hearing tasks is unknown. In fact, many commercial-grade headphones have been verified to have good audio quality, such

[a)]Part of this work was presented during the 2021 Midwinter Conference for the Association for Research in Otolaryngology and the 181st Meeting of the Acoustical Society of America in Seattle.

[b)]Also at: Boys Town National Research Hospital, Boys Town, NE 68010, USA. Electronic mail: Ellen.Peng@Boystown.org

[c)]Also at: Department of Psychology, University of Wisconsin-Lacrosse, La Crosse, WI 54601, USA.

 © 2023 Acoustical Society of America

as flat frequency responses and consistent test-retest reliability documented by independent labs (RTINGS.Com, 2023). Most recently, there have been several published studies utilizing binaural hearing tasks that successfully collected data remotely with varying degrees of control on the audio equipment utilized (Merchant *et al.*, 2021; Milne *et al.*, 2021; Padilla-Ortiz and Orduña-Bustamante, 2021; Lelo de Larrea-Mancera *et al.*, 2022). Task-related attention has been studied to some extent with general best practices available, such as using short instructions and implementing attention checks during the experiment to help participants avoid fatigue and promote on-task attention (Anwyl-Irvine *et al.*, 2020; Gijbels *et al.*, 2021; Milne *et al.*, 2021). Additionally, online auditory screeners have been developed to check proper headphone use by participants in order to further safeguard audio quality and stimulus presentation (Milne *et al.*, 2021; Woods *et al.*, 2017). While these advances are promising, validation studies that explicitly compare data collected online to data collected in-lab are crucial to determine the feasibility of web-based testing for binaural psychoacoustic experiments.

The present study assessed the extent to which web-based remote testing affects binaural hearing performance in normal-hearing (NH) young adults by replicating two studies previously published in the Journal of Acoustical Society of America (JASA). In experiment I, we replicated procedures by Goupell *et al.* (2013) to measure intracranial (i.e., within-the-head) lateralization to a range of ITDs and ILDs. In experiment II, we replicated a similar procedure by Goupell *et al.* (2016) to measure speech reception thresholds in various conditions to quantify dichotic and contralateral unmasking of target speech in the presence of masking speech. These datasets were chosen as the "gold standards" because timing and level differences between the stereo headphone channels were applied to all of the auditory stimuli in both studies, making them appropriate for probing the impact of variable audio quality from commercial-grade hardware on the group level.

We included two groups of participants: (1) verified NH using a hearing screen and (2) self-reported NH, to determine the accuracy of self-report for NH verification. Similar to in-lab studies, participants were recruited through word-of-mouth and job ad postings on university campus. All participants filled out an initial screening questionnaire on Qualtrics to indicate general hearing status and age to determine whether they were eligible to participate. Eligible individuals were sent a link to a second Qualtrics form to obtain online consent. After consenting, participants completed a demographic questionnaire and then received a unique link via email to the online experiment hosted on Gorilla.sc. We chose to deploy all online experimental tasks on Gorilla.sc for two reasons. First, Gorilla.sc has established technical details that meet general Institutional Review Board (IRB) and data privacy compliance. Second, the platform provides various options for easy experimental building, with published studies supporting good data quality (Anwyl-Irvine *et al.*, 2020; Milne *et al.*, 2021).

We anticipated several factors that may contribute to differences in group-level lateralization and unmasking measured online versus in-lab. First, imbalanced outputs between commercial-grade headphone channels may shift the lateralization curve, such that 0 dB ILD stimuli are biased toward one side if the output channels contain an intensity difference. Imbalanced intensities may also reduce or increase the perceived intracranial spatial separation between target and masker speech, resulting in an inaccurate measure of unmasking. Similarly, if playback timing between headphone channels is not aligned, this has the potential to bias ITD lateralization curves. Second, despite the instruction to choose a quiet room for the experiment, unexpected background noise or distractions may affect stimulus audibility and on-task attention that could lead to worse or more variable performance among participants. Last, the absence of a traditional hearing screen may result in the inclusion of participants with poorer hearing sensitivity that did not represent typical binaural hearing abilities.

II. GENERAL METHODOLOGY

A. Experimental setup

All online experiment tasks were built on the Gorilla.sc platform using the "Code Editor Tasks" functions, which allowed for custom JavaScript codes to implement complex procedures, such as adaptive staircase tracking for threshold measurements. All tasks were self-guided. For all online testing, participants were asked to provide their own equipment, including computers and headphones. We limited hardware use to either tablets or computers and software use to Chrome, Firefox, or Safari web browsers. These limitations were enforced by built-in functionality on Gorilla.sc to detect the hardware and software when participants accessed the experiment. We asked all participants to use wired headphones or earphones.

B. Participants and recruitment

Participants were instructed to complete the entire online experiment in a quiet room. Each participant was paid $10/h for their time completing the study; payment was delivered via check sent by postal mail. All experimental protocols, including testing on Gorilla.sc, were approved by the Institutional Review Board at University of Wisconsin-Madison.

C. General procedure

Participants were provided instructions about the hardware and software in the study emails prior to testing. Once they opened the study URL, participants were prompted to go through a multi-step perceptually based procedure to set presentation volume and verify headphone quality.

Step 1—System volume setting: Participants were asked to set the initial system volume to ∼20% on their computer. They were then instructed to play the "Carrot Passage" from Verifit (Audioscan, Dorchester, Canada) and

21 January 2024 17:37:28

https://doi.org/10.1121/10.0020567

adjust the system volume until it was playing at a "comfortably loud" level. Once the level was reached, participants were asked to fix the system volume setting and not to re-adjust for the remainder of the experiment. For each participant, this step set the "comfortable level (CL)" which was also the 0 dB reference level in tasks adaptively changing level from trial-to-trial.

Step 2—Headphone stereo balancing: This step was used to screen for volume imbalance between the stereo channels in participants' headphones. Participants were asked to choose one of the eight audio files that produced an intracranial image closest to the middle of their head. Half of the audio files were louder in the left channel by 1, 3, 6, and ∞ (i.e., sound was played only to the left ear) dB; the other half were louder in the right channel. If participants identified stereo imbalance of ± 3 dB, a correction of 3 dB was applied accordingly to the web browser stereo output via Web Audio API for all subsequent audio presentations. If ± 1 dB stereo imbalance was identified, participants proceeded to the task with no correction. If the stereo imbalance was greater than or equal to ± 6 dB, the participant failed this step and was asked not to participate further due to the inability to reliably correct for such a large imbalance and the possibility of additional distortions from the audio hardware. While it was possible that the stereo imbalance found in this step was due to asymmetrical hearing sensitivity of a participant, the correction (up to 3 dB) ensured that participants always received a centered image for zero ITD and ILD regardless of the source of level imbalance between ears.

Step 3—Maximum level check: Participants were asked to listen to the "carrot passage" again and use a slide bar on the screen to adjust the volume (i.e., –10 to +15 dB) to a "loud but okay" level. The slide bar was coded in JavaScript to control the web browser audio volume using the Web Audio API (Mozilla, 2023). This step was used to identify the "uncomfortable level (UCL)" beyond the CL in Step 1. There were no minimum distance or steps in the slide bar required to set the UCL beyond the CL. All tasks involving trial-to-trial level adjustment were checked for audio output to be under the UCL to avoid discomfort or distortion in the signal. If the volume output was prompted to be above the UCL for any trial, the actual playback volume was reset to UCL as a hard limit.

Step 4—Headphone screen: Participants were asked to perform the task outlined by Woods *et al.* (2017) that screens for proper headphone use. On each trial, participants heard three intervals of binaurally presented 200 Hz pure tones, one of which had a 180° phase shift, and were asked to identify which of the three intervals sounded the softest. The target interval with the 180° phase shift was easily detectable when participants were properly wearing headphones but was much more difficult to detect when listening over loudspeakers. Thus, poor performance on this task suggested that the participant was likely listening over loudspeakers rather than headphones. Participants passed this step by correctly identifying five or more out of six trials. If they scored less than five trials, they could repeat the task up to three times. If a participant failed the task, they were automatically prompted to the debrief screen and asked not to participate further in the study.

