Published in final edited form as:
Behav Res Methods. 2024 March ; 56(3): 1433–1448. doi:10.3758/s13428-023-02101-9.

Web-based Psychoacoustics: Hearing Screening, Infrastructure, and Validation

Brittany A. Mok[a], **Vibha Viswanathan**[b], **Agudemu Borjigin**[b], **Ravinderjit Singh**[b], **Homeira Kafi**[b], **Hari M. Bharadwaj**[a,c,*]

[a]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN, United States

[b]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, United States

[c]Department of Communication Science and Disorders, University of Pittsburgh, PA, United States

Abstract

Anonymous web-based experiments are increasingly used in many domains of behavioral research. However, online studies of auditory perception, especially of psychoacoustic phenomena pertaining to low-level sensory processing, are challenging because of limited available control of the acoustics, and the inability to perform audiometry to confirm normal-hearing status of participants. Here, we outline our approach to mitigate these challenges and validate our procedures by comparing web-based measurements to lab-based data on a range of classic psychoacoustic tasks. Individual tasks were created using jsPsych, an open-source javascript front-end library. Dynamic sequences of psychoacoustic tasks were implemented using Django, an open-source library for web applications, and combined with consent pages, questionnaires, and debriefing pages. Subjects were recruited via Prolific, a subject recruitment platform for web-based studies. Guided by a meta-analysis of lab-based data, we developed and validated a screening procedure to select participants for (putative) normal-hearing status based on their responses in a suprathreshold task and a survey. Headphone use was standardized by supplementing procedures from prior literature with a binaural hearing task. Individuals meeting all criteria were re-invited to complete a range of classic psychoacoustic tasks. For the re-invited participants, absolute thresholds were in excellent agreement with lab-based data for fundamental frequency discrimination, gap detection, and sensitivity to interaural time delay

[*]Correspondence: hari.bharadwaj@pitt.edu.

Author Contributions
HMB conceived the project, designed the protocols, and implemented the infrastructure. BAM performed the meta-analysis to guide the design of our hearing screening procedure. VV, AB, RS, and HK provided design input and contributed validation data. All authors contributed to manuscript writing.

Competing Interests
The authors declare no competing financial or non-financial interests.

Ethics approval
All human subject procedures were conducted in accordance with protocols reviewed and approved by the Purdue University IRB.

Consent to participate
All subjects provided affirmative informed consent.

Consent for publication
Human subjects provided consent to publish their de-identified data.

and level difference. Furthermore, word identification scores, consonant confusion patterns, and co-modulation masking release effect also matched lab-based studies. Our results suggest that web-based psychoacoustics is a viable complement to lab-based research. Source code for our infrastructure is provided.

Keywords

Web-based testing; Psychoacoustics; Remote hearing screening; Crowd-sourcing; Online auditory experiments

Introduction

Behavioral experiments that assess the ability of human listeners to detect and discriminate various acoustic cues have provided considerable insight into the psychology of human auditory perception (Moore, 2012), placed useful constraints on theories about the neural coding of sound (Oxenham, 2018), revealed our ability to use acoustic regularities for perceptual organization of sound (Bregman, 1990), and yielded models of speech intelligibility (Bronkhorst, 2000; Elhilali et al., 2003; Steinmetzger et al., 2019; Viswanathan et al., 2022). Conventionally, such psychoacoustic experiments have been performed on a small number of listeners with known audiological characteristics, and in controlled laboratory environments. In this realm, individual variability is a nuisance parameter. In recent years, studies with a larger number of participants have exploited individual differences to understand foundational aspects of auditory processing (McDermott et al., 2010; Bharadwaj et al., 2015; Whiteford et al., 2020). However, large-N studies are difficult to perform in a laboratory environment. Web-based platforms provide an attractive alternative to quickly and cost-effectively perform studies with a large number of participants (Woods et al., 2015; Gosling and Mason, 2015). Furthermore, online platforms offer tremendous advantages in allowing researchers to access diverse subject pools, conduct cross-cultural research, and serve populations that may otherwise be unable to participate by visiting laboratory facilities (Henrich et al., 2010; Woods et al., 2013).

In the auditory domain, online platforms are being increasingly used to complement lab-based work (Lavan et al., 2019; McPherson and McDermott, 2018; Zhao et al., 2019). The COVID-19 pandemic has also hastened the auditory research community's trend towards adopting online platforms. Indeed, the Psychological and Physiological Acoustics Technical Committee of the Acoustical Society of America established a Remote Testing Task Force to collect and curate information aimed at enhancing the practicality and quality of research conducted via remote testing approaches (Peng et al., 2022). Yet, to date, the literature on web-based testing of auditory perception is limited and consists primarily of examples probing higher-level cognitive aspects of auditory perception, such as voice identity learning (Lavan et al., 2019), lexical influences on speech perception (Giovannone and Theodore, 2021), and schema learning/attentive tracking (Woods and McDermott, 2018). In contrast, online studies of low-level sensory processing (e.g., measurements of absolute sensitivity to subtle spectrotemporal cues) are scarce. This may, in part, be because although studies in other domains have demonstrated that online measurements can yield high-quality data

(Clifford et al., 2014; Woods et al., 2015), there is cause for skepticism about the quality of web-based psychoacoustics data for low-level aspects of auditory processing as these measures may be more sensitive to the testing parameters and participant characteristics. Beyond the typical concerns regarding participant engagement and compliance that apply in all domains, web-based auditory experiments also have to contend with sacrificing control of sound calibration and the listening environment (e.g., there may be ambient noise). With web-based testing, participants use their own computers and perform tasks in the environments available to them. Furthermore, given that hearing loss is highly prevalent – for example one in eight United States individuals aged 12 or older have hearing loss in both ears (Lin et al., 2011) – it is likely that subject pools recruited online will consist of a significant number of individuals with hearing loss. Because hearing loss influences many aspects of auditory perception (Moore, 2007), procedures that can classify the hearing status of anonymous participants will considerably enhance the interpretability of web-based psychoacoustic measurements.

To address these gaps, we developed approaches to screen anonymous participants for engagement, headphone use, and putative normal-hearing status in the present study. While still far from the highly controlled settings in the lab, the use of headphones (or insert earphones) by participants improves the standardization of sound presentation. Headphone use can attenuate ambient noise, circumnavigate acoustic filtering from the physical configuration of speakers and room reverberation, and allow for ear-specific stimulus presentation. Woods et al. (2017) developed a loudness-perception-based procedure to check for headphone use by exploiting the fact that antiphasic stimuli presented from two free-field speakers partially cancel each other, reducing the overall intensity. We supplemented this procedure with a binaural hearing task that is difficult to perform with one or two free-field speakers, and difficult with monaural use of earphones, but easy with stereo headphones. In addition to headphone-check procedures, we develop procedures to classify the hearing status of the online participants. Although conventional audiometry is infeasible without calibrated hardware, there is considerable literature on performing hearing screening over the telephone using suprathreshold stimuli (Smits et al., 2004; Watson et al., 2012). Inspired by this literature, we implemented a procedure to classify participants' hearing status (i.e., normal hearing or not) that was based on a combination of two measures obtained from each participant: (1) scores in a web-based word-recognition-in-babble task, and (2) responses in a self-report survey on their hearing. The choice of cutoff scores for the word recognition task were guided by a meta-analysis of the literature and validation experiments. The efficacy of this hearing-status classification procedure was confirmed by testing a cohort of individuals with an audiological diagnosis of hearing loss. Taken together, our screening procedures help to filter our subject pool for good engagement, putative normal-hearing status, and compliance with instructions to use stereo headphones.

The multistep screening approach for headphone use and normal-hearing status introduces logistical challenges given that the trajectories followed by different participants through our study can different, and the sequences of web pages seen by each participant will need to be dynamically chosen. Moreover, the web application must anonymously track the participant through various components of the study, provide a simplified experience and study flow, and calculate compensation that is tailored to each participant's trajectory. The application must also integrate seamlessly with the subject recruitment workflow provided by Prolific (https://www.prolific.co). To meet these requirements, we developed a web application using the Django framework (https://www.djangoproject.com/, Django Software Foundation) with custom features. In the following sections, we describe our custom infrastructure. Source code for the infrastructure is also provided as a public GitHub repository (see section on Code Availability).

We validate our overall sequence of procedures by comparing the performance of screened participants on a range of classic psychoacoustic tasks with lab-based results obtained from subjects with known normal-hearing status (as measured with standard audiometry) and prior literature. Given that the main challenge with web-based psychoacoustics is limited control of acoustics, our validation experiments focused on measuring sensitivity (i.e., absolute thresholds) to subtle spectrotemporal cues. Specifically, we measured thresholds for fundamental-frequency discrimination and gap detection with diotic stimuli, and binaural sensitivity to interaural level and time differences. We then measured the effect of modulation statistics on auditory scene analysis by quantifying the co-modulation masking release effect for tone-detection in a background of modulated noise. Finally, we measured participant performance for speech identification in noise, quantifying both overall intelligibility and consonant confusion patterns. Across all measures, our results show that web-based psychoacoustics can yield comparable results to lab-based studies.

Materials and Methods

Participants for Lab-Based Testing

All lab-based human subject procedures were conducted prior to the COVID-19 pandemic, and in accordance with protocols approved by the Institutional Review Board (IRB) and Human Research Protection Program (HRPP) at Purdue University (Protocol # 1609018209). All participants had thresholds of 25 dB HL or better in both ears at audiometric frequencies in the 0.25—8 kHz range. Participants provided informed consent, and were compensated for their time. Data were acquired in the lab for a co-modulation masking release (CMR) task and compared with the CMR data obtained with identical stimuli from anonymous participants using web-based procedures.

