Robust and efficient online auditory psychophysics with the right auditory hygiene

Sijia Zhao[a], Christopher A. Brown[b], Lori L. Holt[c,d], and Frederic Dick[e,f]

a. Department of Experimental Psychology, University of Oxford, Oxford OX1 3PH, UK sijia.zhao@psy.ox.ac.uk
b. Department of Communication Science and Disorders, University of Pittsburgh, 4028 Forbes Tower Pittsburgh, PA 15260, cbrown1@pitt.edu
c. Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15232, loriholt@cmu.edu
d. Neuroscience Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15232, loriholt@cmu.edu
e. Department of Psychological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK, f.dick@bbk.ac.uk
f. Department of Experimental Psychology, PALS, University College London, 26 Bedford Way, London, WC1H 0AP, UK, f.dick@ucl.ac.uk

Corresponding Author:
Sijia Zhao
sijia.zhao@psy.ox.ac.uk
Department of Experimental Psychology, University of Oxford, Oxford OX1 3PH

Acknowledgements
We thank Christi Gomez, Erin Smith, and Sydney Sepkovic for their assistance in collecting in-person data. We thank Dr. Alessandro Rinaldo, Carnegie Mellon University Department of Statistics and Data Science, for statistical consultation.

Data Availability
Test implementations of the volume setting (Expt 1) are available in JavaScript [https://gitlab.pavlovia.org/sijiazhao/volumechecking_demo] and Gorilla [https://gorilla.sc/openmaterials/261557]. A demo can be found on Pavlovia [https://run.pavlovia.org/sijiazhao/volumechecking_demo].

The guide to implement the staircase procedure (Expt 2a) is available at [https://sijiazhao.github.io/how-to-staircase/]. A demo is available on Pavlovia [https://run.pavlovia.org/sijiazhao/threshold_demo/] with publicly available in-house code [https://gitlab.pavlovia.org/sijiazhao/threshold_demo].

The probe-signal task (Expt 2b) can be tried at [https://run.pavlovia.org/sijiazhao/probesignal_demo] and its code can be found at [https://gitlab.pavlovia.org/sijiazhao/probesignal_demo].
The raw data of this study are available on GitHub [https://github.com/sijiazhao/TPS_data].

Abstract

Most human auditory psychophysics research has been conducted with extreme ‘auditory hygiene’ in carefully controlled environments, with calibrated audio equipment, and potentially hours of repetitive testing with expert listeners. The incompatibility of web-based platforms with such experimental regimes would seem to preclude online auditory psychophysical paradigms, where success may hinge on absolute sound presentation level, reliably estimated perceptual thresholds, and sustained motivation and attention. Here, we introduce and validate a set of procedures that address these challenges and facilitate successful online auditory psychophysics. First, we establish a simple means of setting sound presentation levels where online participants serve as their own sound level meter. Across a set of three experiments conducted in person, we demonstrate the stability and robustness of this volume setting procedure both ‘in the wild’ and in controlled settings. Second, we test participants' tone-in-noise thresholds using widely adopted online experiment platforms, and demonstrate that reliable threshold estimates can be derived in approximately one minute of testing. Third, using these sound level setting and thresholding procedures to establish participant-specific stimulus conditions, we show that frequency-selective attention can be reliably demonstrated in individual participants with an online implementation of the classic, yet challenging, probe signal psychophysics paradigm. Finally, we show how threshold and attentional measures relate to well-validated assays of participants' in-task motivation, fatigue, and confidence. This demonstrates the promise of asking new questions in auditory neuroscience with classic, yet challenging, online psychophysics paradigms. Code for implementing the tests is publicly available through Pavlovia (pavlovia.org) and Gorilla (gorilla.sc).

Keywords

Online testing; Probe-signal; Psychophysics; Auditory Thresholds; Motivation

Introduction

Much of what we know about the function of the auditory system is thanks to a half-century of clever and carefully designed auditory psychophysical behavioral paradigms in human listeners. The methodologies of auditory psychophysics - and indeed, much of auditory behavioral neuroscience - tend to rely on strongly sound-attenuated environments, finely calibrated equipment, and small numbers of expert or highly trained listeners who are motivated and compliant with task demands. This high level of 'auditory hygiene' is important: seemingly minute differences in stimulus delivery and timing, background noise levels, or participant engagement during an arduous thresholding procedure can dramatically affect experimental results (D. M. Green, 1995; Manning, Jones, Dekker, & Pellicano, 2018; Rinderknecht, Ranzani, Popp, Lambercy, & Gassert, 2018).

The COVID pandemic has challenged our assumptions about auditory hygiene. With lab facilities inaccessible, researchers have been prompted to abandon their deep traditions of tight experimental control, expert subjects, and carefully calibrated acoustic environments to embark on the seemingly foolhardy endeavor of online auditory experimentation. This disruption has inspired unexpectedly successful innovations in porting auditory research to participants' home offices and living rooms, using only the internet bandwidth, computers, and headphones at participants' immediate disposal. For instance, several groups have created clever tests for ensuring participants are using headphones rather than speakers (Milne et al., 2020; Woods, Siegel, Traer, & McDermott, 2017), and that they are engaging with the experimental task, rather than haphazardly pressing buttons (Bianco, Mills, de Kerangal, Rosen, & Chait, 2021; Mok et al., 2021; Zhao et al., 2019). Such innovations notwithstanding, uncontrolled online experimental situations are particularly challenging for auditory paradigms that deliver stimuli at a given sound pressure level or that require sustained vigilance to respond consistently to an ever more difficult-to-perceive target sound, as is often found in psychophysical thresholding procedures.