Once a participant completed the system level calibration and passed the perceptual headphone screening, they were automatically prompted to start the main experiment. For each participant, the entire self-guided protocol typically took no more than 30 min for experiment I and 60 min for experiment II to complete, including instructions, screening, and main experiment with short breaks. Once the protocol was completed, a debriefing screen appeared to thank them for their participation and provide instructions to exit the experiment by closing the browser tab.

D. Analysis

All statistical analyses were conducted using R (version 3.6.0; R Core Team 2022). As a general approach, we first checked data distribution against the assumptions of parametric ANOVA, including normal residual via the Shapiro-Wilk test and homogeneity of variance via the Levene test ("car" package). When data failed to meet assumptions, non-parametric models were fit to the data. When non-parametric models were used, we also screened for homogeneity of variance to determine the need to apply additional statistical transformations on the data. Mixed-effects models ("lme4" package, v1.1–21 and "lmerTest," v3.1–0) were implemented by including a random effect of the listener, with other fixed effects as appropriate.

III. EXPERIMENT I: LATERALIZATION

For this experiment, we replicated a binaural hearing task by Goupell *et al.* (2013) which assessed intracranial lateralization to ITDs and ILDs. By comparing with data collected in-lab, we assessed the impact of a web-based testing protocol on sound lateralization at supra-threshold levels.

A. Methods and procedure

1. Participants

Fifty NH adults participated in the experiment. Five participants were excluded from the final analysis: Three participants were excluded from the study after failing the online perceptual screen and two participants aborted the study early with incomplete datasets. The remaining 45 participants with complete datasets formed two listener groups: NH as verified either by pure tone audiometry [n = 22; mean age = 21.6 years, standard deviation (SD) = 2.8] or by self-report (n = 23; mean age = 20.2 years, SD = 1.4). The former had audiograms previously collected in the laboratory and some experience with in-lab psychoacoustic research participation. All listeners with verified NH status had pure tone thresholds $\leq$ 20 dB HL from 250 to 8000 Hz. Each participant was randomly assigned to be tested with ITD cues or ILD cues in a lateralization task using the integrated random generator function on Gorilla.sc. Approximately the

21 January 2024 17:37:28

same number of participants from each group were tested on ITD and ILD lateralization tasks. Nine participants had a 3 dB correction applied to one of the headphone channels as a result of the pre-experiment perceptual headphone screen. Of these nine participants, six individuals were in the self-reported NH group.

2. Stimuli

Stimuli were 300 ms transposed tones with a 4 kHz carrier tone and a 125 Hz envelope modulation. ITDs were imposed by shifting the whole waveform in either the left or right channel. With such stimulus design, previous experiments have demonstrated that listeners rely upon the envelope for lateralization (Bernstein and Trahiotis, 2002). ILDs were imposed by attenuating level in one ear. Various ITDs (0, ±100, ±200, ±400, or ±800 μs) and ILDs (0, ±1.5, ±3, ±6, ±9 or ±15 dB) were tested. Positive values of ITDs or ILDs are defined as leading (ITD) or having a higher level in the right ear (ILD), with negative values defined to be leading or have a higher level in the left ear, respectively. Fifteen trials were tested for each cue magnitude, resulting in a total 150 trials presented in randomized order for the ITD task, and 180 trials presented in randomized order for the ILD task.

Note that the stimulus was different than that used in Goupell *et al.* (2013), in which Gaussian-envelope tone (GET) pulses were presented at a rate of 100 Hz. Windows systems are known to impose additional processing on audio with sharp onset. We chose transposed tones due to the slightly more gradual onset compared to the much sharper GET pulses because of potential system-related distortion to the signals depending on participants' choice of hardware. Further, GET versus transposed tones at 125 Hz modulation rate were shown to have similar lateralization curves even among NH school-age children (Ehlers *et al.*, 2016). There were no additional methodological deviations in the stimulus creation between the study design in this study and that used by Goupell *et al.* (2013)

3. Procedure

Once participants passed the perceptual screening steps, they were provided instructions on the lateralization task. They were presented with sample ITDs or ILDs (corresponding to the cue being tested) in a directional sequence from left to right, then right to left; a sequence of all the test magnitudes in the main task was included. To provide a visual reference, a cartoon image of a head with a horizontal blue shaded bar between ears was displayed with an arrow to indicate the direction of the ITD or ILD sequence (i.e., from far left to far right then back). This was intended to familiarize participants with the full range of cue magnitudes to be tested and instruct participants as to the direction they should attend to. During the experiment, the same cartoon image with the blue shaded bar was always visible but without the reference arrow. On each trial, participants self-initiated the presentation of a single interval of ITD or ILD and indicated the perceived intracranial position along the blue shaded bar. Responses on the horizontal blue bar were coded proportionally between −1 (at the left ear) and +1 (at the right ear), with 0 being at the center of the head. For each participant, all stimuli were presented at the fixed comfortable level that was self-identified during the perceptual screening. Because the pre-experiment perceptual screen did not measure individual participants' full dynamic range of sensation level, level roving was intentionally excluded during online testing in an effort to ensure audible and safe presentation levels.

B. Results

A lateralization curve was fitted to each participant's data using the MATLAB curve-fitting toolbox and the following equation:

$$\text{Position} = \frac{\text{Upper Bound} - \text{Lower Bound}}{1 + e^{\beta(\alpha - x)}} + \text{Lower Bound}. \quad (1)$$

Equation (1) allowed for optimization of upper and lower bound and produced three parameter estimates of interest for each curve fit: shift α, slope β, and range. Shift is the cue magnitude derived at the center of the head or the zero intracranial position. Slope is the change of intracranial position per unit change of binaural cue. Range is the distance between upper and lower bounds. Figure 1 illustrates individual lateralization curve fits to participants in Goupell *et al.* (2013) and the two groups tested in the present experiment using Gorilla.sc.

Two participants tested on Gorilla.sc showed reversed ITD lateralization curves (B23 and C28) and four showed reversed ILD lateralization curves (B21, C11, C22, and C27). Individual curve fit was assessed using R^2, which was computed based on the sum of squared errors between the participant response and the predicted lateralization values. The average R^2 among online participants was 0.94 with a standard deviation of 0.15. Five out of these six participants had an R^2 between 0.71 and 0.93, which is within two SD of the in-lab participants, except for participant C11 who had a $R^2 = 0.24$. This suggests good model fit to the raw trial data among the five participants identified. Because none of the steps during the perceptual screen confirmed the correct headphone channel placement on the corresponding ears, it is likely that these participants had worn the headphones flipped with the left/right channels on the wrong ears during testing. We recoded these six listeners' responses by multiplying their perceived intracranial positions by a factor of −1.