Participant Recruitment and Pre-Screening for Web-Based Testing

All web-based human subject procedures were conducted in accordance with protocols approved by the Institutional Review Board (IRB) and Human Research Protection Program (HRPP) at Purdue University (Protocol # IRB-2020–876). Participants were recruited anonymously via Prolific (https://www.prolific.co). Prolific is an online platform for human subject recruitment where researchers can post web-based tasks (Peer et al., 2017). Anyone

Code Availability Statement
Our infrastructure was developed using open-source toolboxes (jsPsych and Django). The code for our Django application is publicly available on GitHub (https://github.com/haribharadwaj/SNAPlabonline) and archived using Zenodo (Bharadwaj, 2021). A working demo of the app is hosted at https://snaplabonline.com. The license for the code is highly permissive and interested investigators are welcome to adapt as needed for their purposes. The audio files used to create the MRT stimuli were obtained from the National Institute of Standards and Technology PSCR Audio Source Files repository (https://www.nist.gov/ctl/pscr/pscr-audio-source-files). The speech materials used for the consonant confusion task were from the commercially available STeVI corpus (Sensimetrics Corporation, Gloucester, MA).

can sign up to be a participant and complete the tasks in exchange for payment. Individuals who sign up for participant accounts are requested to answer a series of “About You” questions. Prolific allows researchers to pre-screen participants based on their responses to different items in the ”About You” section, such that only participants meeting the researcher’s criteria are shown the study. Prolific also implements an approval process where researchers can “reject” participant submissions if there is clear evidence of poor engagement or poor compliance with instructions. Prolific tracks participants’ proportion of approved and rejected submissions. Prolific’s documentation provides examples of fair and unfair “attention checks” to guide the design of procedures to probe engagement. For our experiments, the pre-screening criteria required participants to:

1. be US/Canada residents and native speakers with English as first language (because we use speech stimuli spoken with North-American accents),
2. be in the 18–55 year age range,
3. have prior participation in at least 40 other studies on Prolific, and
4. have more than 90% of their previous Prolific submissions approved

For studies intended to recruit participants with normal hearing, another Prolific pre-screening criterion was that they should have answered “No” to the question “Do you have any hearing loss or hearing difficulties?”

Each participant’s Prolific ID served as the identifier for all data associated with that participant. This allowed for tracking the same individual across multiple visits to our web application. Participant compensation was a fixed amount for each task in the range of $8 to $11 per hour based on the median time taken (across participants) to complete the task.

Infrastructure

Individual psychoacoustic tasks were created using jsPsych (https://www.jspsych.org/), a free and open-source JavaScript library for running behavioral experiments in a web browser (De Leeuw, 2015). This library is well-suited for trial-based task design and includes an array of “plugins”, each implementing a different interface for stimulus delivery and participant interaction. We adapted the audio plugins available in jsPsych v6.1.0 for our purposes.

Using jsPsych, all audio-based tasks were set up to begin with a set of instruction screens that introduced the participant to the task at hand. Following the instructions, participants landed on a volume-setting screen. This volume-setting screen instructed them to set the volume level on their computer to a low value (10–20% of the maximum on their computers). On the next screen, they were instructed to hit “play” on a calibration sound, and adjust the volume level up to a clearly audible and comfortable level while the calibration sound is playing. The calibration sound for each task was chosen to have similar spectrotemporal characteristics as the stimuli in the task (often just long strings of stimuli from the task). For all audio trials, the stimulus level was restricted to be no higher than 6 dB above the calibration sound. All audio tasks reported in this study involved stimulus presentation followed by a request for a button-click response (classic n-alternatives forced-choice trials). Feedback was provided in all audio tasks except the headphone screening using a green thumbs up for correct responses, and a red thumbs down for incorrect responses. Each audio trial was implemented using a modified version of the “audio-button-response” plugin, where the modification was done using javascript to disable (grey out) the response buttons until the audio was done playing. Because all audio tasks had a similar structure and used the same jsPsych plugins, the Django app (described below) was set up to take a JSON format (https://www.json.org/) text file with stimulus information and automatically convert it to appropriate jsPsych javascript code. This allowed for lab members who were unfamiliar with javascript to easily design their tasks.

While individual tasks were designed using jsPsych, the full-featured study design involved a combination of (1) a consent and Prolific ID confirmation page, (2) a demographics and hearing-status survey, (3) multiple individual audio tasks implemented using jsPsych with dynamic constraints on task sequence (e.g., present task #2 if subject scores more than X% in condition 1 of task #1, else conclude and debrief), (4) informational pages that showed the participants where they were within the study sequence and how they could exit the study at any time with partial compensation, and finally (5) a debriefing page at the end that displayed the compensation amount for each subject based on their particular trajectory through the study, and redirected them to submit a “completion code” back to Prolific for study completion (or alternately request them to return the study with partial completion under certain conditions described in the Headphone-use Screening section). On the experimenter side, it was necessary to set up individual web pages to upload task information, and a page with an interface that allowed for stringing different tasks into a full study with conditional flow constraints. Pages were also set up for downloading response data, and tracking study progress and subject compensation amounts.

This complex study flow was made possible using Django (https://www.djangoproject.com/), which is a free and open-source Python-based framework for developing web applications and controlling server behavior (the so-called “back-end” logic that complements our“front-end” jsPsych JavaScript). The Django application was set up to serve the consent form, the survey pages, and the individual jsPsych task pages as the subject proceeded through the study. Pages where lab members can upload the JSON files with task information and create studies were also set up within the Django application. Pages other than jsPsych tasks were created using simple HTML and styled using Bootstrap (https://getbootstrap.com/). For the jsPsych tasks, the Django app was also set up to automatically and asynchronously (i.e., without refreshing the page) extract single-trial data as it was generated within jsPsych and write it to an SQLite (https://www.sqlite.org/) database on our server. The app also performed calculations (e.g., of scores for screening tasks) to decide task flow and compensation. Each study was assigned a random alphanumeric slug as URL, and this link was posted on Prolific. Prolific allows the use of URL parameters to automatically extract the Prolific ID of each participant, which the Django app was set up to take advantage of. As a result, participants were just asked to confirm their Prolific ID rather than enter it manually.

Note that while Django is a general purpose web app framework, it offers many out-of-the-box features that made it a convenient choice to complement the capabilities of jsPsych. In particular, Django comes with functionality for working with databases using python objects (knowledge of SQL is not necessary), secure authentication capabilities that allows for logins for lab members, fine-grained permission control on creating tasks/studies and viewing results, and the ability to use sessions and cookies to track participants anonymously through various parts of the study. Importantly, the API provided by Django cleanly separates the rendering of front-end HTML/javascript from the back-end calculations and data handling, making it possible to seamlessly mix jsPsych tasks with other kinds of pages and server control. The Django app was hosted on a virtual private server (VPS) rented from Linode (https://www.linode.com/) and served by Apache2 (https://httpd.apache.org/) on Ubuntu Linux (https://ubuntu.com/). Communications between our server and subject browsers were encrypted using free SSL capabilities provided by Let's Encrypt (https://letsencrypt.org/). A working example of our Django app, which includes a ”demo” study, can be viewed at https://www.snaplabonline.com. The source code for the application is available on GitHub (see section on Code Availability). The key resources used to build the infrastructure for web-based testing and recruit participants are illustrated in Figure 1A.

Headphone-Use Screening

Participants were instructed to use stereo headphones in a quiet room. While compliance can generally be expected, it is possible that some participants do not comply and instead use either single-channel headphones/earphones or free-field speakers. Therefore, to objectively check for stereo headphone use, we used a combination of two tasks (six trials each). In the first task, participants were instructed to identify the softest of a sequence of three low-frequency tones. The target tone was 6 dB softer than the two foil tones, but one of the foil tones was presented with opposite phases to the left and right channels (Woods et al., 2017). Woods et al. (2017) reasoned that if a participant used a pair of free field speakers, acoustic cancellation would result in a greater attenuation of this “antiphase” foil tone compared to the target, which would cause the subject to report the wrong tone as the softest. While this is true to some extent for typical two-channel free field speakers, in general, the effectiveness of the acoustic cancellation would depend on the speaker configuration and the physical set up (e.g., distance, orientation, number of channels, etc.). Crucially, if the participant used a single-channel free-field speaker or just one-channel headphone/earphone, the foil would be ineffective. To catch participants who use a single-channel set up, we added a second task where participants had to report whether a low-frequency chirp (150–400 Hz) embedded in background low-frequency noise was rising, falling, or flat in pitch. The stimulus was designed such that chirp was at pi-phase between the left and right channels, whereas the noise was at zero phase (i.e., the so-called “$N0S\pi$” configuration). Importantly, the signal-to-noise ratio (SNR) was chosen such that the chirp would be difficult to detect with just one channel, but easily detected with binaural headphones owing to the so-called binaural masking level difference (BMLD) effect that would yield a substantial masking release (Licklider, 1948). The use of a low-frequency chirp was advantageous in two ways: (1) If a subject used two-channel free-field speakers, not only would the BMLD benefit be absent, but the same acoustic cancellation effect as described in Woods et al. (2017) would result in a further reduction in the SNR. This is in contrast to another recently described procedure based on Huggins' pitch (Milne et al.,

2020) where free-field acoustic cancellation could enhance the perception of the target. (2) The test would be applicable even to individuals with audiometric hearing loss. This is because most individuals with hearing loss have a sloping audiogram with high-frequency loss (Parthasarathy et al., 2020), and the BMLD effect is known to be largely preserved when the low-frequency loss is not substantial (Jerger et al., 1984). Participants had to correctly respond to five out of six trials in each task to pass the headphone screening and proceed to the other tasks in the study. One exception to this was the study used to validate our hearing-screening procedures where we explicitly sought to recruit individuals with a diagnosis of audiometric hearing loss; the cut-off was relaxed to four out of six trials for that cohort. Milne et al. (2020) showed that combining complementary tests can boost the overall selectivity of headphone screening.

Note that headphone screening procedures place a logistical challenge when subjects are recruited through Prolific. Our interpretation of Prolific's current policies (which we fully support) was that we can only exclude participants mid-way through a study without pay if there is clear evidence of non-compliance with instructions or clear evidence of inattentive engagement. However, the headphone checks of the nature used in this study cannot be expected to have 100% sensitivity and specificity. Indeed, it is possible that a small number of participants fail the headphone screening even if they comply with the instructions to use stereo headphones. This could happen if for instance, unbeknownst to the subject, their headphone (or computer settings) did not separate left and right channels adequately, or if they had asymmetric hearing loss, or if they did not understand the task well enough right from the first trial to meet threshold for passing. Thus, when a participant failed headphone screening, the Django app concluded the study and showed a debriefing page explaining that we were unable to verify headphone use ("quality check failed") and instructed the participant to "return" the study to Prolific. Such participants were manually compensated for the time spent on the headphone screening (compensation amount tracked by the Django app) using the "bonus payment" feature on Prolific.