Control of the range of sound pressure levels is important not only for ensuring participants' well-being (to make sure they are not exposing themselves to overly loud sounds) but also because neuronal responses from the cochlea to cortex are known to differ as a function of overall level. For instance, subpopulations of auditory nerve fibers differing in spontaneous firing rates respond at different acoustic stimulation levels (Horst, McGee, & Walsh, 2018; Taberner & Liberman, 2005). Across the peripheral and central auditory systems, single neuronal responses tend to be level-dependent, with frequency selectivity typically broadening with increasing sound amplitude levels (Bizley, Nodal, Nelken, & King, 2005; Schreiner, Read, & Sutter, 2000); behaviorally derived critical bands have also been shown to be level-dependent (Glasberg & Moore, 2000; Pick, 1980).

Obtaining reliable auditory psychophysical thresholds can be challenging even in laboratory conditions with experienced and motivated adult listeners. For example, quiet thresholds have been shown to be affected by the duration of time spent in a 'quiet' environment (Bryan, Parbrook, & Tempest, 1965; Steed & Martin, 1973) such as an audiometric booth. Even supra-threshold detection tasks performed by experienced listeners can be affected by presentation level (Williams, Elfner, & Howse, 1978). Converging on reliable psychoacoustical thresholds may seem nigh unto impossible with inexperienced listeners

(Kopiez & Platz, 2009) or in the presence of distracting events (Ruggles, Bharadwaj, & Shinn-Cunningham, 2011) typical of a home environment.

The potentially reduced levels of engagement and vigilance due to online listeners' motivation, fatigue, and confidence inject additional noise and bias into threshold estimates. This is particularly the case if the thresholding paradigms required to set levels for the actual experiments of interest are themselves tedious and unrewarding. Multiple long thresholding tracks also add considerable expense to online experiments, which tend to rely on shorter experimental sessions with larger numbers of participants to compensate for participant variability.

Here, we outline novel approaches for improving auditory hygiene in online psychoacoustics experiments that necessitate threshold measurements. In Experiment 1, we first ask whether we can reliably control the range of experimental stimulus amplitude levels (within ± 10 dB SPL) in online testing conducted in uncontrolled environments by asking online participants to set the volume on their own computer. In this way, participants act as their own sound level meter. This is quite a departure from the exquisite control afforded in laboratory studies.

We examine the effectiveness of this online approach with in-person tests and careful laboratory measurements on standardized equipment. In Experiment 1, participants recruited and tested in public outdoor spaces adjust the volume setting of the same consumer laptop and headphones to just-detectable levels when listening to a segment of broadband Gaussian white noise generated to have a particular root-mean-square (RMS) amplitude. The RMS level of this noise segment serves as the reference for experimental stimuli differing in RMS amplitudes. We then measure the range of at-ear dB SPL levels of this Gaussian white noise using the same hardware setup. This provides us with a ‘ground truth’ against which to compare in-person participants’ volume settings in terms of dB SPL. It also represents the worst-case scenario in testing environments - in that the outdoor testing area was noisy (between 57-67 dB with many distractors and people) and participants were heterogeneous community members.

In Experiment 2, we ask whether the volume setting procedure established in Experiment 1 is sufficient to allow for reliable online threshold estimates for a tone-in-noise detection task conducted ‘in the wild’. In an online experiment, we roll out the volume setting paradigm validated in Experiment 1, then ask whether small tweaks to standard thresholding procedures for a classic psychophysical task - tone detection in white noise - will permit fast (2-3 minute) and reliable estimation of thresholds among participants recruited and tested online.

In Experiment 2a we evaluate:1) the reliability of estimates over three short (40-trial) staircase-based thresholding tracks; 2) whether a simple estimator of psychophysical threshold - the mode of levels across a thresholding track - is as robust or more robust at estimating threshold as traditional estimators based on staircase reversals; and 3) whether and how online psychophysical thresholds are related to established assays of participant fatigue, apathy, and task confidence.

In Experiment 2b, we apply the online tone-in-noise thresholding procedure from Experiment 2a to measure frequency-selective auditory attention using a classic probe-signal paradigm (Botte, 1995; Dai, Scharf, & Buus, 1991; Greenberg & Larkin, 1968; Moore, Hafter, & Glasberg, 1996; Scharf, Quigley, Aoki, Peachey, & Reeves, 1987). After completing the online thresholding procedure of Experiment 2a, the same online participants heard continuous noise in which an above-threshold tone was followed by two listening intervals. Participants reported the interval in which a near-threshold tone was embedded in the noise, with the tone frequency matching the cue on 75% of trials and mismatching the cue at one of four other frequencies on 25% of the trials. We sought to determine whether patterns of frequency-selective attention: 1) can be replicated in uncontrolled online testing environments with inexpert listeners; 2) are evident in the short testing sessions necessitated by online testing; 3) change and develop over testing trials; and 4) are related to established assays of participant fatigue, apathy, and task confidence.

We provide code for each of these approaches to facilitate improved ‘auditory hygiene’ in online experiments, and to demonstrate the possibilities for asking new questions in auditory neuroscience with classic, yet challenging, online psychophysical paradigms.