Figure 2 shows the averaged ITD and ILD lateralization curves for each group with shaded area indicating the 98.3% confidence interval, equivalent to maintaining a family-wise 95% confidence interval by applying the Bonferroni correction. When comparing the shaded 95% confidence interval between the online data vs in-lab data from Goupell *et al.* (2013), in-lab participants demonstrated shallower lateralization curves on average, with intracranial positions closer

21 January 2024 17:37:28

https://doi.org/10.1121/10.0020567

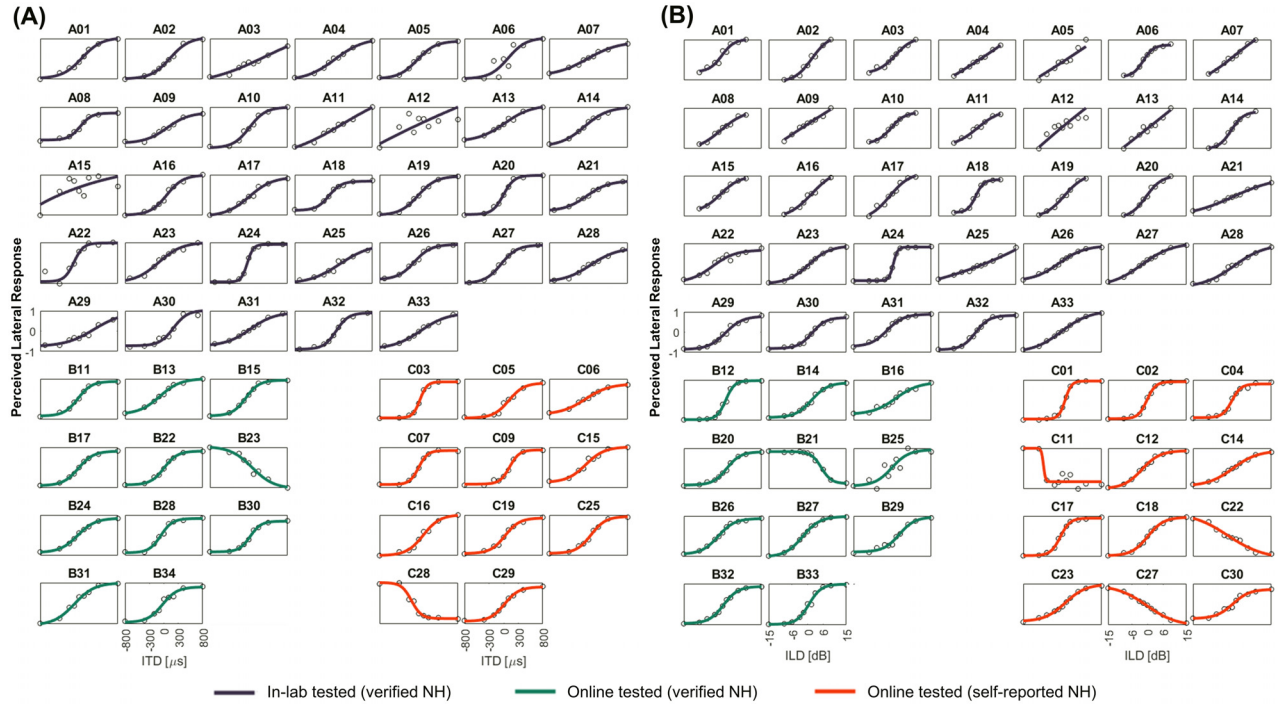

FIG. 1. (Color online) Individual data showing lateralization to interaural time differences (A) and to interaural level differences (B). Data for in-lab tested listeners in black are replotted from Goupell *et al.* (2013) and Anderson *et al.* (2019) with permission. Individual curves for participants tested on Gorilla.sc are plotted in green for those with verified NH and in red for those with self-reported NH.

to the center of head across all ITD magnitudes than either of the online groups. For ILD lateralization, the three listener groups showed largely overlapping intracranial positions along the measured magnitudes.

We further compared the parameter estimates derived from lateralization curves. Figure 3 illustrates the shift, slope and range calculated from individual lateralization curves. Two participants, A12 and A15 from the in-lab

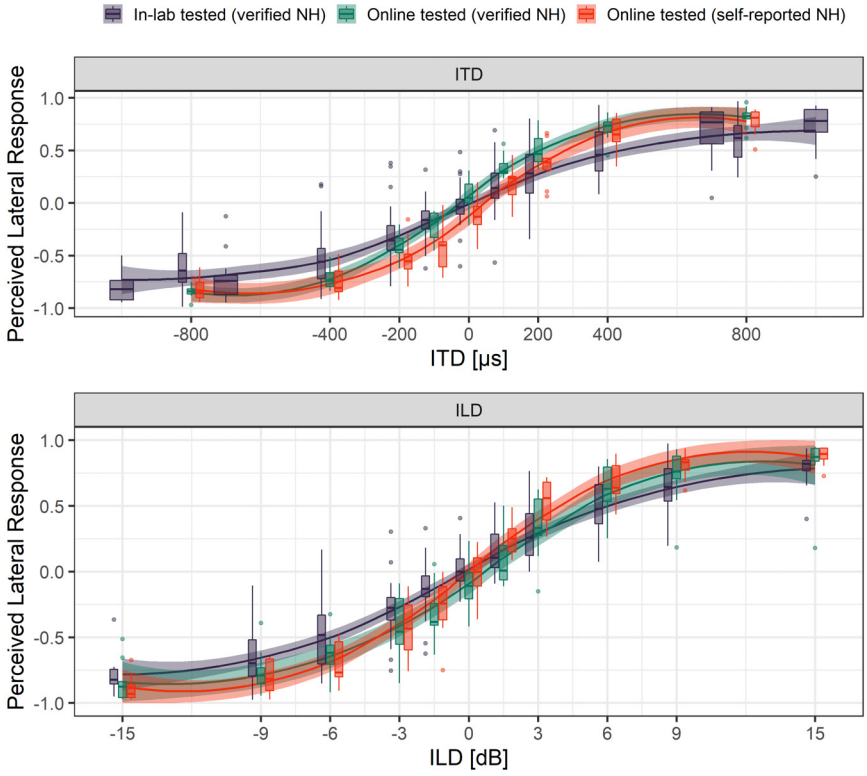

FIG. 2. (Color online) Averaged lateralization curves to interaural time and level differences for each listener group. Data for in-lab tested listeners in black are replotted from Goupell *et al.* (2013) and Anderson *et al.* (2019) with permission. Shaded area indicates 98.3% confidence intervals around the curve (for maintaining a family-wise 95% confidence interval). Listeners with flipped lateralization curves are excluded.