Hearing Screening

It is well known that individuals with audiometric hearing loss show reduced performance in a range of suprathreshold tasks, the best known example of which is perhaps speech understanding in noise. Difficulty understanding speech in noise is the most common audiological complaint among individuals who are eventually diagnosed with hearing loss (Blackwell et al., 2014). Thus, while conventional threshold audiometry is infeasible without calibrated hardware, such suprathreshold tasks may plausibly be used to screen for hearing loss. Indeed, identification of spoken digits in background noise has been validated for hearing screening via telephone in many countries by comparing the threshold SNR for digit identification to audiometric thresholds (Smits et al., 2004; Watson et al., 2012). Along the same lines, we developed a hearing screening procedure using material from the modified rhyme test (MRT) (House et al., 1963) presented in 4-talker babble. The MRT materials include 50 lists of monosyllabic words, where each list consists of six monosyllabic word alternatives that differ in just the first or the last phoneme. The MRT is thus convenient for administration with a web-based interface by virtue of allowing for a 6-alternatives forced-choice response. Despite the closed-set nature of the response, the MRT is known to capture the suprathreshold deficits resulting from hearing loss (Elkins, 1971; Bell et al., 1972; Brungart et al., 2014; Miner and Danhauer, 1976), as evidenced by the finding that individuals with hearing loss exhibiting poorer performance even at high stimulus levels (Brungart et al., 2021). To use the MRT-in-babble as a hearing screener, suitable cutoff scores must be determined below which an individual would be considered as failing the screening. Any given performance cutoff represents a tradeoff between sensitivity and specificity for detecting hearing loss, with the precise trade-off relationship dependent on the effect size of hearing loss on speech-in-noise scores. To determine a suitable performance cutoff, we conducted a meta-analysis of 15 studies that reported speech-in-noise scores for various materials across normal-hearing listeners and individuals with hearing loss (pure-tone average thresholds > 30 dB HL). Inverse variance pooling was used to estimate the effect size using the bias-corrected procedure outlined in Hedges (1982). The efficacy of the hearing screening procedure was then validated by measuring the pass rates among individuals with a history of hearing loss as diagnosed by an audiologist or otologist.

Psychoacoustic Tasks

The primary experimental strategy used to investigate the validity and viability of web-based psychoacoustics was to compare results from web-based administration of a range of classic psychoacoustic tasks with corresponding data from the same (or substantially similar) tasks in controlled lab-based settings. Specifically, we used the following measures to validate our web-based procedures:

1. Fundamental-frequency (F0) discrimination (F0 difference limens; F0DLs)
2. Gap detection (duration thresholds)
3. Interaural time difference (ITD) threshold sensitivity
4. Interaural level difference (ILD) threshold sensitivity
5. co-modulation masking release (CMR) effect
6. Effect of SNR on word identification, and
7. Consonant confusion patterns in background noise

The tasks employed target a range of phenomena pertaining to sensory processing that have been classically studied in the psychoacoustics literature. These include detection of subtle diotic and dichotic cues (F0DLs, Gap/ITD/ILD thresholds), the use of modulation cues for auditory scene analysis (CMR), and speech categorization with no syntactic or semantic context (monosyllabic word identification in noise). The specific stimuli employed for each task are described alongside the measured performance trends in the Validation Experiments and Results section.

Overall study Sequence

To integrate our screening procedures and psychoacoustic tasks with the workflow facilitated by Prolific, the data collection procedures were mapped to two separate studies on Prolific (Figure 1B). In addition to the pre-screening based on Prolific's "About You" bank of questions, Prolific also allows for custom pre-screening where researchers can upload an “allowlist” of Prolific IDs to restrict which participants are shown the study. This allowlist feature enables longitudinal designs where researchers can store the IDs of participants from one study, apply any custom criteria to the data from that first study, and re-invite participants who meet criteria for follow-up studies. We leveraged this feature to implement hearing screening within an entry study; the participants who passed the screening were re-invited to participate in studies that contained the other, main psychoacoustic tasks (Figure 1B). Headphone screening procedures were included in both the entry study and the follow-up studies.

Validation Experiments and Results

Validation of hearing screening procedure

For hearing screening, we applied cutoffs to percent-correct scores obtained for recognition of words from the modified rhyme test (MRT; House, 1963). The words were presented with the carrier phrase “Please select the word ____ ” in 4-talker babble with matching long-term average spectrum at four different SNRs (10, 5, 0, and −5 dB). To help choose appropriate cutoffs, a meta-analysis of the literature was conducted. This helped obtain an initial estimate of the effect-size with which speech-in-noise tests generally separate listeners with normal hearing (NH; pure-tone average/PTA thresholds ≤25 dB) from those with hearing loss (HL; PTA ≥ 30 dB). 15 studies were selected for inclusion based on the following criteria:

1. Use of English-language materials
2. Reporting of either percent-correct speech identification for a fixed SNR away from floor and ceiling, or the SNR at which a fixed percent-correct was obtained
3. Inclusion of NH and HL groups within the same study
4. Explicit reporting of mean and standard deviation in each group
5. Use of one or more of the following standard speech-testing materials: Bamford-Kowal-Bench Speech-in-Noise Test/BKB-SIN (Niquette et al., 2003), the Quick Speech-in-Noise Test/QuickSIN (Killion et al., 2004), Words-in-Noise test/WIN (Wilson, 2003), Hearing in Noise Test/HINT (Nilsson et al., 1994), NU-6 words in noise (Auditec Inc.), or the MRT

Although the meta-analysis was not exhaustive, and the speech-testing material was non-uniform across the studies chosen, we considered the analysis adequate to provide an initial estimate of cutoffs for our purposes. This is because the actual selectivity of our web-based hearing screening was separately quantified by applying the chosen cutoffs to data from a cohort of individuals with a diagnosis of hearing loss. The number of studies (i.e., 15) included in the analysis was also considered adequate because the effect-size was estimable with a narrow confidence interval (Figure 2A) using bias-corrected inverse variance pooling (Hedges, 1982). The meta-analysis revealed a large effect size of g = 2.05 based on the pooled-variance estimate (Figure 2A), and that the standard deviation of scores in the HL group was a factor of three larger than the standard deviation in the NH group. Using this standard-deviation factor and overall effect-size estimate, it is possible to theoretically calculate the receiver-operator characteristics (ROC curve) that can be expected when using a cutoff procedure to blindly classify the hearing status of an individual participant. This theoretical ROC curve is shown in Figure 2B. These initial estimates suggested that if we choose a cutoff such that we exclude 20–40% of NH participants, we would be able to exclude around 90% of participants with hearing loss. Because our initial pool in web-based testing would likely include some participants with hearing loss, and because it is possible that web-based procedures yield larger across-participant variance, we chose our initial cutoff conservatively to exclude 35% (i.e., closer to 40% rather than 20%) of the entry pool. Accordingly, we set initial cutoffs for word recognition scores at each SNR at the 65th percentile for that SNR. Then the cutoffs were relaxed in steps of 2 percentile points at each SNR such that 35% of the subjects would be excluded with all SNRs taken together. This resulted in a screening procedure where participants who scored 100% at 10 dB SNR, ≥ 83% at 5 dB SNR, and ≥ 75% at 0 dB SNR were deemed to have “normal” hearing. Note that setting performance cutoffs on a task also inherently filters for participants showing good compliance with instructions and focused attention. To further filter for focused engagement, we also eliminated any participants who registered more than two “blur” events during the course of this 6–7 minute-long MRT test. This was possible because jsPsych automatically records a few different kinds of user interaction events, including the so-called “blue” events. A blur event occurs when the user clicks on another window or tab during the experiment, indicating that they are no longer interacting with the experiment.

To estimate the actual selectivity of our hearing screening procedure, we applied the procedure to a separate pool of participants who met all of the following conditions: They (1) self-identified as having hearing loss or hearing difficulties on Prolific, (2) were subsequently able to confirm through our survey that they had indeed received a diagnosis of hearing loss from a medical professional (audiologist or otologist), and (3) were able report their diagnosed degree of HL with high confidence. Subjects were instructed to *not* use any hearing aids or assistive listening devices they may have, and instead listen to the stimuli via regular hearphones or earphones. However, they were given the same instructions for setting the computer volume as the other experiments; they were asked to adjust the volume to an audible and comfortable level while a calibration sound was playing. Of this pool of $N = 72$ participants, only 19% exceeded our cutoff (i.e., the overall sensitivity for detecting any degree of HL was about 81%). An examination of the degree of HL among those who passed our hearing screening revealed that the majority of hearing-impaired participants that our test failed to catch only had a mild HL. Specifically, while 31% of subjects reporting a clinical diagnosis of mild HL were able to meet cutoff, only 3% of those with more-than-mild HL (i.e., clinical diagnoses of moderate, moderately severe, or severe HL combined) were able to meet cutoff (Figure 2C). Taken together, our results confirm that we obtain > 65% specificity (by choice of cutoff, less than 35% of NH individuals will fail hearing screening), > 80% sensitivity for correctly excluding subjects with any degree of HL, and > 95% sensitivity for correctly excluding subjects with more-than-mild HL. Thus, although by no means a substitute for audiometric screening, our suprathreshold screening based on MRT scores is quite successful in filtering the subject pool for near-normal hearing.

To improve our hearing screening even further, we supplemented this MRT-based screening with the requirement that participants had to explicitly deny having hearing loss in our demographics survey. We also required that participants had to deny having any neurological disorders or persistent tinnitus. A Prolific-style “allowlist” of entry-study participants who met all of our criteria was maintained, and the “main” studies were only open to individuals on this list (Figure 1B). Next, we measured the performance of re-invited participants on a range of classic psychoacoustic tasks to test whether web-based psychoacoustics yields results comparable to lab-based data.