Experiment 1

In Experiment 1, we examine the extent to which it is possible to reliably control the range of experimental stimulus amplitude levels in online testing conducted in uncontrolled environments. Our approach involves having online participants set the volume on their own computer so that they serve as their own sound level meter in the context of their unique testing equipment and acoustic environment. The RMS amplitude of the white noise stimulus used for setting this detection threshold is then used as a reference value for setting the amplitude of subsequent experimental stimuli during the same session.

Methods

Participants

Validation of the online volume setting procedure required testing in-person participants on a common consumer laptop with consumer headphones. For Experiments (Expts) 1a and 1b, recruitment was primarily conducted via informal in-person solicitation in outdoor environments due to COVID-related restrictions on indoor activities, and because the total task duration was approximately 2 minutes. For Expt 1a, participants (N=24) were recruited in an open lawn on the Carnegie Mellon University campus; a subset of participants were graduate students at a departmental gathering, others were undergraduate students as well as parents visiting for graduation ceremonies. Expt 1a participants were asked only whether they were at least 18 years of age, and considered their hearing to be within normal ranges, similar to the information that is solicited in many online studies. For Expt 1b, participants (N=28) were recruited in a central Pittsburgh park from a more heterogenous pool; here, participants were asked to note their age (mean age = 27.9 years (SD 10.2), ranging

between 18 and 55 years). One of these participants mentioned that they occasionally wore hearing aids. For Expt 1c, all participants were Carnegie Mellon or University of Pittsburgh students or staff (N=20; mean age = 30.1 years (SD 9.2), ranged between 17 and 47 years of age); here the same individuals were tested in the outdoor environment as well as in an anechoic sound booth under well-controlled laboratory conditions.

For all experiments, outdoor sound levels were measured using a Lutron SL-4022 (Expt 1a) or a Radio Shack Cat No. 33-2055 sound level meter (Expt 1b and 1c). For Expt 1a, Baseline average sound levels were ~50 dB SPL A-weighting; for Expt 1b, they were somewhat higher, with an average of ~57dB SPL, ranging between ~53-67 dB SPL. For Expt 1c, sound levels were an average of 28 dB SPL indoors in the anechoic chamber (using a Bruel & Kjaer 2231 precision sound level meter) and 57 dB SPL outdoors. As with many real-world listening environments, the outdoor environments included frequent sound events of somewhat higher amplitude (bird chirps, conversations of passing people, motorized skateboards, and helicopters flying overhead).

The study was approved by the Birkbeck College ethics committee, and took approximately 2-3 minutes to complete, including reading and completing the consent form, reading instructions, and setting the volume.

Stimuli, Procedure, and Calibration

Using Praat 6.0.17 (Boersma & Weenink, 2021) a 1-second Gaussian white noise was generated, band-pass filtered between 80-8000 Hz, and adjusted to achieve a nominal RMS level of 0.000399. The audio data were stored in the WAV file format, then exported in Sox (http://sox.sourceforge.net/) to a stereo sound file in the lossless FLAC format. 100-ms cosine ramps were added to both ends of the noise, so when it played on a continuous loop without gaps, it would sound like a sequence of gentle noise bursts. This stimulus file serves as the RMS amplitude reference for the sound stimuli in Expts 2 and 3.

Stimuli were presented using Beyer DT-150 circumaural headphones connected to the on-board headphone jack of an older consumer-grade Apple laptop computer (MacBook Pro, mid 2012) running macOS 10.15.7. Sounds were presented with the Pavlovia.org (Peirce et al., 2019) online experimental platform using Google Chrome version 09.0.4430.212 via wireless connections to various broadband providers.

Written instructions presented on the laptop screen asked participants to set the computer's volume to about 50% and then to click a button labeled 'play' to hear the pulsed noise played on a continuous loop until the participant pressed pause or proceeded to the next page. The continuous loop was achieved by in-house JavaScript code and no gap was inserted between the repetitions. Next, participants were instructed to use the computer's volume setting buttons to adjust the volume to a level at which the noise was barely audible. Specific instructions directed participants to slowly lower the volume until they could no longer hear the noise, and then to increase the volume a very small amount until they could again just barely hear the noise. After the participant was satisfied with their setting, the experimenter manually entered the final volume. As with many computers, the Mac volume buttons permit

a discrete range of percentage values. The only possible percentage settings were [0 6 12 19 25 31 38 44 50 56 62 69 75 81 88 94 100]. A demonstration of the procedure is available at [https://run.pavlovia.org/sijiazhao/volumechecking_demo]. The implementation is available in JavaScript [https://gitlab.pavlovia.org/sijiazhao/volumechecking_demo] and via the Gorilla experimental platform [https://gorilla.sc/openmaterials/261557].

For each volume setting on the MacBook Pro, dB SPL measurements were obtained using a Bruel & Kjaer 2231 precision sound level meter and Bruel & Kjaer 4152 artificial ear, coupled to the same set of Beyer DT-150 headphones used for data collection, while the stimuli were played with exactly the same procedure as used for participant testing. A headphone calibration with the Gaussian white noise stimulus was conducted in an anechoic chamber (a small room treated with 14-in fiberglass wedges on all six surfaces, which is anechoic down to 220 Hz) located on the University of Pittsburgh campus.