21 January 2024 17:37:28

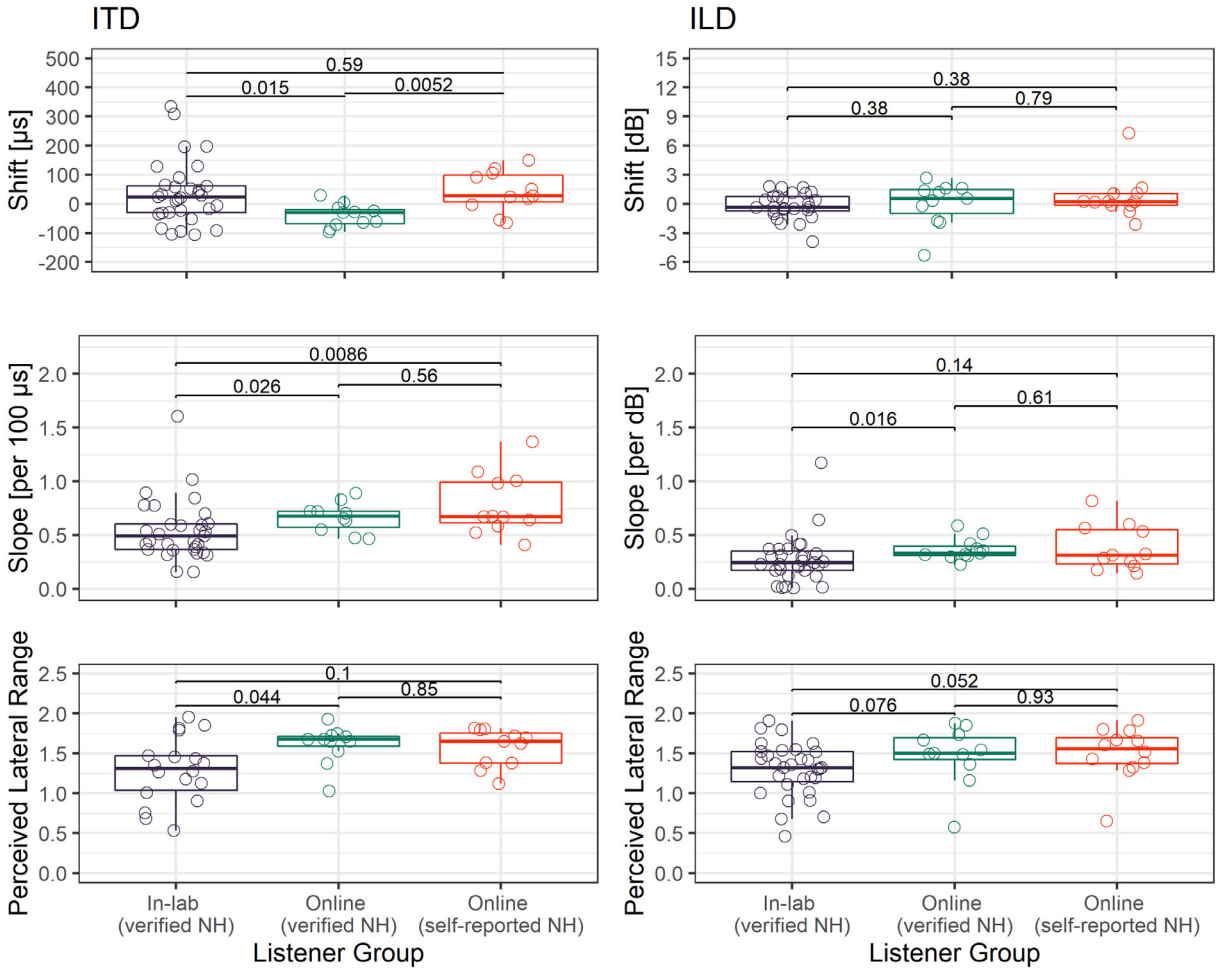

FIG. 3. (Color online) Boxplot overlaid with individual data of parameter estimates derived from the individually fit lateralization curve for ITDs (left column) and ILDs (right column) for the three listener groups. Data for in-lab tested listeners in black are replotted from Goupell *et al.* (2013) and Anderson *et al.* (2019) with permission. Listeners with flipped lateralization curves are included with corrected parameter estimates. Shift is the cue magnitude derived at the center of the head or the zero intracranial position. Slope is the change of intracranial position per unit change of binaural cue. Intracranial range is the span between ±800 μs for ITD and ±9 dB for ILD. Uncorrected p-values from pairwise Wilcoxon Mann Whitney tests are listed for each pair.

group, were removed from the analysis on ITD lateralization due to poor curve fitting that resulted in the shift estimated at $>1000\,\mu$s ITD. Because an ITD $\gtrsim 750\,\mu$s is not physically possible for human listeners, any ITD shift beyond this value is not meaningful (Hartmann, 2021). Data were first fit using a mixed-effects ANOVA including fixed-effects of group and cue (ITD or ILD). Results showed that model residuals were non-normally distributed but variance was homogeneous across groups for all three parameters. A non-parametric ANOVA showed that there was a significant main effect of group on ITD [$\chi^2(2) = 8.04$, $p = 0.018$] but not ILD [$\chi^2(2) = 2.25$, $p = 0.116$] for the shift parameter. As a follow-up, pairwise Wilcoxon Mann Whitney tests were performed to compare each parameter estimate among the three listener groups for ITD and ILD cues separately. After Bonferroni corrections,[1] several significant pairwise comparisons were identified. For lateralization to ITDs, the online listeners with verified NH on average (Median = –30.6 μs, IQR = 48.4) had a significantly different left-ward shift than the other two listener groups (vs In-lab: Median = 23.3 μs, IQR = 92.8, W = 255, Bonferroni-corrected $p = 0.005$; vs online tested with self-reported NH: Median = 27.2 μs, IQR = 91.2, W = 19, corrected $p = 0.015$). Further, a Wilcoxon Sign Rank test comparing each group mean to zero suggested that online tested, NH verified listeners' left-ward shift was significantly away from the zero intracranial position ($V = 6$, $p = 0.014$) but not for participants tested in-lab or those tested online with self-reported NH (both p's > 0.05). The average shift of ILD lateralization curves was −0.36, 0.56, and 0.21 dB for the in-lab, online NH verified, and online NH self-reported groups, accordingly. Further, there was no significant left- or right-ward shift away from the zero intracranial position for any of the groups.

Next, we assessed lateralization slope. Figure 2 shows the group-average lateralization curves between three listener groups. For ITD, there was a trend that the in-lab participants produced shallower lateralization than the two online groups; but for ILD, the lateralization curves show much larger overlaps between groups. A non-parametric ANOVA showed that there was a significant main effect of group on ITD [$\chi^2(2) = 7.87$, $p = 0.020$] and ILD

21 January 2024 17:37:28

https://doi.org/10.1121/10.0020567

[$\chi^2(2) = 7.16$, $p = 0.028$] for the slope parameter. By comparing the calculated slopes among listener groups, we confirmed that in-lab participants (Median = 0.49 per 100 μs, IQR = 0.24) had significantly shallower ITD lateralization curves than online participants with self-reported NH (Median = 0.65, IQR = 0.21, W = 80, corrected $p = 0.027$) but not online participants with verified NH (Median = 0.67, IQR = 0.38, W = 113, corrected $p = 0.31$). However, the averaged slope of ITD lateralization curves did not differ significantly between the two groups of online participants (corrected $p > 0.05$). Further, in-lab participants (Median = 0.24, IQR = 0.18) had significantly shallower ILD lateralization than the online group with verified NH (Median = 0.33, IQR = 0.09, W = 87, corrected $p = 0.048$) even though ILD lateralization slopes did not significantly differ between the two online groups (online listeners with self-reported NH, Median = 0.32, IQR = 0.33, corrected $p > 0.05$).

The third parameter we assessed was lateralization range. A non-parametric ANOVA showed that there was no effect of group on ITD [$\chi^2(2) = 4.56$, $p = 0.102$] or ILD [$\chi^2(2) = 5.28$, $p = 0.071$] lateralization range. All other between-group comparisons of fitted parameters for lateralization to ITDs and ILDs yielded non-significant results (all corrected p's > 0.05).

To verify the overall effect of remote testing on the three lateralization parameters, we re-analyzed the data by combining the two remote groups as one and compared them with those from in-lab participants. Comparing in-lab versus remote testing, the statistical results between-group difference were upheld for lateralization shift [$\chi^2(2) = 7.13$, $p = 0.0076$ significant for ITD, and $\chi^2(2) = 0.14$, $p > 0.05$ non-significant for ILD] and the lateralization slope [$\chi^2(2) = 7.14$, $p = 0.0076$ significant for ITD, and $\chi^2(2) = 7.02$, $p = 0.0080$ significant for ILD]. However, the non-significant group effect on lateralization range became significant comparing between in-lab versus remote testing for both ITD [$\chi^2(2) = 4.46$, $p = 0.035$] and ILD [$\chi^2(2) = 5.13$, $p = 0.023$]. Note that in the online testing experiment, we provided a practice phase where the full range of ITD/ILD stimuli were demonstrated to all participants, whereas the in-lab participants did not go through such practice which likely explained this change. Overall, this finding supports the notion that the effects were not driven by differences in either remote group and that separating them into two groups did not underpower the analysis to detect the underlying effect.

C. Discussion

This experiment tested lateralization of high-frequency transient pulse trains on a web-based platform for young adults with verified and self-reported NH. Results showed remarkably overlapping lateralization curves from online and in-lab participants (Anderson *et al.*, 2019; Goupell *et al.*, 2013). On the group-level, when compared with in-lab data, lateralization curves measured online were steeper for both ITD and ILD cues. Further, besides a left-ward shift observed for ITD cues, NH verification (i.e., verified vs self-reported) for young adults tested online did not produce substantial differences in lateralization curves.