Absolute sensitivity measurements using diotic stimuli

The first series of measurements probed absolute psychoacoustic sensitivity to subtle fundamental frequency shifts (perceived as pitch variations) and gap duration cues for $N = 100$ participants in our allowlist. Fundamental-frequency (F0) discrimination was tested for a harmonic complex tone with an F0 of 110 Hz, using stimulus parameters resembling a recent large-scale lab-based study (Madsen et al., 2017). Specifically, F0 difference limens were measured using a three-interval 3-alternatives forced-choice (3AFC) paradigm. Each interval consisted of a sequence of four 200-ms-long complex tones ramped on and off using 20-ms-long Hann tapers. The first and third tone in the target interval had a symmetric F0 shift centered at 110 Hz. All four tones in the two foil intervals had a constant F0 of 110 Hz. F0 difference sensitivity was measured for each participant using the method of constant stimulus for various values of F0 shift from 0.05 Hz to 3.2 Hz in geometric steps. Psychometric functions were fitted using a Bayesian approach with a beta-binomial model (Schütt et al., 2016). Numerical integration procedures implemented in Psignifit 4 (Schütt et al., 2016) were used to estimate threshold, slope, and lapse-rate parameters using the default priors in the software, and with chance level set to 33%. The average and 95% (posterior) confidence interval for the F0 difference limen, defined as the 79.4% point on psychometric function, were estimated (Figure 3A). The F0-difference limen was about 0.5%, showing that the exquisite F0 sensitivity of typical normal-hearing human subjects was estimable using web-based testing. These results are a close match with lab-based estimates (Madsen et al., 2017). To quantitatively compare lab-based and web-based results, the effect size of the lab-web difference was computed using inverse-variance pooling (Hedges, 1982). To obtain the reference lab-based values, data from both musicians and non-musicians from Madsen et al. (2017) were pooled together to generate a single median and standard-deviation estimate. The computed effect size is provided in Table 1.

Gap-duration sensitivity was measured using a simple 3AFC detection task for short gaps in 4 kHz tones embedded in background noise. The stimulus parameters were chosen to resemble a recent large-scale lab-based study (Patro et al., 2021). The background noise ranged from 0.5 octaves below to 0.5 octaves above 4 kHz and was 10 dB below the 4 kHz tone in intensity (i.e., SNR = 10 dB). The tonal segments before and after the short gap were each 125-ms long and were ramped on and off using 1-ms-long Hann tapers. The noise itself started 50 ms before the first tone segment and was gated on with a 10-ms Hann taper. The gap duration in the target interval was varied, whereas the two foil intervals always had a gap duration of 0 ms. Detection accuracy was measured for target gap durations ranging from 1 to 32 ms in geometric steps. Once again, psychometric functions were fitted using the Bayesian approach. Results demonstrated exquisite sensitivity to gaps with mean 79.4%-thresholds of about 6 ms (Figure 3B), which are well in line with lab-based results from Patro et al. (2021). A quantitative comparison in provided in Table 1. Note that mean and standard error values were pooled across the young and middle-aged subjects in Patro et al. (2021) to reference lab-based values for comparison.

Taken together, our results suggest that despite using participants' own hardware in their own listening environments, absolute psychoacoustic sensitivity measures matched highly controlled lab-based data for diotic stimuli. An effect size of 0.2 is considered small by conventional criteria (Ferguson, 2009). For the F0DL and gap-detection tasks, Hedges g values were less than 0.2 (mean of −0.02 and 0.01, respectively).

Absolute sensitivity to interaural differences

Next we asked whether web-based measurements can replicate lab-based data when ear-specific stimuli need to be delivered. To this end, we probed absolute sensitivity to interaural cues. Humans are known to be sensitive to interaural time differences in low-frequency sounds that are an order of magnitude or more smaller than the width of typical neural spikes (about 1 ms), and sensitive to interaural level differences that occur in the environment for a source that is displaced from the azimuthal center by as few as a couple of degrees for high-frequency sounds (Moore, 2012). We sought to test whether such sensitivity can be robustly demonstrated with web-based testing. Interaural time difference (ITD) sensitivity for 500 Hz tones was assessed using a two-interval task where the leading ear was switched between the intervals. Participants were asked to report whether the sound jumped from left to right, or right to left in a 2-alternatives forced-choice (2AFC) design. Accuracy was measured for ITD values ranging from 2 to 128 μs in geometric steps. Psychometric curves were fitted using the Bayesian approach as before for data collected from $N = 200$ subjects who had passed our screening procedures. ITD thresholds (defined as 75% correct points) in the vicinity of 28 μs were obtained (Figure 4A), consistent with classic literature (Zwislocki and Feldman, 1956; Klumpp and Eady, 1956), and a recent lab-based study using identical stimuli and task (Borjigin et al., 2022).

Along the same lines, interaural level difference (ILD) sensitivity for 4 kHz tones was assessed using a two-interval task, where the ear with the more intense tone was switched between the intervals. As with the ITD task, participants were asked to report whether the sound jumped from left to right, or right to left (2AFC). Data from $N = 200$ subjects revealed that ILD thresholds were around 0.8 dB, consistent with classic literature (Mills, 1960). The psychometric data for ITD and ILD detection are shown in Figure 4 and confirm the validity of web-based testing for probing sensitivity to binaural cues. Quantitative comparisons of lab-based and web-based thresholds for ITD and ILD sensitivity are provided in Table 1. Similar to thresholds measured in diotic tasks, the effect size of the the lab-web difference was also less than small (i.e., less than 0.2) for the binaural measures (mean of 0.06 and −0.14, respectively).

Measuring scene analysis using co-modulation masking release

Humans use statistical regularities in sounds to perceptually organize sound information that arrives simultaneously from multiple sound sources in the environment (Bregman, 1990). One acoustic cue that promotes the grouping of different components of a sound source into a single perceptual object is co-modulation (Darwin, 1997). Indeed, it is thought that masker sounds that contain acoustic components that are temporally coherent with a target source will perceptually interfere with target perception (Shamma et al., 2011; Viswanathan et al., 2021a, 2022). The phenomenon of co-modulation masking release (CMR) using simple tone-in-noise stimuli is thought to capture aspects of such co-modulation-based scene analysis. For tones masked by on-frequency noise, the addition of flanking bands of noise, or a widening of the bandwidth of the on-frequency masking noise significantly improves detection thresholds for the target tones, but only if the different masker frequency components are co-modulated (Schooneveldt and Moore, 1989). Qualitatively, co-modulated masker components are perceived as a single broadband object thereby allowing for the target (narrowband) tone to stand out. Here, we tested whether the CMR effect can be illustrated using web-based psychoacoustic measurements.

Detection thresholds were measured for 4 kHz tones embedded in narrowband 1-ERB-wide (Glasberg and Moore, 1990) 10-Hz modulated noise, with various configurations of flanking noise bands (1 ERB bandwidth, 2 ERB gap between bands; Figure 5A). CMR was calculated as the difference in the detection threshold across the different flanker conditions. In-lab (pre-COVID-19) measurements using an adaptive procedure on $N = 40$ subjects with normal audiograms (and similar age as our Prolific participants) suggested that CMR > 3 dB for CORR - REF conditions, and > 12 dB for CORR - ACORR conditions. The psychometric functions obtained with identical stimuli using web-based measurements are shown in Figure 5B ($N = 203$). The CMR measured via web-based testing was about 4 dB for CORR - REF conditions, and about 17 dB for CORR - ACORR conditions, establishing that strong CMR effects are demonstrable with web-based testing as in lab-based tests.

Word-recognition in noise and consonant identification

Finally, we sought to compare word recognition performance and consonant categorization patterns measured using web-based testing with corresponding lab-based data from the literature. Figure 6A shows the word-recognition scores in background babble from the MRT test that was used for hearing screening. Overall performance levels and trends with SNR for participants who passed our web-based hearing-screening replicate lab-based results obtained from individuals with normal audiometric thresholds (Miner and Danhauer, 1976).

To probe consonant categorization and confusions (Miller and Nicely, 1955), participants were asked to identify consonants (i.e., C/a/ syllables) from the STeVI corpus (Sensimetrics Corporation, Gloucester, MA) presented in speech-spectrum stationary noise using the carrier phrase “You will mark ____ please”. An SNR of −8 dB was chosen, which yielded an overall intelligibility of 66% (and hence a sufficient number of confusions for analysis). Speech materials were available in the voices of four different talkers. Consonant-confusion data were obtained from about N=75 subjects for each of the four talkers, but subject overlap across the different talkers was not controlled. In total, consonant confusion matrices were obtained from a total of N=295 participants. The average confusion matrix across all N=295 measurements is shown in Figure 6B. This data was compared with controlled lab-based results using similar stimuli (Phatak and Allen, 2007). Phatak and Allen (2007) found that for a fixed overall intelligibility, recognition scores varied across different consonants. Specifically, they found that errors were most common on consonants {f, θ, v, ð, b, m}, fewer errors were made on {n, p, g, k, d}, and even fewer errors on {t, s, z, ʃ, ʒ }. They referred to these three groups of consonants as “C1” (low-score group), “C3” (intermediate-score group), and “C2” (high-score group), respectively. Our results replicate this trend for the specific consonant groups that Phatak and Allen (2007) identified based on their measurements. That is, the ordering of scores for the three consonant groups was replicated (Figure 6C). Separately, based on a graphical analysis of confusion patterns, Phatak and Allen (2007) also identified confusion *clusters*, i.e., sets of consonants where any consonant in a particular set is most confused with another consonant in the same set. To examine whether their clustering results are replicated in our web-based measurements, we used the probability of confusion between a pair of consonants (from Figure 6B) as the distance metric to perform hierarchical agglomerative clustering. With agglomerative clustering, each consonant starts out as its own cluster, and then clusters are progressively merged by slowly relaxing the distance criterion defining the clustering. When we set our distance threshold for agglomeration to yield the same *number* of clusters as reported in Phatak and Allen (2007), the clusters from our data were identical to their report. Specifically, the indentified confusion clusters were {f, v, b, θ, ð}, {s, z}, {ʃ, ʒ}, and {m, n}.