Results and Discussion

Calibration

The ambient noise floor in the anechoic chamber was measured to be about 28 dB SPL. In order to estimate the sound pressure level of the noise at low volume settings, pure tones were generated with several zero-to-peak amplitude values, and SPL values were then recorded at all volume settings. Volume setting was determined to be linear on the MacBook Pro used in the volume setting experiment; a given change in volume setting generated a consistent change in dB SPL at both high and low overall levels. This result gave us confidence that we could extrapolate downward to and below the noise floor. The noise segment was played at relatively high-volume settings to ensure accurate SPL measurements, and SPL values at low-volume settings were extrapolated. The results of this acoustic analysis indicated that the highest volume setting (100%) produced a presentation level of 46.4 dB SPL, and the lowest (6%) corresponded to 3.5 dB SPL. Figure 1 shows dB SPL values for each volume setting.

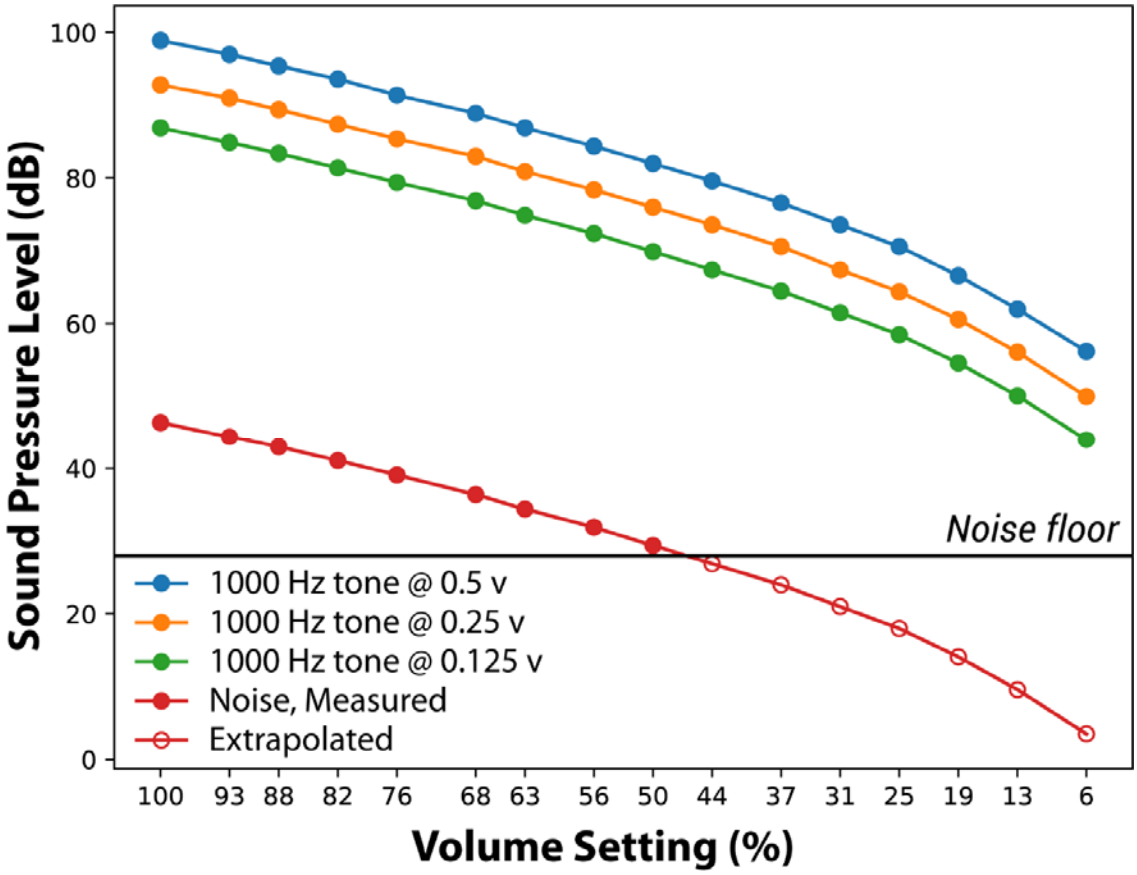


Figure 1. Sound pressure levels of the noise segment as a function of computer volume setting percentages. *Calibration was established in an anechoic chamber; the same stimulus, computer and headphones were used for the perceptual thresholding of Experiments 1a-1c.*