Prior studies have shown that NH listeners demonstrate ITD sensitivity in the signal envelope with just-noticeable-difference (JND) thresholds between ~70 and 180 μs (Anderson *et al.*, 2019; Bernstein and Trahiotis, 2002; Ehlers *et al.*, 2016; Goupell *et al.*, 2013; Peng *et al.*, 2020). The average shift to ITD lateralization demonstrated by all three groups of listeners between −30.6 and 27.2 μs, including those tested online, is likely undetectable from the zero intracranial position. Further, the sampling frequency of the stimuli was 44.1 kHz or ~23 μs between any two samples. Most commercial audio hardware samples at 44.1 kHz or lower, such that any ITD < 23 μs would not have been represented in the physical signals presented to listeners tested online. Even though both groups of online participants showed higher ITD shift away from the zero intracranial position than in-lab participants, the magnitude of up to ~30 μs off center on average was in fact rather small and less than two audio samples. Interestingly, a consistent, stimulus-irrelevant shift toward one side of the head has been observed in previous laboratory experiments (Goupell *et al.*, 2021), which suggests that the difference in shift observed in web- vs laboratory-based experiments may have occurred simply due to sample size. With ILD lateralization, the average shifts of < 0.6 dB were also likely perceptually undetectable across all listener groups, as the magnitudes are much smaller than previously reported ILD JNDs between ~1–3 dB (Ehlers *et al.*, 2016; Goupell *et al.*, 2013).

In-lab participants demonstrated a shallower average slope of the psychometric function than one group of online participants in lateralization to ITD but not ILD. This is likely due to the larger overall sample size from the two in-lab studies (Goupell *et al.*, 2013 and Anderson *et al.*, 2019), where the individual psychometric functions were more variable under ITD [see Fig. 1(A)].

Note that there is large variability in the intracranial ranges to both ITD and ILD cues among participants tested in-lab (Fig. 3). Online testing was able to capture some of these individual variabilities, particularly for the higher end of intracranial ranges. While this may be due to smaller sample sizes in the two online groups, the practice phase for online participants to scan all ITD or ILD magnitudes prior to testing might have prompted them to map larger magnitude cues to more extreme intracranial positions.

IV. EXPERIMENT II: SPEECH IN SPEECH RECOGNITION

For this experiment, we replicated the binaural listening task by Goupell *et al.* (2016) which assessed spatial attention in NH listeners. By comparing with the data collected in-lab, we assessed the impact of a web-based testing protocol on the signal-to-noise ratio corresponding to a 50% correct speech reception threshold (SRT), and the improvement

21 January 2024 17:37:28

in SRT resulting from a perceived spatial separation between the target and masker speech.

A. Methods and procedure

1. Participants

Fifty-seven NH adults participated in the experiment. Seventeen participants were excluded from the final analysis; five participants failed the online perceptual screen and 12 participants aborted the study early with incomplete datasets (six in a specific test condition with monaural listening of same-sex target and masker, six aborted during other steps). The remaining 40 participants with complete datasets formed two listener groups: NH verified by pure tone audiometry ($n = 20$; mean age = 25.7 years, SD = 5.3) and by self-report ($n = 20$; mean age = 20.0 years, SD = 1.6). All listeners with verified NH status had pure tone thresholds ≤ 20 dB HL from 250 to 8000 Hz. Seven participants had a 3 dB correction applied between ears during SRT measurements due to a channel imbalance identified during the perceptual headphone screen. Of these seven participants, five individuals were from the self-reported NH group.

2. Stimuli

The same speech stimuli from Goupell *et al.* (2016) were used for the online experiment, which consisted of five-word sentences that were comprised of combinations of a name, verb, number, adjective, and object (Kidd *et al.*, 2008). There were eight possible names, verbs, numbers, adjectives, and objects. On each trial, one word was randomly selected from each category to create a five-word sentence for the target talker (e.g., "Jane took two new toys"). At the same time, another five-word sentence was created from the remaining pool of words for the masker talker. We introduced two masker conditions: different-sex versus same-sex as the target talker. The different-sex masker data were used to compare with data collected in-lab (Goupell *et al.*, 2016), whereas the same-sex masker data provided additional opportunities to examine the effect increased informational masking.

The target talker was a female with a fundamental frequency (F0) of 182 Hz and the male different-sex masker had a lower F0 of 103 Hz. These were the same stimuli used by Goupell *et al.* (2016). The same-sex masker was another female who had slightly higher F0 of 193 Hz. All speech tokens were root mean square normalized to the same level. Each target and masker token under the same category was pre-processed to have the same duration, with the utterances time-aligned at the mid-point and zero-padding around the onset and offset.

3. Procedure

Listeners in the verified NH group were audiologists or audiology students who had routine hearing screens performed in a clinic or verified in the lab as part of the study; listeners in the self-reported NH group were screened from the recruitment survey. After the participant passed the perceptual headphone screen, the experimental protocol on Gorilla.sc automatically directed them to the main experiment. Participants were tested in six conditions, including a quiet condition and five conditions with an interfering masker talker. The target was always in one ear, randomly chosen for each participant. Three conditions were tested with a different-sex masker: (1) monaural, masker ipsilateral to or in the same ear as the target, (2) contralateral, masker in the opposite ear as the target, and (3) dichotic, masker in both ears. Two conditions were tested with a same-sex masker: (1) monaural, masker ipsilateral to the target and (2) dichotic, masker in both ears. The quiet condition was always tested first to familiarize participants with the target talker's voice. The conditions with an interfering masker were presented in blocks by the masker sex, with the test blocks randomized between participants and condition order randomized within each block. For each trial, participants were instructed to choose from a 25-word matrix on the web-browser screen containing all possible words to form the five-word target sentence while ignoring the masker. Participants could not respond until after the audio finished playing and the response buttons were activated. The masker was always fixed at –3 dB (re CL). For each test condition, the target was presented at 0 dB (re CL) for the first trial and then the level was changed from trial-to-trial following a one-down-one-up adaptive procedure (Levitt, 1971) to identify the SNR corresponding to 50% correct. The initial step size was 8 dB, then reduced to 4 and 2 dB after the first and second reversal. The adaptive track terminated after the participant reached six reversals. For each participant, a total of six SRTs (one for each condition) were measured in the self-guided experiment. Once they completed all testing, Gorilla.sc automatically directed them to a debriefing screen.

Note that the adaptive tracking method used for measuring threshold was different than the method of constant stimuli used by Goupell *et al.* (2016). The choice to use adaptive procedures for threshold measurement was made based on the need for an automated testing protocol within a shorter testing duration. To maintain consistency in threshold estimation between studies, we used the same maximum-likelihood estimation algorithm [i.e., psignifit (Frund *et al.*, 2011)] as in Goupell *et al.* (2016) to extract SRT thresholds. For each participant, unmasking was calculated as the improvement in SRT (i.e., reduction) from the monaural condition to (1) the contralateral separation condition with different-sex masker, (2) the dichotic separation condition with different-sex masker, and (3) the dichotic separation condition with same-sex masker.

B. Results

Figure 4 illustrates the SRT thresholds (re masker level) for the three listener groups under each test condition. Data were first analyzed using a mixed-effects ANOVA with