Summary and Discussion

Although web-based experiments are used in many domains of behavioral research, they have not received much adoption in psychoacoustics research. This is understandable given the challenges stemming from lack of control of stimulus hardware or listening environment, and from the inability to know the hearing status of anonymous participants. Psychoacoustic phenomena pertaining to low-level sensory processing of sounds is generally thought to be sensitive to such testing parameters. Here, we developed and implemented a range of strategies to mitigate these challenges and tested whether web-based psychoacoustics is in fact viable.

For hearing screening, we used a suprathreshold word-recognition-in-babble task and documented evidence for selectivity for near-normal hearing status. Specifically, we found that by setting specificity to around 65% (i.e., exclude about 35% of all participants), we can detect hearing loss of any degree with > 80% sensitivity, and detect more-than-mild hearing loss with > 95% sensitivity. We supplemented this hearing screening with procedures to check for stereo headphone use, focused engagement (i.e., not clicking on other tabs/windows during the experiment), screening for demographics characteristics (age and English language fluency status), and self-report of normal hearing (including no persistent tinnitus) and neurological status.

We then evaluated web-based psychoacoustics by comparing data from web-screened participants with lab-based data from individuals with known audiometric status, and found excellent agreement. Specifically, we found that with web-based testing, absolute sensitivity to diotic and binaural cues were robustly estimable, the main effects of changes in modulation statistics of the auditory scene were readily apparent, and that speech recognition and consonant category confusion patterns closely matched lab-based results.

In particular, F0 difference limens for web-based data were about 0.5%, gap detection thresholds in tonal markers were about 6 ms, ITD thresholds for low-frequency tones about 28 *μs*, ILD thresholds for high-frequency tones about 0.8 dB, and consonant confusions were in agreement with highly controlled studies. Indeed, when we quantitatively compared these psychoacoustic metrics between web- and lab-based studies, we found that the effect size of the difference was less than 0.2 (see Table 1). A effect size of at least 0.2 is conventionally thought to be a small effect (Ferguson, 2009). One exception to this was that the scores for word identification in babble using the MRT task were higher (better) in our web-based studies compared to a previous lab-based study with audiometrically normal-hearing subjects (Miner and Danhauer, 1976). This is likely because we used high performance in the MRT task as a criterion for putative normal-hearing status, thus likely excluding many participants with normal hearing but happen to fall within the lower end of the distribution of MRT scores. Indeed, the specificity of of hearing-screening procedure was designed to be around 65% (i.e., exclude 35% of participants).

Taken together, our results show that web-based psychoacoustics is viable and an excellent complement to lab-based work. Our screening procedures and psychoacoustic tasks are implemented using open-source tools (jsPsych and Django), which provide capabilities for trial-based experiment design with complex study flow. We make the source code for our custom web app available via GitHub for interested investigators to adapt to their needs. We also make a working demo of our infrastructure available. This infrastructure has since been successfully used in several recent studies (Viswanathan et al., 2021b; Singh and Bharadwaj, 2021; Viswanathan et al., 2022; Borjigin, 2022).

Remote psychoacoustic testing can help overcome many challenges with lab-based testing. One important example of a challenge that remote testing can help mitigate is the restrictions on in-person experiments during the COVID-19 pandemic. Other key advantages of remote testing include the ability to collect large datasets at lower costs, the ability to include diverse and representative cohorts of subjects, to reach and serve certain groups who may otherwise find themselves excluded from lab-based research in the convenience of their homes (Peng et al., 2022). Another important advantage is that large datasets accumulate at rapid rates with web-based studies compared to lab-based studies in that participants can perform the tasks in parallel. For instance, when we post a study on Prolific, we readily obtain participation from N=100 participants over a weekend even when we break down the data acquisition into smaller segments of 20–25 subjects in each posting. Web-based testing also allows for conducting studies that involve longitudinal training or monitoring over extended periods without the need to have subjects repeatedly visit the testing site. Indeed, web-based testing has been leveraged for longitudinal studies in other domains (Ruano et al., 2016). While a range of approaches can be adopted within the remote-testing umbrella (Peng et al., 2022), the present study focused on web-based testing on anonymous participants using their own devices and web browsers in whatever listening environments were available to them. Other remote-testing possibilities that allow for greater hardware control include shipping lab-calibrated hardware such as tablets/smartphones and a headphone to remote participants (Lelo de Larrea-Mancera et al., 2020), or having participants install apps on specific consumer-grade devices with known average acoustic characteristics. We chose to focus on browser-based testing here because it requires minimal co-operation from subjects for installing specialized software and is easily scalable to a range of devices and operating systems. Although browser-based testing is perhaps the least controlled of remote testing options, our results using carefully designed procedures show that excellent outcomes can be obtained with such web-based testing.

One potential limitation of our study is that our validation experiments focused on absolute sensitivity measures (e.g., F0DLs, ITD, etc.), performance scores (e.g., MRT), and within-subject effects (e.g., CMR). We did not directly assess the viability of web-based testing for between-subject comparisons (e.g., age effects or effects of hearing loss or other psychoacoustic outcomes). Although we found that our hearing-screening procedure can separate individuals with normal hearing and hearing loss, the procedure was designed for subject screening. Separate validation studies need to be conducted to examine whether between-subject differences can be reliability assessed with web-based studies. For instance, although age information is easy to obtain through a web-based survey, differences between age groups or different demographics may indicate differences in facility/experience in interacting with web-based technologies rather than differences in auditory processing *per se*. Furthermore, our study focused on whether mean sensitivities and within-subject effects and were not designed to directly address whether the distributions themselves are similar between web- and lab-based studies. Indeed, the limited control of the acoustics and the departure from conventional audiometric assessments could translate to greater variance in web-based scores, thus reducing between-subject effects of interest. Nonetheless, our results suggest that web-based studies can complement lab-based work in many ways.

Funding

This work was supported by National Institutes of Health (NIH) grants R01DC015989 (HMB), F31DC017381 (VV), and T32DC016853 (RS).

Data Availability Statement

Data to reproduce all figures will be made publicly on Figshare (https://figshare.com/) upon peer-reviewed publication.

References

Bell DW, Kreul EJ, Nixon JC (1972) Reliability of the modified rhyme test for hearing. Journal of speech and hearing research 15(2):287–295 [PubMed: 5047866]

Bharadwaj H (2021) SNAPlabonline, a Django-based web application for conducting psychoacoustics on the web from the Systems Neuroscience of Auditory Perception Lab (SNAPlab) [pre-print release]. Zenodo 10.5281/zenodo.4743850.

Bharadwaj HM, Masud S, Mehraei G, Verhulst S, Shinn-Cunningham BG (2015) Individual differences reveal correlates of hidden hearing deficits. J Neurosci 35(5):2161–2172 [PubMed: 25653371]

Blackwell DL, Lucas JW, Clarke TC (2014) Summary health statistics for us adults: national health interview survey, 2012. Vital and health statistics Series 10, Data from the National Health Survey (260):1–161 [PubMed: 24819891]

Borjigin A (2022) The role of temporal fine structure in everyday hearing. Purdue University Thesis DOI 10.25394/PGS.19673883.v1, URL https://hammer.purdue.edu/articles/thesis/The_Role_of_Temporal_Fine_Structure_in_Everyday_Hearing/19673883

Borjigin A, Hustedt-Mai AR, Bharadwaj HM (2022) Individualized assays of temporal coding in the ascending human auditory system. eNeuro 9(2), DOI 10.1523/ENEURO.0378-21.2022, URL https://www.eneuro.org/content/9/2/ENEURO.0378-21.2022, https://www.eneuro.org/content/9/2/ENEURO.0378-21.2022.full.pdf

Bregman A (1990) Auditory scene analysis: the perceptual organization of sound. Cambridge, MA: MIT Press

Bronkhorst A (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica 86:117–128

Brungart D, Makashay MJ, Summers V, Sheffield BM, Heil TA (2014) Assessing functional auditory performance in hearing-impaired listeners with an updated version of the modified rhyme test. The Journal of the Acoustical Society of America 135(4):2391–2391

Brungart DS, Makashay MJ, Sheffield BM (2021) Development of an 80-word clinical version of the modified rhyme test (mrt80). J Acoust Soc Am 149(5):3311–3327 [PubMed: 34241116]

Clifford S, Jerit J, et al. (2014) Is there a cost to convenience? an experimental comparison of data quality in laboratory and online studies. Journal of Experimental Political Science 1(2):120–131

Darwin C (1997) Auditory grouping. Trends Cogn Sci 1(9):327–333 [PubMed: 21223942]

De Leeuw JR (2015) jspsych: A javascript library for creating behavioral experiments in a web browser. Behavior research methods 47(1):1–12 [PubMed: 24683129]

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. Speech communication 41(2–3):331–348

Elkins EF (1971) Evaluation of modified rhyme test results from impaired-and normal-hearing listeners. Journal of speech and hearing research 14(3):589–595 [PubMed: 5163893]

Ferguson CJ (2009) An effect size primer: a guide for clinicians and researchers. Prof Psychol Res Pr 40(5):532–538

Giovannone N, Theodore RM (2021) Individual differences in lexical contributions to speech perception. J Speech Lang Hear 64(3):707–724

Glasberg BR, Moore BC (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47(1):103–138 [PubMed: 2228789]

Gosling SD, Mason W (2015) Internet research in psychology. Annual review of psychology 66:877–902

Hedges LV (1982) Estimation of effect size from a series of independent experiments. Psychol Bull 92(2):490

Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? Behavioral and brain sciences 33(2–3):61–83 [PubMed: 20550733]

House AS, Williams C, Hecker MH, Kryter KD (1963) Psychoacoustic speech tests: A modified rhyme test. J Acoust Soc Am 35(11):1899–1899

Jerger J, Brown D, Smith S (1984) Effect of peripheral hearing loss on the masking level difference. Arch Otolaryngol 110(5):290–296 [PubMed: 6712516]

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S (2004) Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America 116(4):2395–2405 [PubMed: 15532670]

Klumpp R, Eady H (1956) Some measurements of interaural time difference thresholds. J Acoust Soc Am 28(5):859–860

Lelo de Larrea-Mancera ES, Stavropoulos T, Hoover EC, Eddins DA, Gallun FJ, Seitz AR (2020) Portable automated rapid testing (part) for auditory assessment: Validation in a young adult normal-hearing population. J Acoust Soc Am 148(4):1831–1851 [PubMed: 33138479]