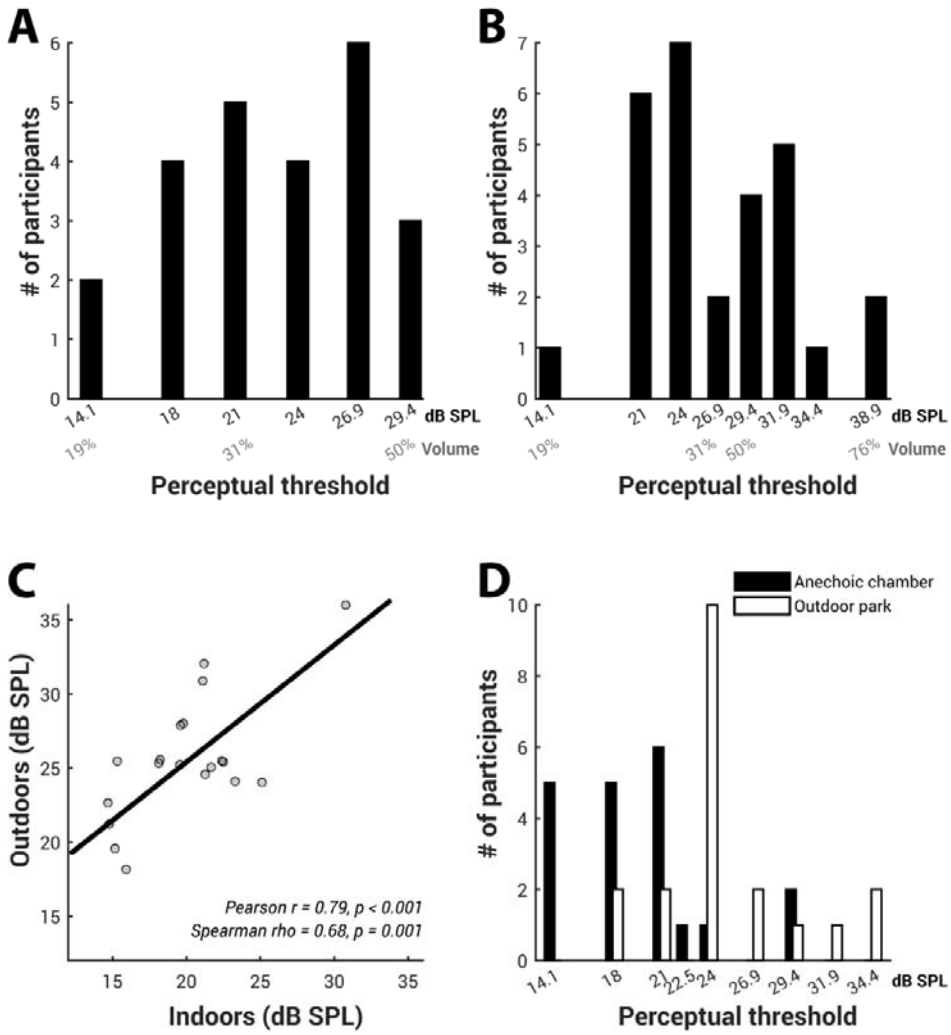


Figure 2. Perceptual thresholds set in Expt 1. *(A) Frequency histogram showing the number of Expt 1a participants who set their perceptual threshold at each volume setting/dB SPL level, as established in the anechoic calibration procedure. The top row of the x-axis shows estimated dB SPL level; the bottom row of the x-axis shows the range of the corresponding MacBook Pro volume settings. (B) Frequency histogram showing the number of Expt 1b participants who set their perceptual threshold at each dB SPL. The top row of the x-axis shows estimated dB SPL; the bottom row of the x-axis shows the range of the corresponding MacBook Pro volume settings. (C) Scatterplot showing Expt 1c data from the same participants, collected indoors in the anechoic chamber (x-axis), and outdoors in a Pittsburgh park (y-axis). The black line shows best linear fit; individual data points are slightly jittered to show all 20 individuals. (D) Frequency histogram showing the number of Expt 1c participants who set their perceptual threshold at each dB SPL, for indoor (anechoic chamber) and outdoor (park) settings.*

Experiment 1a

In the sample of people on the Carnegie Mellon University campus, participants set their audibility thresholds by choosing volume settings that were between 19 - 50%, a range that corresponds to 14.1 - 29.4 dB SPL, with a mean dB SPL setting of 22.95 (standard deviation (SD) 4.65, Figure 2a).

Experiment 1b

In the sample of people recruited in the central Pittsburgh park, participants' white noise perceptual thresholds were somewhat broader than in Expt 1a. Volume settings were between 19 and 76%, a range corresponding to 14.1 - 38.9 dB SPL, and a mean dB SPL of 26.82 (SD 5.86, Figure 2b).

Experiment 1c

As with the previous experiments, participants' white noise detection thresholds were converted from the MacBook Pro percent volume setting to dB SPL using the data and extrapolation shown in Figure 1. Results in both settings replicated the previous experiments, with participants' indoor volume settings ranging between 19 - 50% (14.1 - 29.4 dB SPL, mean 19.59 dB SPL, SD 4.56), and outdoor settings ranging between 25 - 63% (18-34.4 dB SPL, mean 25.1 dB SPL, SD 4.54).

Figure 2c shows that participants' noise detection thresholds in anechoic and outdoor conditions were highly correlated (Pearson $r = 0.79$, $p < 0.001$, verified using nonparametric Spearman rho = 0.68, $p = 0.0011$). There was a quite modest average increase of 5.5 dB SPL threshold values from anechoic to outdoor settings (Figure 2d). This mean increase in threshold is considerably lower than would be expected if the difference between ambient dB SPL in the anechoic chamber and outdoor park environments were used as a guide (28 dB SPL indoors versus 57 dB SPL outdoors).

Because participant age can interact with both pure-tone hearing thresholds as well as listening in noise, we assessed the potential effects of age on estimated thresholds in outdoor settings by combining data from Expts 1b and 1c (Figure 3). Using a regression analysis including age in years as well as cohort (participants in Expt 1b or Expt 1c), the overall model was significant (ANOVA, $F(2,45) = 4.81$, $p < 0.0128$), with no significant effect of cohort ($t = 1.49$, $p > 0.14$), and a significant and moderate effect of age ($t = 2.86$, $p < 0.0063$, slope estimate 0.215, omega-squared 0.13). There were two people who had quite high thresholds (38.9 dB SPL); one participant (age 40) mentioned s/he occasionally wore hearing aids.