21 January 2024 17:37:28

https://doi.org/10.1121/10.0020567

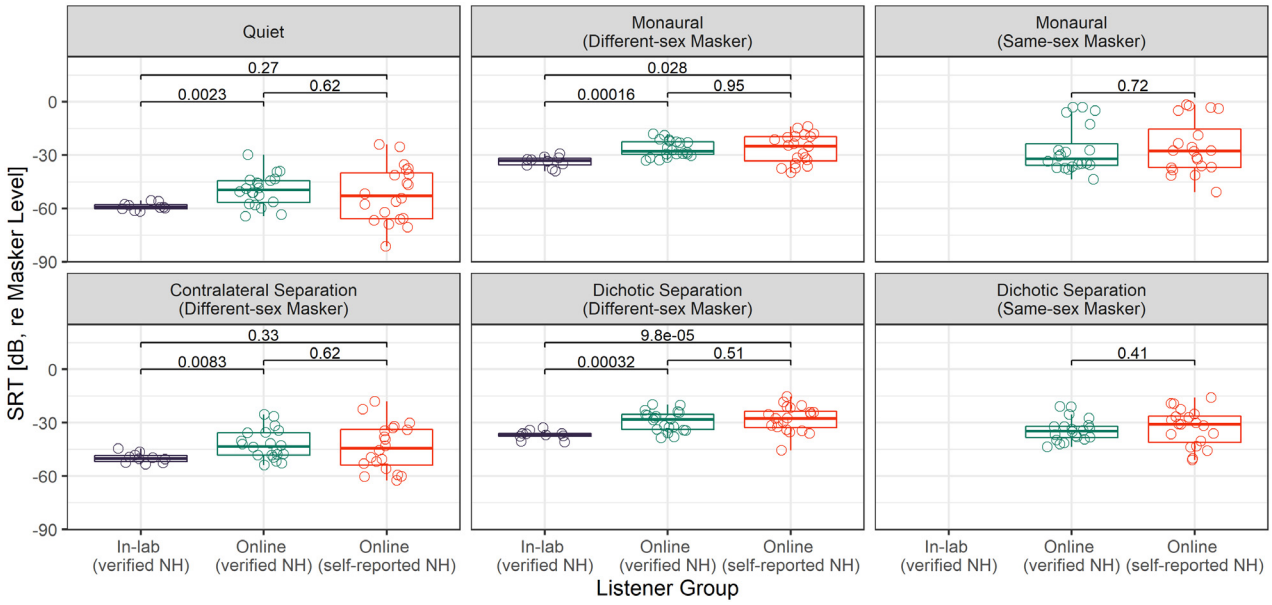

FIG. 4. (Color online) Boxplots with individual data showing SRT (re masker level) at 50% correct for three listener groups. Data for in-lab participants with verified NH are replotted from Goupell *et al.* (2016) with permission. Masker levels were presented at 70 dB SPL (re 20 μPa) in Goupell *et al.* (2016) and at –3 dB full-scale [re self-identified comfortable level (CL)] for the two online listener groups in the present study. Uncorrected p-values from pairwise Wilcoxon Mann Whitney tests are listed for each pair.

fixed-effects of group and test condition. Results revealed significant deviations from the assumption of normality and equal variance, but only between test conditions and not listener groups. Since the focus was on comparing group-level differences, no transformation on SRT was necessary. A non-parametric ANOVA revealed a significant effect of group [$\chi^2(2) = 13.23$, $p = 0.0013$]. Non-parametric Wilcoxon Mann Whitney tests were conducted to compare the SRT distributions between each pair of listener groups. In-lab participants had significantly lower SRTs than online participants with verified NH across all test conditions: Quiet[2] ($W = 33$, corrected $p = 0.0069$), Monaural Different-sex ($W = 20$, corrected $p < 0.001$), Contralateral Separation Different-sex ($W = 41$, corrected $p = 0.025$), Dichotic Separation Different-sex ($W = 23$, corrected $p < 0.001$). In contrast, SRTs for in-lab participants were similar to self-reported NH participants for all conditions except Dichotic Separation Different-sex ($W = 18$, corrected $p < 0.001$). Between the two online groups, there was no significant difference in SRT distributions between any test conditions (all corrected p's > 0.05).

Figure 5 illustrates the relationship between SRTs in the masker conditions versus in quiet. In contrast to the in-lab participants, SRTs in quiet and masker conditions were

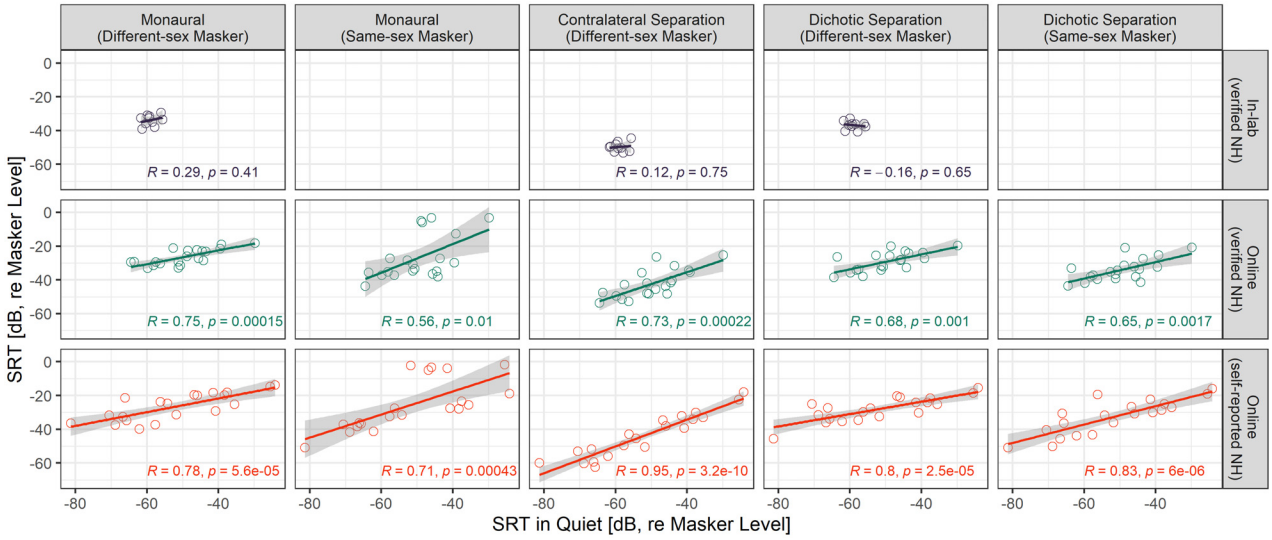

FIG. 5. (Color online) Scatter plots showing SRTs in each masker condition as a function of SRTs in Quiet for the three groups of listeners. Data for in-lab participants with verified NH are replotted from Goupell *et al.* (2016) with permission.

21 January 2024 17:37:28

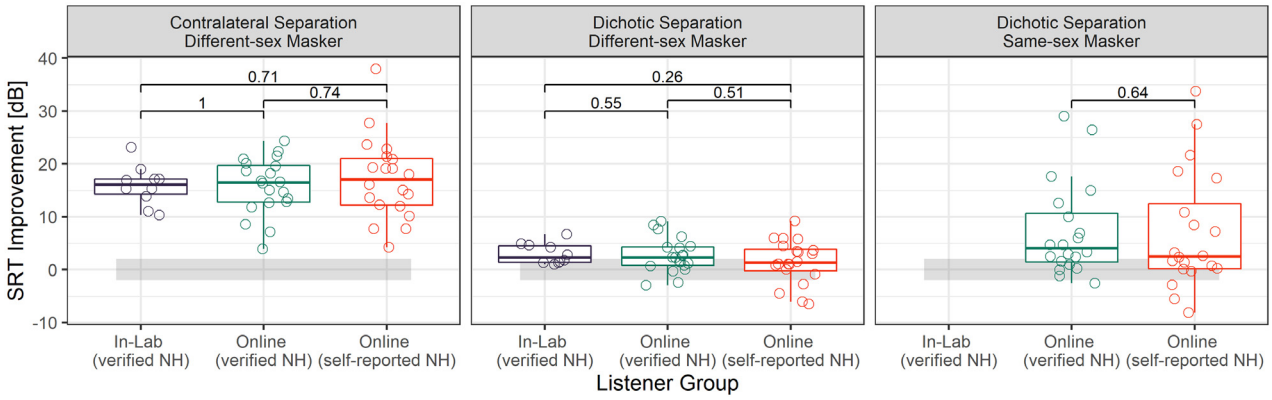

FIG. 6. (Color online) Boxplots with individual data showing unmasking as SRT improvements for the three listener groups under each masker separation condition, as compared to the monaural condition. Data for in-lab participants with verified NH are replotted from Goupell *et al.* (2016) with permission. Uncorrected p-values from pairwise Wilcoxon Mann Whitney tests are listed for each pair. The shaded area indicates ±2 dB improvement.

highly correlated for both groups of online participants. The correlation R-value between quiet SRT and SRT under masker conditions ranged from 0.58 to 0.83 (all p's < 0.05, individual p-values reported in Fig. 5) for both groups of online participants.