Lavan N, Knight S, Hazan V, McGettigan C (2019) The effects of high variability training on voice identity learning. Cognition 193:104026 [PubMed: 31323377]

Licklider J (1948) The influence of interaural phase relations upon the masking of speech by white noise. J Acoust Soc Am 20(2):150–159

Lin FR, Niparko JK, Ferrucci L (2011) Hearing loss prevalence in the United States. Arch Intern Med 171(20):1851–1853 [PubMed: 22083573]

Madsen SM, Whiteford KL, Oxenham AJ (2017) Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. Scientific Reports 7(1):1–9 [PubMed: 28127051]

McDermott JH, Lehr AJ, Oxenham AJ (2010) Individual differences reveal the basis of consonance. Current Biology 20(11):1035–1041 [PubMed: 20493704]

McPherson MJ, McDermott JH (2018) Diversity in pitch perception revealed by task dependence. Nature human behaviour 2(1):52–66

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some english consonants. J Acoust Soc Am 27(2):338–352

Mills AW (1960) Lateralization of high-frequency tones. The Journal of the Acoustical Society of America 32(1):132–134

Milne AE, Bianco R, Poole KC, Zhao S, Oxenham AJ, Billig AJ, Chait M (2020) An online headphone screening test based on dichotic pitch. Behavior Research Methods pp 1–12

Miner R, Danhauer JL (1976) Modified rhyme test and synthetic sentence identification test scores of normal and hearing-impaired subjects listening in multitalker noise. J Am Audiol Soc 2(2):61–67 [PubMed: 977431]

Moore BC (2007) Cochlear hearing loss: physiological, psychological and technical issues, 2nd edn. John Wiley & Sons

Moore BC (2012) An introduction to the psychology of hearing.

Brill Nilsson M, Soli SD, Sullivan JA (1994) Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am 95(2):1085–1099 [PubMed: 8132902]

Niquette P, Arcaroli J, Revit L, Parkinson A, Staller S, Skinner M, Killion M (2003) Development of the bkb-sin test. In: Annual meeting of the American Auditory Society, Scottsdale, AZ

Oxenham AJ (2018) How we hear: The perception and neural coding of sound. Annual Review of Psychology 69(1):27–50, DOI 10.1146/annurev-psych-122216-011635, URL 10.1146/annurev-psych-122216-011635, 10.1146/annurev-psych-122216-011635

Parthasarathy A, Pinto SR, Lewis RM, Goedicke W, Polley DB (2020) Data-driven segmentation of audiometric phenotypes across a large clinical cohort. Scientific reports 10(1):1–12 [PubMed: 31913322]

Patro C, Kreft HA, Wojtczak M (2021) The search for correlates of age-related cochlear synaptopathy: Measures of temporal envelope processing and spatial release from speech-on-speech masking. Hear Res 409:108333 [PubMed: 34425347]

Peer E, Brandimarte L, Samat S, Acquisti A (2017) Beyond the turk: Alternative platforms for crowd-sourcing behavioral research. Journal of Experimental Social Psychology 70:153–163

Peng ZE, Waz S, Buss E, Shen Y, Richards V, Bharadwaj H, Stecker GC, Beim JA, Bosen AK, Braza MD, et al. (2022) Remote testing for psychological and physiological acoustics. J Acoust Soc Am 151(5):3116–3128 [PubMed: 35649891]

Phatak SA, Allen JB (2007) Consonant and vowel confusions in speech-weighted noise. The Journal of the Acoustical Society of America 121(4):2312–2326 [PubMed: 17471744]

Ruano L, Sousa A, Severo M, Alves I, Colunas M, Barreto R, Mateus C, Moreira S, Conde E, Bento V, et al. (2016) Development of a self-administered web-based test for longitudinal cognitive assessment. Scientific Reports 6(1):1–10 [PubMed: 28442746]

Schooneveldt GP, Moore BC (1989) Comodulation masking release (cmr) as a function of masker bandwidth, modulator bandwidth, and signal duration. The Journal of the Acoustical Society of America 85(1):273–281 [PubMed: 2921409]

Schütt HH, Harmeling S, Macke JH, Wichmann FA (2016) Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. Vision research 122:105–123 [PubMed: 27013261]

Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34(3):114–123 [PubMed: 21196054]

Singh R, Bharadwaj H (2021) Cortical temporal integration window for binaural cues accounts for sluggish auditory spatial perception. bioRxiv 20211214472656 DOI 10.1101/2021.12.14.472656

Smits C, Kapteyn TS, Houtgast T (2004) Development and validation of an automatic speech-in-noise screening test by telephone. Int J Audiol 43(1):15–28 [PubMed: 14974624]

Steinmetzger K, Zaar J, Relaño-Iborra H, Rosen S, Dau T (2019) Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations. The Journal of the Acoustical Society of America 146(4):2562–2576 [PubMed: 31671986]

Viswanathan V, Bharadwaj HM, Shinn-Cunningham BG, Heinz MG (2021a) Modulation masking and fine structure shape neural envelope coding to predict speech intelligibility across diverse listening conditions. J Acoust Soc Am 150(3):2230–2244 [PubMed: 34598642]

Viswanathan V, Shinn-Cunningham BG, Heinz MG (2021b) Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble. J Acoust Soc Am 150(4):2664–2676 [PubMed: 34717498]

Viswanathan V, Shinn-Cunningham BG, Heinz MG (2022) Speech categorization reveals the role of early-stage temporal-coherence processing in auditory scene analysis. J Neurosci 42(2):240–254 [PubMed: 34764159]

Watson CS, Kidd GR, Miller JD, Smits C, Humes LE (2012) Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a us version. J Am Acad Audiol 23(10):757–767 [PubMed: 23169193]

Whiteford KL, Kreft HA, Oxenham AJ (2020) The role of cochlear place coding in the perception of frequency modulation. Elife 9:e58468 [PubMed: 32996463]

Wilson RH (2003) Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. J Am Acad Audiol 14(9):453–470 [PubMed: 14708835]

Woods AT, Spence C, Butcher N, Deroy O (2013) Fast lemons and sour boulders: Testing crossmodal correspondences using an internet-based testing methodology. i-Perception 4(6):365–379 [PubMed: 24349696]

Woods AT, Velasco C, Levitan CA, Wan X, Spence C (2015) Conducting perception research over the internet: a tutorial review. PeerJ 3:e1058 [PubMed: 26244107]

Woods KJ, McDermott JH (2018) Schema learning for the cocktail party problem. Proc Natl Acad Sci USA 115(14):E3313–E3322 [PubMed: 29563229]

Woods KJ, Siegel MH, Traer J, McDermott JH (2017) Headphone screening to facilitate web-based auditory experiments. Attention, Perception, & Psychophysics 79(7):2064–2072

Zhao S, Yum NW, Benjamin L, Benhamou E, Yoneya M, Furukawa S, Dick F, Slaney M, Chait M (2019) Rapid ocular responses are modulated by bottom-up-driven auditory salience. Journal of Neuroscience 39(39):7703–7714 [PubMed: 31391262]

Zwislocki J, Feldman R (1956) Just noticeable differences in dichotic phase. The Journal of the Acoustical Society of America 28(5):860–864

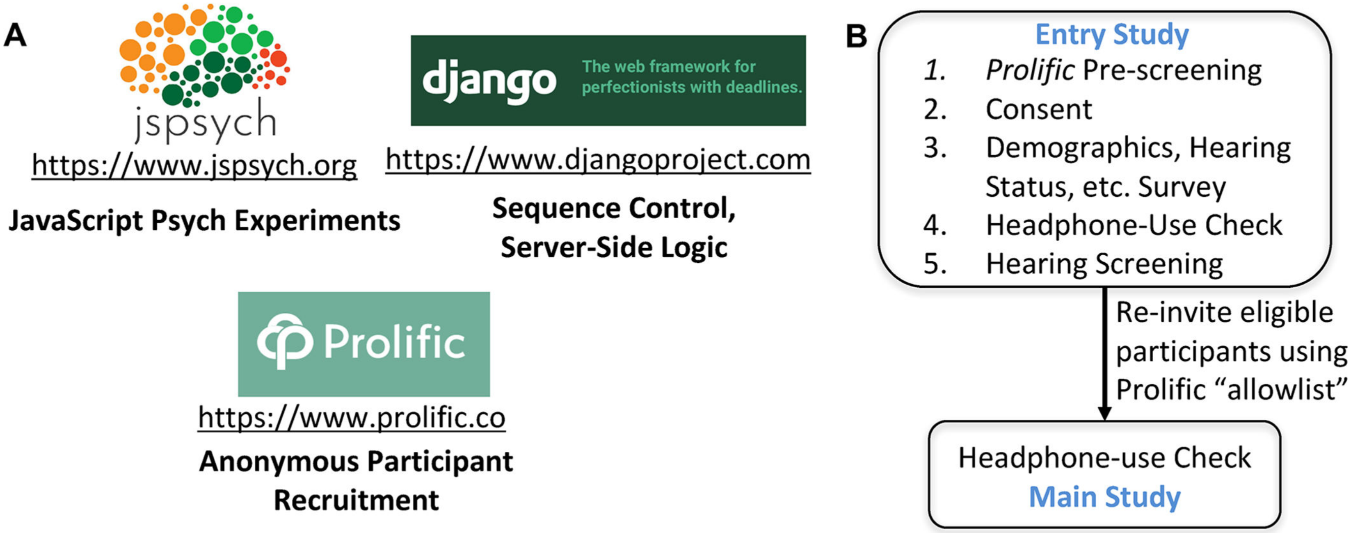
A
jspsych
https://www.jspsych.org
JavaScript Psych Experiments
django The web framework for perfectionists with deadlines.
https://www.djangoproject.com
Sequence Control, Server-Side Logic
Prolific
https://www.prolific.co
Anonymous Participant Recruitment
B
Entry Study
1. Prolific Pre-screening
2. Consent
3. Demographics, Hearing Status, etc. Survey
4. Headphone-Use Check
5. Hearing Screening
Re-invite eligible participants using Prolific "allowlist"
Headphone-use Check
Main Study

Figure 1. Components of the infrastructure for web-based psychoacoustics.