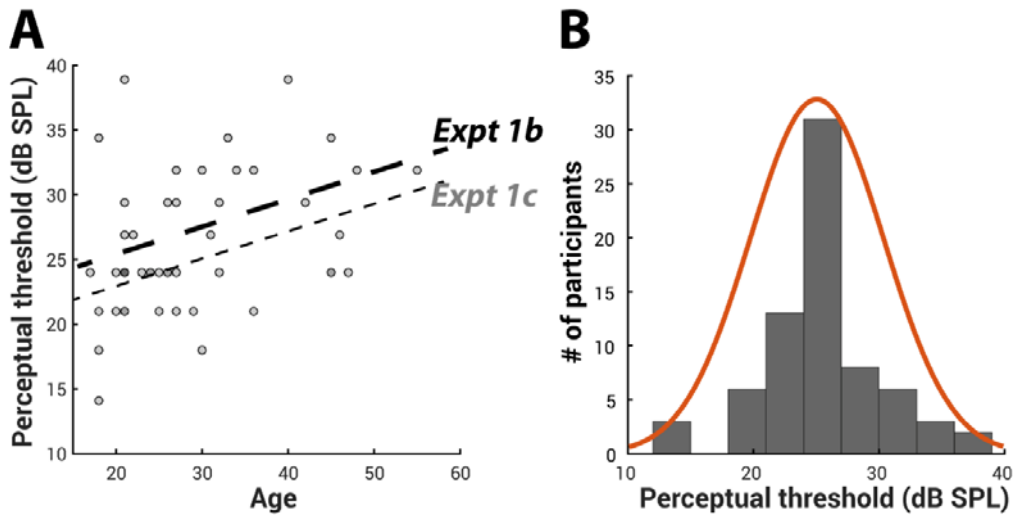


Figure 3. Perceptual thresholds in Expt 1. *(A) Scatterplot showing the relationship between Expt 1b and 1c participant age (on the x-axis) and estimated dB SPL threshold on the y-axis. The grey circles present the individual data pooled from both Expts 1b and 1c (N = 48 in total). Some circles are darker indicating overlapping data points. The black and grey lines show the best fit between age and dB SPL threshold when cohort (participants in Expt 1b or 1c) is included in the regression model. (B) Histogram of perceptual thresholds set by participants who were tested outdoors in all Expts 1a-1c (n=72) is shown, with the black bins indicating the number of participants who set their perceptual threshold at each dB SPL and the red solid curve indicating a normal distribution fit.*

Across Expts 1a-1c (N=72 total participants tested outdoors, Figure 3b), the median noise detection threshold was 24 dB SPL, with the 10th and 90th percentiles at 18 and 31.9 dB SPL. For very quiet indoor settings, extrapolating from the Expt 1c outdoor versus indoor within-subjects experiment showing a 5.5 dB level difference, we would expect a median detection threshold of 18.5 dB SPL with 10th and 90th percentiles of 12.5 and 26.4 dB SPL. The 24.8 dB (14.1 to 38.9 dB SPL) range of sound detection thresholds is consonant with that of normal hearing adults in clinical test settings (range of 30 dB, between -10 to 20 dB SL, Moore, 2013).

For experimenters who need to present auditory stimuli within a given range of intensities or at a particular level above perceptual threshold, the presentation level can be referenced to the RMS level of the white noise stimulus used in the volume setting procedure. For instance, if it is expected that stimulus presentation will be in somewhat noisy households at ~ 60dB SPL, the average stimulus dB RMS (as measured in Praat) should be *((60 - [median threshold in dB SPL = 24]) + [dB RMS of white noise stimulus in thresholding procedure = 26]) = 62*. Assuming our Exp 1 noise detection threshold results generalize to the online population, the 10th and 90th percentiles of presented levels across all participants should be approximately 54- and 68-dB SPL. Alternatively, the stimulus could simply be set 30 dB RMS above each individual participant's white noise threshold level *(e.g., 26 + 30 = 56 dB RMS)* to ensure that stimuli are sufficiently audible to the vast majority of participants.

In sum, Expt 1 establishes the feasibility of having participants act as their own sound level meter under worst-case listening conditions in public outdoor spaces. Although the approach is quite a departure from the high level of control typical of laboratory studies, it presents a practical alternative for online auditory psychophysical paradigms in which stimulus amplitude must fall within a constrained range of audibility.

Experiment 2

Experiment 2 makes use of the noise detection threshold setting procedure, validated in Expt 1, to set stimulus levels for a classic psychophysical task -- tone detection in noise -- among online participants. In Expt 2a, we first ask if reliable, well-behaved psychophysical threshold tracks can be obtained online. Second, we examine whether small adjustments to traditional thresholding procedures might permit fast (1-3 minutes) and reliable threshold estimates online. Given the risk of reduced participant vigilance and attentiveness during online studies, reducing the amount of time devoted to establishing a psychophysical threshold is particularly important. Thus, the first goal of Expt 2a is to investigate the minimum number of trials needed to derive a reliable threshold estimate. Modern online testing platforms also make the study of human psychophysics available to a wide cross-section of would-be researchers, including students and other non-experts. In this light, another goal of Expt 2a is to determine whether the standard method of estimating a threshold -- the mean of a number of reversals -- can be simplified. In particular, we examine whether a simple estimate -- the mode across all levels encountered in the staircase procedure -- is as robust or more robust at estimating threshold as traditional estimators based on staircase reversals.