We subsequently calculated the amount of unmasking as the improvement in SRT in the conditions with masker separation from the monaural condition. Figure 6 illustrates the SRT improvements for the three listener groups under three separation types: Contralateral separation with different-sex masker, dichotic separation with different-sex masker, and dichotic separation with same-sex masker. Data were initially analyzed using a mixed-effects ANOVA with fixed-effects of group and masker (different-sex, contralateral vs dichotic). Diagnostics revealed significantly non-normal residuals but homogeneity of variance. Results of a non-parametric between-groups ANOVA revealed no significant main effect of listening group [$\chi^2(2) = 0.12$, $p = 0.941$]. Follow-up pairwise comparisons with Wilcoxon Mann Whitney tests with Bonferroni corrections indicated that there were no significant differences in the distributions of SRT improvements obtained across listener groups under any separation type ($p > 0.05$ for all three groups). The group-level difference in SRT improvements between in-lab and online testing, when averaging verified NH and self-reported NH groups was 0.6 dB for contralateral separation with different-sex masker and 0.9 dB for dichotic separation with different-sex masker. The new condition of Dichotic Separation Same-sex masker tested among online participants revealed no significant difference of same- versus different-sex masker in SRT improvement, $p > 0.05$. The unmasking effect was larger from Contralateral Separation than Dichotic Separation with both same-sex ($p = 0.0094$) and different-sex maskers ($p \leqslant 0.001$).

C. Discussion

Experiment II examined the influence of remote testing using a binaural task that assessed contralateral unmasking. Two groups of listeners were tested online, one with verified NH and the other with self-reported NH. When compared with the in-lab tested group, both groups of online participants showed elevated average SRTs by ∼7 dB and larger individual variabilities in their speech-in-speech performance. The elevated SRT and larger individual variabilities may be partially explained by the limited target audibility from individual CLs set by online listeners. Across the two groups of online participants, we observed SRTs in Quiet mostly between −25 and −70 dB re masker level at –3 dB CL. Assuming that NH adults had SRT in Quiet between 10 and 15 dB SPL [similar to in-lab participants in Goupell *et al.* (2016)], the CLs set by online listeners are estimated to be between 40 and 80 dB SPL. Although an exact mapping of CL onto SPL is unknown, this range suggests that most online listeners set their CLs below 70 dB SPL.[3] For participants who identified CLs at more conservative (lower) SPLs, the target speech would have reduced to an inaudible level during the adaptive SRT measurement more quickly than those who set CLs at higher SPLs. Similarly, individual CLs at lower SPLs would have also restricted access to lower SNRs during SRT measurements with a masker.[4] The large range of SPLs from CLs set by individual listeners contributed to the range of SRTs in Quiet observed among online participants, which further influenced SRTs under conditions with a masker (Fig. 5).

We observed similar group-level SRT improvements between in-lab and online participants, where the difference was <1 dB. This is a difference substantively smaller than the measurement resolution where the smallest step size was 2 dB during the adaptive procedure. Thus, even with elevated SRTs, the access to auditory cues for contralateral and dichotic unmasking was similar between in-lab and online testing. Additionally, the unmasking effects observed among in-lab participants were replicated by both groups of online listeners, regardless of the method of NH verification.

V. GENERAL DISCUSSION

Remote testing offers an abundance of opportunities for psychoacoustic research beyond the period of restricted lab access during COVID-19. It has the potential to provide better access to under-represented groups and clinical

21 January 2024 17:37:28

https://doi.org/10.1121/10.0020567

populations to participate in research. To investigate data quality from remote testing, particularly web-based testing with more variable audio quality from commercial hardware and home listening environments, the present study aimed to verify phenomena reported in published studies with data collected in-lab (Goupell *et al.*, 2013; Goupell *et al.*, 2016). Through two experiments, we replicated spatial hearing effects measured in-lab using two groups of listeners who were tested online: One group with verified NH and the other with self-reported NH. We observed evidence that, with the proper hardware screenings and carefully designed instructions and experimental procedure, automated online testing is a viable option for measuring spatial hearing abilities, specifically lateralization of binaural cues and speech unmasking among NH adults, even with a sample size similar to that for in-lab studies.

The perceptual headphone screen at the start of online testing provided several checkpoints to safeguard data quality. Approximately 6%–9% of online participants failed the perceptual screen due to poor audio quality from their chosen headphones. This may have been due to excessive imbalance (>3 dB) between channel outputs or improper headphone placement. For the self-reported NH group, it is also possible that they had asymmetric hearing sensitivity between ears. The perceptual screen also resulted in small adjustments of 3 dB in the stereo outputs for ∼20% of the participants. This procedure allowed us to include a substantial portion of participants whose commercial audio hardware might not have research grade quality. Small adjustments to stimulus output within a limited range can be a useful consideration for future work to reach many underrepresented participant groups who do not have access to expensive commercial audio hardware.

Once the online participants began the automated procedure, a small portion aborted the experiment early. The attrition rate was 4% for experiment I and ∼10% for experiment II. The difference in attrition rate between experiments may be due to the fact that experiment I only involved a single condition (e.g., lateralization to either ITDs or ILDs), whereas experiment II involved six conditions with adaptively changing task difficulty, which required a longer testing period and potentially resulting in increased fatigue. Unlike in-lab testing where participants are often monitored for fatigue, the design of automated online experiment should consider both the task complexity and duration to avoid elevated attrition.

In addition, we discovered several opportunities for additional control in designing remote research testing protocols. One valuable lesson learned from experiment I was the need to check for correct headphone channel placement on ears (i.e., left vs right) during screening. We observed two participants from the verified NH group and four from the self-reported NH group to have reversed lateralization curves. Although online participants could complete both tasks in this study with flipped headphones, other tasks that rely on correct identification of hemifield may suffer from incorrect channel placement on ears. The verified NH group generally had more experience with auditory tasks and therefore were more likely to verify headphone placement and choose a quiet home environment for online testing. In experiment II, the online group with self-reported NH showed an extended range of SRTs in Quiet below −65 dB (re masker level at −3 dB CL), suggesting that they may have been more likely to identify high dB SPLs as comfortable during the subjective calibration. Since this group likely represents naive listeners who may be recruited from online platforms (Milne *et al.*, 2021) for future remote research studies, the automated experimental procedure may benefit from additional steps to check for high stimulus presentation levels (e.g., from SRT in Quiet). This step may provide additional opportunities to safeguard stimulus quality (e.g., to avoid clipping), protect participants against exposure to loud sounds, and identify individuals with poor hearing sensitivity from their needs for louder than typical presentation levels.

The home environments are arguably more distracting than sound booths in laboratories for participants during psychoacoustics tasks. In experiment II we observed elevated speech-in-speech thresholds by ∼7 dB from online testing as compared to in-lab testing on the group level. While the majority of the threshold elevation is likely due to differences in audibility as a result of individually set presentation levels, we do not yet have a way to fully ascertain the potential role of attention shifts during testing in the home environment. However, overall, our results suggest that online testing in the home environment yields similar results to in-lab testing for binaural cue lateralization and speech unmasking experiments. Future work is warranted to fully understand the impact of attention shifting and different types of hardware (e.g., wireless headphones) on spatial hearing tasks and to further refine hardware screenings and subjective calibration procedures for online auditory experiments.

VI. CONCLUSION

The present study validated online testing as a viable option to measure spatial hearing abilities through two tasks that replicated recently published in-lab studies: lateralization to binaural cues (experiment I) and speech unmasking (experiment II). We recruited two groups of NH adults, one with verified NH and the other with self-reported NH. SRT thresholds were elevated by ∼7 dB in the online tested groups, for which the perceptual calibration might contribute to such elevation. Group-level comparisons suggested that access to auditory cues for lateralization to ITD/ILD and contralateral and dichotic speech unmasking are similar between online and in-lab testing. Our results provide encouraging evidence that, with proper perceptual-based headphone screening and stimulus level calibration, online testing is a viable option to test spatial hearing tasks that involve delivering stimuli with interaural level and timing differences and aim to elicit perceptual separation of target and masker speech.