(A) Individual tasks were created using jsPsych, an open-source javascript front-end library. Dynamic sequences of psychoacoustic tasks were implemented using Django, an open-source library for web applications, and combined with consent pages, questionnaires, and debriefing pages. Subjects were recruited via Prolific.co, a platform for recruiting human subject for web-based studies. (B) A number of screening procedures were implemented to select participants for stereo headphone use, (putative) normal-hearing status, and attentive engagement via an entry study. Participants meeting criteria were re-invited to participate in the validation experiments using Prolific's allowlist feature.

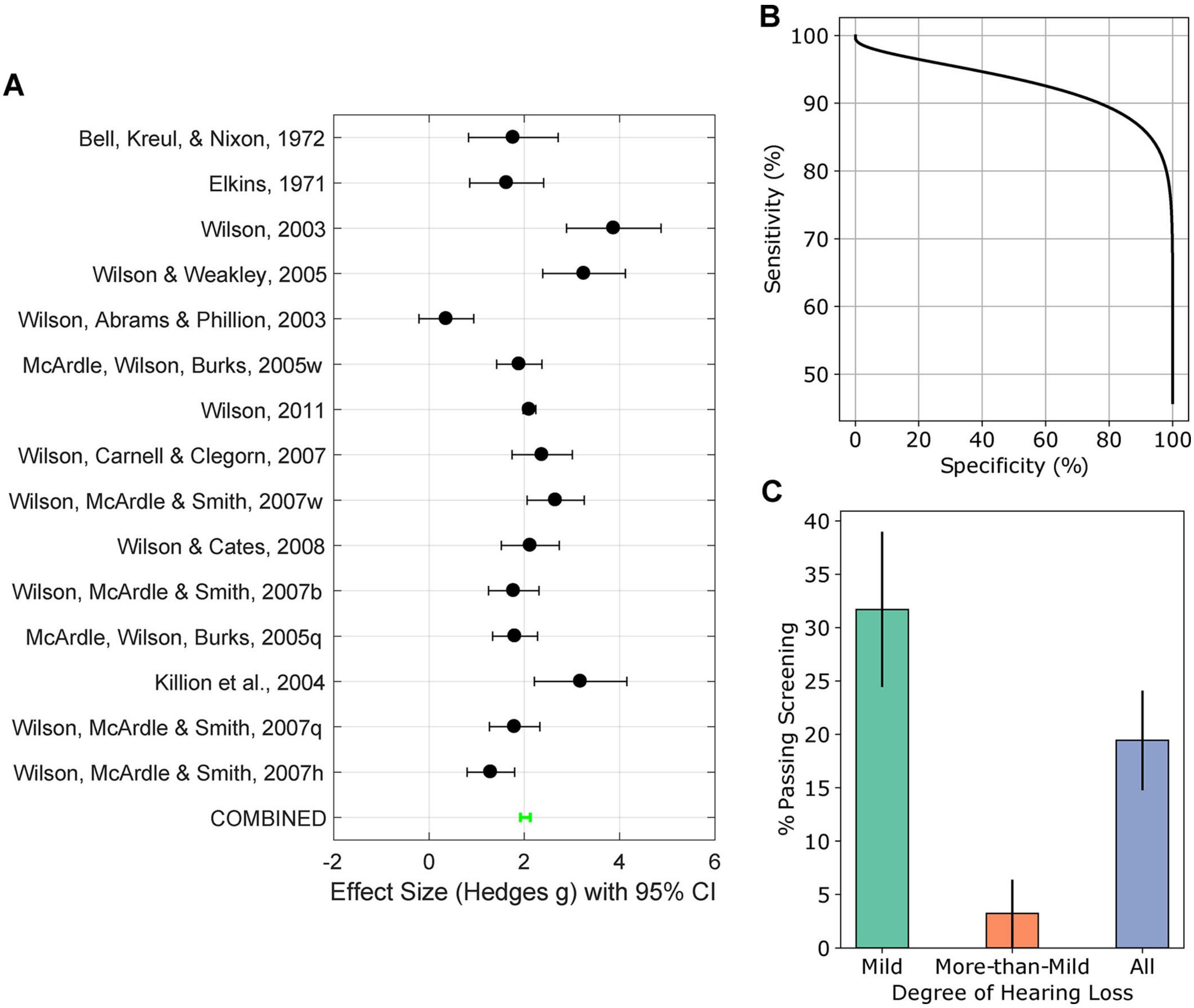

Figure 2. Validation of the hearing screening procedure.

(A) A meta-analyis of 15 studies in the literature suggested that speech-in-noise tasks yield a large effect size separating individuals with normal hearing (NH) and hearing loss (HL). (B) Theoretical calculations based on results from panel (A) predict the expected tradeoff between sensitivity and specificity for the detection of hearing loss. This was used to determine cutoff scores for the hearing screening procedure. (C) Actual selectivity of our hearing screening procedure was quantified in a cohort of subjects with a professional (clinical) diagnosis of hearing loss. Sensitivity for any degree of HL was about 81%. Sensitivity for more-than-mild HL (i.e., moderate, moderately-severe, or severe) was > 95%. Error bars in panel (C) indicate standard deviation of estimated pass rates from $N = 72$ subjects.

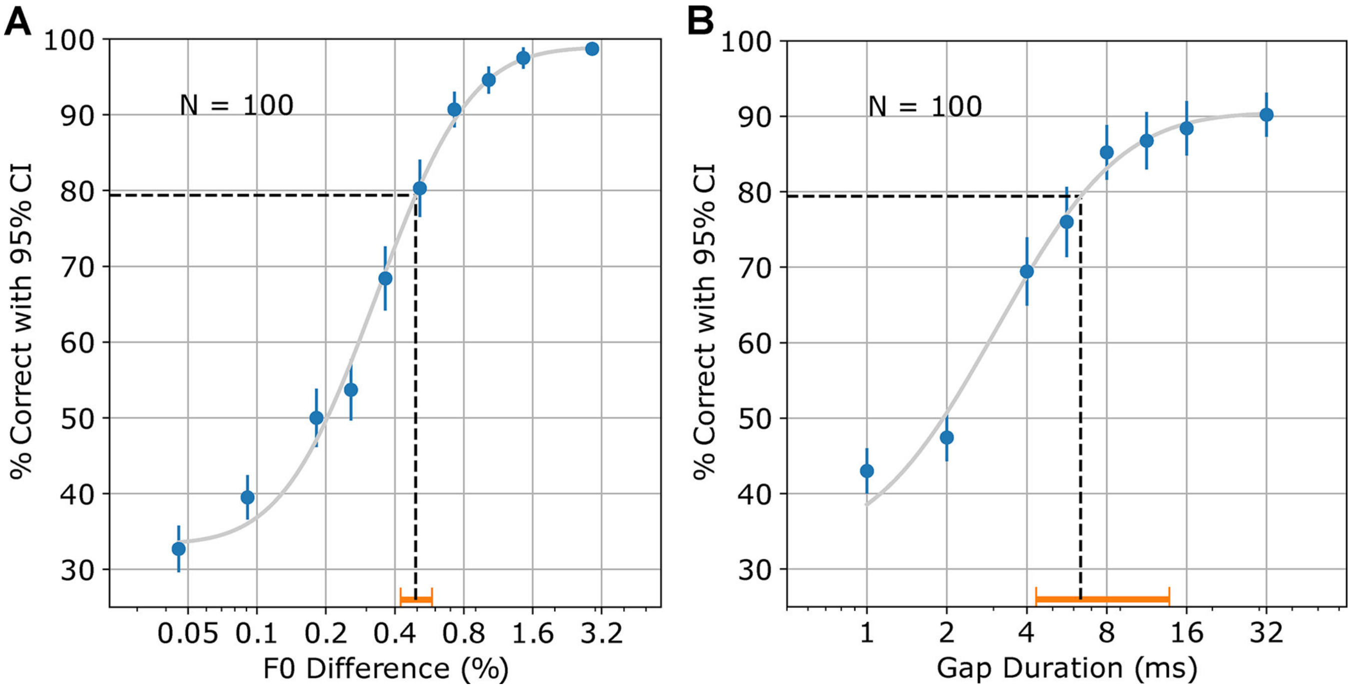

Figure 3. Validation of web-based testing for classic psychoacoustic tasks with diotic stimuli.

Psychometric functions were measured for fundamental-frequency (F0) discrimination (A), and gap detection (B). F0 discmination limens were around 0.5% for F0 = 110 Hz, and gap thresholds were around 6 ms, consistent with lab-based data with similar stimuli. Horizontal error bars in orange indicate the 95% confidence interval of the mean threshold estimated from $N = 100$ subjects.

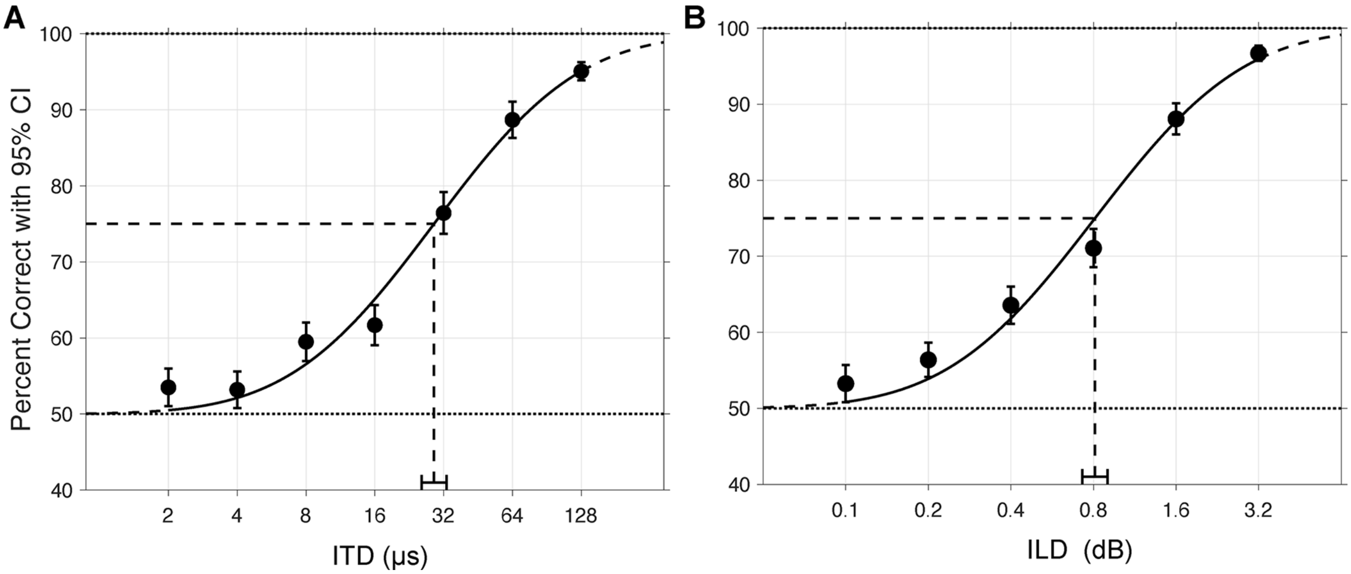

Figure 4. Validation of web-based testing for classic interaural difference detection tasks.
Psychometric functions were measured for interaural time-delay (ITD) detection (A), and interaural level-difference (ILD) detection (B). ITD thresholds were around 28μs for 500 Hz tone bursts, and ILD thresholds were around 0.8 dB for 4 kHz tone bursts, consistent with classic literature. Horizontal error bars indicate the 95% confidence interval of the mean threshold estimated from $N = 200$ subjects.