Expt 2b applies the thresholding procedure from Expt 2a to measure frequency-selective auditory attention using the classic probe signal paradigm(Borra, Versnel, Kemner, van Opstal, & van Ee, 2013; Dai et al., 1991; T. J. Green & McKeown, 2001; Greenberg & Larkin, 1968; Macmillan & Schwartz, 1975; Moore et al., 1996; Scharf et al., 1987). We ask 1) whether this somewhat finicky paradigm can be replicated online in uncontrolled testing; 2) if frequency-selective attention effects can be observed on an individual basis within a single short online testing session (circa 30 minutes); and 3) if these effects change across the course of a testing session. Given widespread concerns about online participants' motivation and vigilance, we also examine whether psychophysical thresholds and frequency-selective attention are related to well-established measures of fatigue, apathy, and task confidence before, during, or after testing.

Since the same participants completed Expt 2a and Expt 2b in the same online session, we report the methods together. Results are separated by experiment to highlight the outcomes for thresholding (Expt 2a) versus frequency-selective auditory attention (Expt 2b).

Methods

Participants

We recruited 60 online participants via the Prolific recruitment platform (prolific.co, Damer and Bradley, 2014; see Table 1 for demographics) who gave electronic informed consent prior to the experiment. Data collection occurred between 11th and 14th May 2021 with participants paid to complete the study. The study was approved by the Ethics Committee in the Department of Psychological Sciences at Birkbeck, University of London and carried out in accordance with the relevant guidelines and regulations.

Participants were selected from a large pool of individuals from across the world. We utilized Prolific.co pre-screening options to refine eligible participants to those who were between 18 and 40 years of age, reported no hearing difficulties, and had a 100% Prolific.co approval rate.

TABLE 1 ABOUT HERE

Stimuli and Procedure

The experiment was implemented using PsychoPy v2021.1.2 and hosted on PsychoPy's online service, Pavlovia (pavlovia.org). A demo is available at [https://run.pavlovia.org/sijiazhao/threshold_demo]. Participants were required to use the Chrome internet browser on a laptop or desktop computer (no smartphone or tablet) to minimize the variance in latency caused by differences among browsers and devices. Operating system was not restricted. Before the start of the online experiment, participants were explicitly reminded to turn off computer notifications.

Volume Setting

Participants first followed the volume setting procedure described for Expt 1. As described above, this brief (<2 min including form-filling) procedure had participants set the volume on their computer to just-detectable levels, thereby serving as their own sound level meter.

Headphone Check

After that, we screened for compliance in wearing headphones using the dichotic Huggins Pitch approach described by Milne et al. (2020). Here, a faint pitch can be detected in noise only when stimuli are presented dichotically, thus assuring that accurate detection of the pitch only if headphones are worn. The code was implemented in JavaScript and integrated with Pavlovia using the web tool developed by author SZ [https://run.pavlovia.org/sijiazhao/headphones-check/].

The headphone check involved 6 trials, each with three one-second-long white noise intervals. Two of the intervals presented identical white noise delivered to each ear. The third

interval, random in its temporal position, was a Huggins Pitch stimulus (Cramer & Huggins, 1958) for which white noise was presented to the left ear and the same white noise, phase shifted 180° over a narrow frequency band centered at 600 Hz (±6%), was presented to the right ear to create a Huggins Pitch percept (Chait, Poeppel, & Simon, 2006; Yost & Watson, 1987).

Participants were instructed that they would hear three noises separated by silent gaps and that their task was to decide which noise contained a faint tone. Perfect accuracy across six trials was required to begin the main experiment. Participants were given two attempts to pass the headphone check before the experiment was terminated. The procedure took approximately 3 minutes to complete.

Adaptive Staircase Thresholding Procedure

Two simple acoustic signals comprised the stimuli for the adaptive thresholding procedure. A 250-ms, 1000-Hz pure tone with 10-ms cosine amplitude onset/offset ramps was generated at a sampling rate of 44.1kHz (16-bit precision) in the FLAC format using the Sound eXchange (SoX, http://sox.sourceforge.net/) sound processing software. This tone served as the target for detection in the thresholding procedure.