21 January 2024 17:37:28

ACKNOWLEGEMENTS

This work was supported by the National Institute of Health (NIH) National Institute on Deafness and Other Communication Disorders Grants Nos. R01DC003083 and R01DC008365 (R.Y.L.) and in part by a core grant to the Waisman Center at UW-Madison from the NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant No. P50HD105353.

[1]For significance test, p-values were corrected by multiplying 3 for comparison against $\alpha = 0.05$ or were uncorrected for comparison against adjusted $\alpha = 0.0167$. Corrected p-values are reported in text, while p-values remain uncorrected in figures.

[2]SRT in Quiet in SPL was unattainable in this task due to the adaptive track starting at a subjective level set by each participant at an unknown SPL (CL-3, also referred to as "masker-level"). Hence, SRT in Quiet was reported as a dB level referencing the masker level as 0 dB even though no masker was present in this condition.

[3]This is also the intensity range of 62–68 dB by normal-speaking to raised voice effort (Pavlovic, 2018) that online listeners might find perceptually comfortable.

[4]For instance, if an NH adult had an SRT in Quiet at 10 dB SPL and an SRT with a monaural, different-sex masker at 35 dB SPL (estimated from −60 and −35 dB re masker level at 70 dB, respectively, from in-lab participants' regression fit in Fig. 5) but identified a CL at 55 dB SPL in online testing with a masker level subsequently set at 52 dB SPL, this individual would likely achieve −42 dB SRT in Quiet and −17 dB SRT with masker (re masker at 52 dB SPL).

Anderson, S. R., Easter, K., and Goupell, M. J. (**2019**). "Effects of rate and age in processing interaural time and level differences in normal-hearing and bilateral cochlear-implant listeners," J. Acoust. Soc. Am. **146**(5), 3232–3254.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (**2020**). "Gorilla in our midst: An online behavioral experiment builder," Behav. Res. **52**(1), 388–407.

Bernstein, L. R., and Trahiotis, C. (**2002**). "Enhancing sensitivity to interaural delays at high frequencies by using 'transposed stimuli,'" J. Acoust. Soc. Am. **112**(3), 1026–1036.

Ehlers, E., Kan, A., Winn, M. B., Stoelb, C., and Litovsky, R. Y. (**2016**). "Binaural hearing in children using Gaussian enveloped and transposed tones," J. Acoust. Soc. Am. **139**(4), 1724–1733.

Frund, I., Haenel, N. V., and Wichmann, F. A. (**2011**). "Inference for psychometric functions in the presence of nonstationary behavior," J. Vision **11**(6), 16.

Gijbels, L., Cai, R., Donnelly, P. M., and Kuhl, P. K. (**2021**). "Designing virtual, moderated studies of early childhood development," Front. Psychol. **12**, 4331.

Goupell, M. J., Best, V., and Colburn, H. S. (**2021**). "Intracranial lateralization bias observed in the presence of symmetrical hearing thresholds," JASA Express Lett. **1**, 104401.

Goupell, M. J., Kan, A., and Litovsky, R. Y. (**2016**). "Spatial attention in bilateral cochlear-implant users," J. Acoust. Soc. Am. **140**(3), 1652–1662.

Goupell, M. J., Stoelb, C., Kan, A., and Litovsky, R. Y. (**2013**). "Effect of mismatched place-of-stimulation on the salience of binaural cues in conditions that simulate bilateral cochlear-implant listening," J. Acoust. Soc. Am. **133**(4), 2272–2287.

Hartmann, W. M. (**2021**). "Localization and lateralization of sound," Auditory Res. **73**, 9–45.

Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., and O'Donnell, T. J. (**2019**). "A thousand studies for the price of one: Accelerating psychological science with Pushkin," Behav. Res. **51**(4), 1782–1803.

Kidd, G., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**(6), 3793–3802.

Larrouy-Maestri, P., Harrison, P. M. C., and Müllensiefen, D. (**2019**). "The mistuning perception test: A new measurement instrument," Behav. Res. **51**(2), 663–675.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2B), 467–477.

Lelo de Larrea-Mancera, E. S., Stavropoulos, T., Carrillo, A. A., Cheung, S., He, Y. J., Eddins, D. A., Molis, M. R., Gallun, F. J., and Seitz, A. R. (**2022**). "Remote auditory assessment using Portable Automated Rapid Testing (PART) and participant-owned devices." J. Acoust. Soc. Am. **152**(2), 807–819.

Mehr, S. A., Kotler, J., Howard, R. M., Haig, D., and Krasnow, M. M. (**2017**). "Genomic imprinting is implicated in the psychology of music," Psychol. Sci. **28**(10), 1455–1467.

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. v., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., and Glowacki, L. (**2019**). "Universality and diversity in human song," Science **366**(6468), eaax0868.

Mehr, S. A., Singh, M., York, H., Glowacki, L., and Krasnow, M. M. (**2018**). "Form and function in human song," Curr. Biol. **28**(3), 356–368.

Merchant, G. R., Dorey, C., Porter, H. L., Buss, E., and Leibold, L. J. (**2021**). "Feasibility of remote assessment of the binaural intelligibility level difference in school-age children," JASA Express Lett. **1**(1), 014405.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., and Chait, M. (**2021**). "An online headphone screening test based on dichotic pitch," Behav. Res. **53**(4), 1551–1562.

Mozilla (**2023**). "Web Audio API - Web APIs | MDN," https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API (Last viewed September 16, 2021).

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (**2014**). "The musicality of non-musicians: an index for assessing musical sophistication in the general population," PLoS One **9**(2), e89642.

Padilla-Ortiz, A. L., and Orduña-Bustamante, F. (**2021**). "Binaural speech intelligibility tests conducted remotely over the internet compared with tests under controlled laboratory conditions," Appl. Acoust. **172**, 107574.

Pavlovic, C. (**2018**). "SII—Speech intelligibility index standard: ANSI S3.5 1997," J. Acoust. Soc. Am. **143**(3), 1906.

Peng, Z. E., Kan, A., and Litovsky, R. Y. (**2020**). "Development of binaural sensitivity: Eye gaze as a measure of real-time processing," Front. Syst. Neurosci. **14**(July), 1–13.

Peng, Z. E., Waz, S., Buss, E., Shen, Y., Richards, V., Bharadwaj, H., Stecker, G. C., Beim, J. A., Bosen, A. K., Braza, M. D., Diedesch, A. C., Dorey, C. M., Dykstra, A. R., Gallun, F. J., Goldsworthy, R. L., Gray, L., Hoover, E. C., Ihlefeld, A., Koelewijn, T., Kopun, J. G., Mesik, J., Shub, D. E., and Venezia, J. H. (**2022**). "FORUM: Remote testing for psychological and physiological acoustics," J. Acoust. Soc. Am. **151**(5), 3116–3128.

Peretz, I., Gosselin, N., TillmannI, B., Gagnon, B., Trimmer, C. G., Paquette, S., and Bouchard, B. (**2008**). "On-line identification of congenital amusia," Music Percept. ***25***(4), 331–343.

Peretz, I., and Vuvan, D. T. (**2017**). "Prevalence of congenital amusia," Eur. J. Hum. Genet. **25**(5), 625–630.

RTINGS.Com (**2023**). "Headphones: Reviews," https://www.rtings.com/headphones/reviews (Last viewed February 15, 2022).

Swanepoel, D. W., de Sousa, K. C., Smits, C., and Moore, D. R. (**2019**). "Mobile applications to detect hearing impairment: Opportunities and challenges," Bull. World Health Organ. **97**(10), 717–718.

Vuvan, D. T., Paquette, S., Mignault Goulet, G., Royal, I., Felezeu, M., and Peretz, I. (**2018**). "The Montreal protocol for identification of amusia," Behav. Res. **50**(2), 662–672.

Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (**2017**). "Headphone screening to facilitate web-based auditory experiments," Atten. Percept. Psychophys. **79**(7), 2064–2072.

21 January 2024 17:37:28