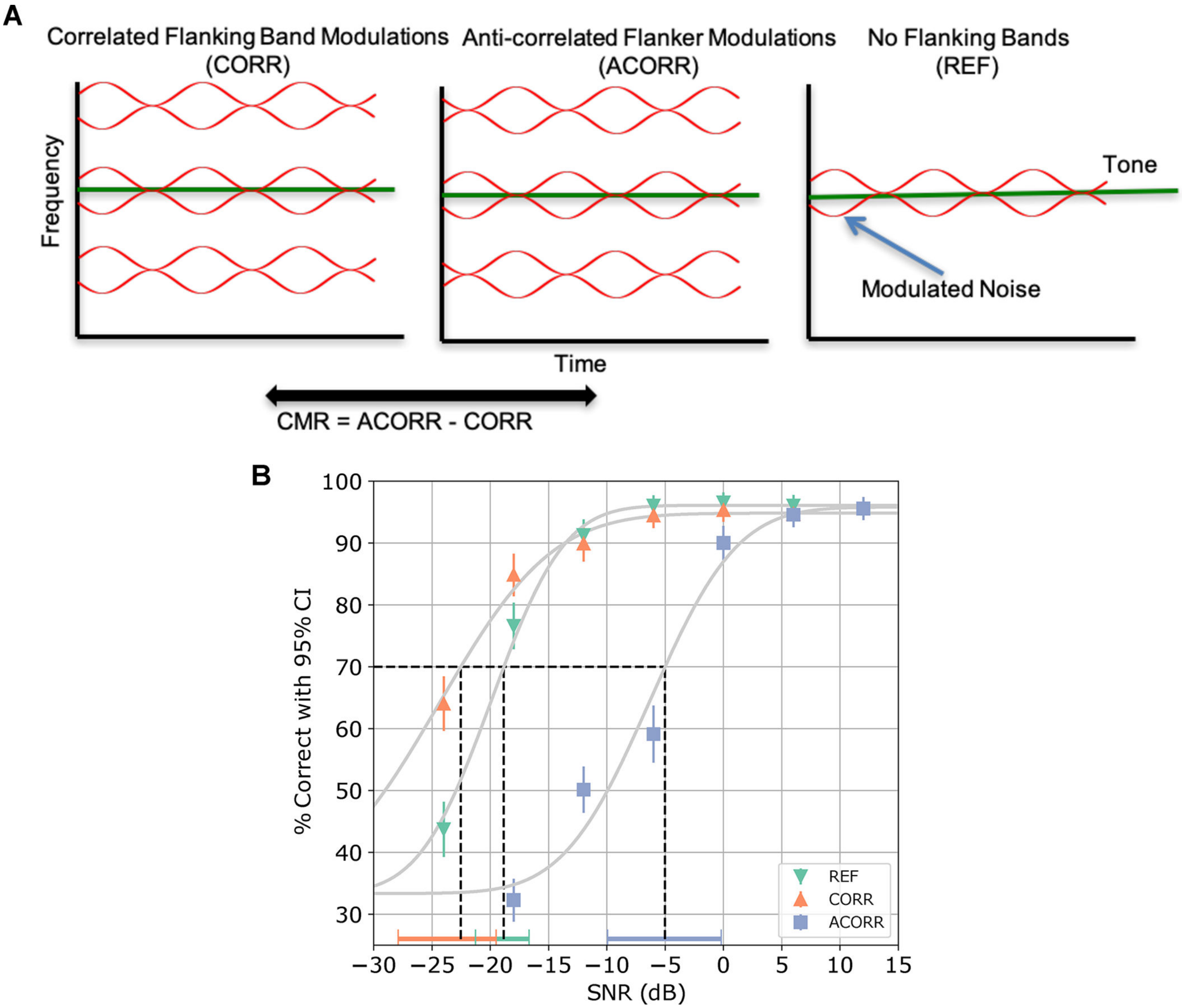

Figure 5. Validity of web-based testing for demonstrating the co-modulation masking release (CMR) effect.

(A) CMR was measured by assessing the SNR thresholds for detecting 4 kHz tones in modulated 1-ERB-wide on-band noise for varying configurations of 1-ERB-wide noise flankers (2-ERB gap between on-band noise and flankers). The flankers were either absent (REF), modulated in a correlated manner with the on-band noise (CORR), or anticorrelated manner with the on-band noise (ACORR). The change in tone thresholds across conditions is quantified as the CMR effect. (B) Psychometric curves for tone detection in different CMR conditions was measured from N=203 subjects, yielding a clear separation between conditions. CMR for CORR-REF was about 4 dB, whereas CMR for CORR - ACORR was about 17 dB consistent with lab-based measurements with identical stimuli. Horizontal errorbars indicate 95% confidence intervals for the tone threshold in each condition.

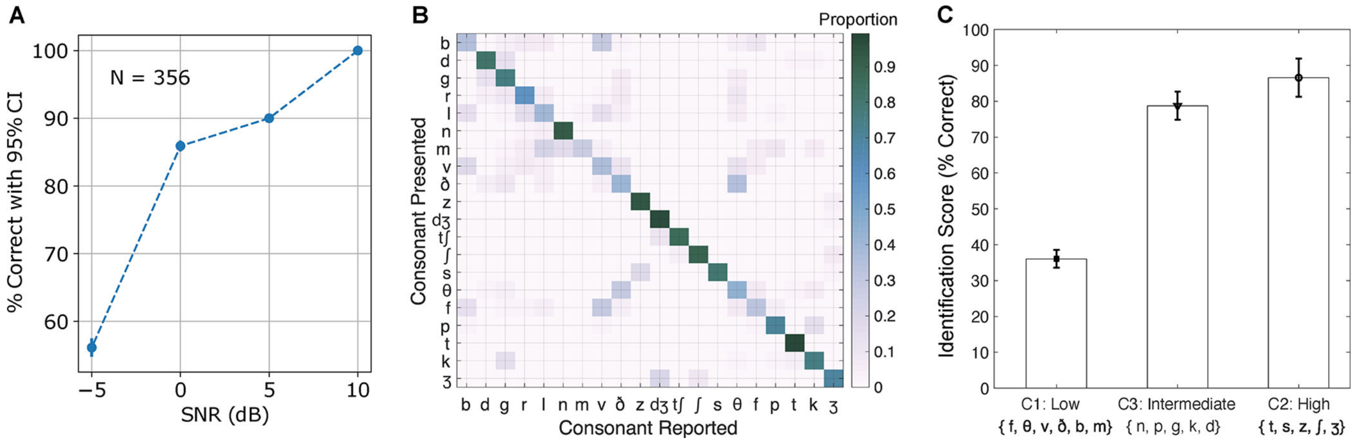

Figure 6. Validation of web-based testing using word-recognition scores and consonant confusion patterns.

(A) Recognition scores for monosyllabic words embedded in spectrally matched (in the long-term average sense) 4-talker babble showed overall performance and SNR dependence consistent with literature (Miner and Danhauer, 1976). (B) Consonant confusion matrix for consonants presented in C/a/ context in speech-shaped stationary noise at −8 dB SNR from $N = 295$ samples across four talker voices. Confusion clusters extracted from this matrix match lab-based results (Phatak and Allen, 2007). (C) A summarization of diagonal entries from panel (B) revealed that recognition scores were not uniform across consonants. Differences across sets of consonants with low, intermediate and high scores match lab-based data (Phatak and Allen, 2007).

Table 1.

Comparison of results from web-based and lab-based measurements. To obtain the reference values, the median was used as an estimate of μ when individual data points or median were reported in the reference study, and sample mean values were used if those were reported instead. The sample standard error of the mean (SEM) along with the number of participants was used to estimate σ, unless standard deviations were directly reported in the reference study. The effect size of the difference (i.e., Lab - Web values) was estimated using the procedure described in Hedges (1982). The values obtained using our infrastructure were indistinguishable except for MRT-in-babble where the performance scores on the web-based measure were higher (better).

Task	Web data ($\mu \pm \sigma$)	Reference source	Reference values ($\mu \pm \sigma$)	Lab - Web difference (Hedges' g ±σ)
Gap detection	6.1 ± 9 ms	Patro et al. (2021)	5.9±4.15 ms	−0.02 ± 0.19
F0 discrimination	0.5 ± 0.79%	Madsen et al. (2017)	0.51 ± 0.8%	0.01 ± 0.16
ITD detection	28 ± 19 μs	Borjigin et al. (2022)	29 ± 11 μs	0.06 ± 0.17
ILD detection	0.8 ± 0.7 dB	Mills (1960)	0.7 ± 0.4 dB	−0.14 ± 0.45
MRT-in-babble (0 dB SNR)	85 ± 15%	Miner and Danhauer (1976)	75 ± 10%	−0.68 ± 0.18