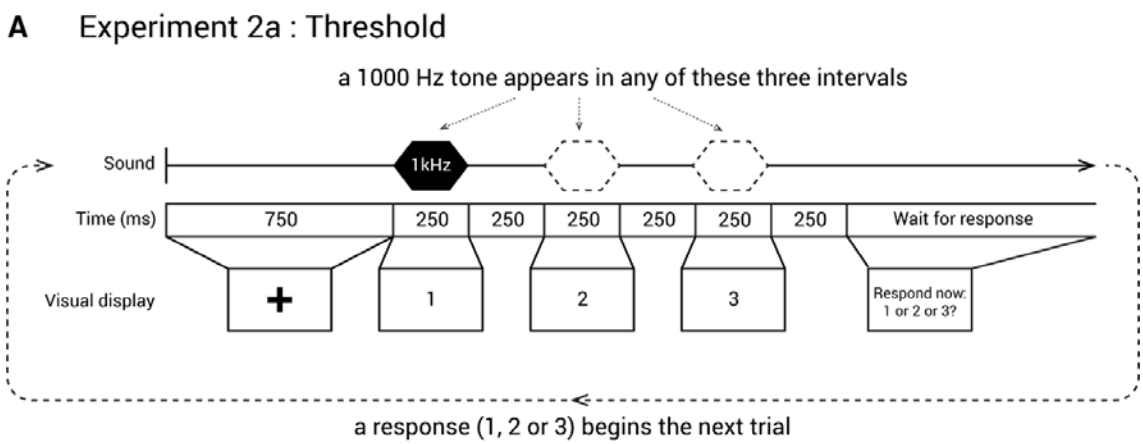
A long 300-sec white noise with 200-ms cosine on/off ramps served as a masker; this was generated using the same procedure as described for Expt 1, except that it was adjusted in intensity using Praat to 0.0402 RMS (nominal 66.06 dB in Praat intensity setting) rather than 0.000399 RMS (Praat 26 dB intensity) as in the volume setting experiment (Expt 1). The white noise masker was thus 40 dB suprathreshold. To estimate the sound pressure level of the masker as delivered to Expt 2 participants, we averaged Expt 1c's indoor and outdoor extrapolated dB SPL (mean 22 dB SPL, SD 4.3) and added 40 dB, arriving at an estimate of 66 (+/- 4.3) dB SPL average masker intensity. This is similar to many probe signal experiments, including the original Greenberg and Larkin (1968) studies (65 dB SPL), as well as a recent replication and extension (65 dB SPL, Anandan et al., 2021).

The noise masker was continuous, with onset commencing as soon as participants began the threshold procedure and looping through the end of the experiment. At the end of each five-minute loop, there was a slight 'hiccup' as the noise file reloaded which occurred at different times for each participant, as a number of the experimental parts were self-paced. Simultaneous presentation of a long masking sound - or indeed any long continuous sound - is challenging for experimental drivers, particularly online. However, transient noise onsets and offsets - for instance, starting and stopping the noise mask for each trial - can have surprisingly large effects on perception (e.g., Dai et al., 1991; Franosch et al., 2003). Indeed, our own pilot work suggested that a transient rather than continuous noise masker can cancel out the probe signal effect.

The staircase threshold procedure followed the headphone check, with the continuous white noise background sound commencing in advance of the first trial and continuing throughout the experiment. The threshold procedure trial design is shown in Figure 4a. Each trial was a three-interval forced choice: the 1000-Hz signal tone could appear during any one of the three 250-ms intervals (250-ms ISI) with equal probability. The intervals were labelled with the digits “1”, “2” and “3” displayed visually at the center of a screen and participants

responded using their computer keyboard by pressing the number corresponding to the interval in which they heard the signal. All symbols and instructions were presented as black text on a white background.

The intensity level of the signal relative to noise that was required to produce 79.4% correct detection was determined using an adaptive “three-down, one-up” staircase procedure (Levitt, 1971). Each track started at a signal-to-noise ratio (SNR) of -13.8 dB, and began with an initial descent to approximate threshold, with every correct response leading to a decrease in signal intensity by 1.5 dB with the decrement reducing to 0.75 dB once the level fell below 22.75 dB SNR or after the first incorrect response. At this juncture, the three-down, one-up staircase procedure started.



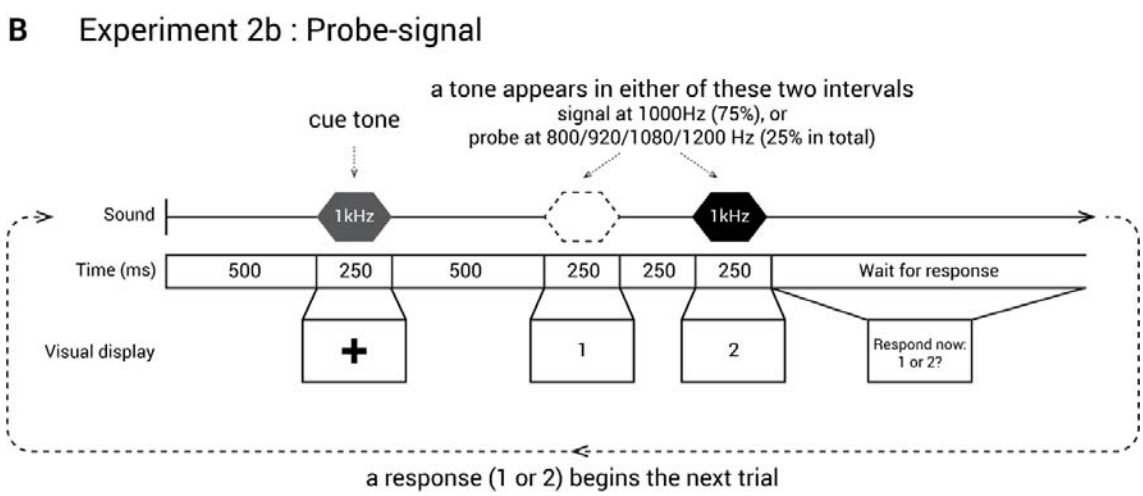


Figure 4. Trial structure in the threshold staircase procedure (A) and the probe-signal task (B). *(A) In Expt 2a, only one of the three intervals (1, 2, or 3) contained the signal, a 250 ms, 1 kHz pure tone. Responses were collected by participants pressing the corresponding numerical key on their computer keyboards. (B) In Expt 2b, the trial always started with a 1 kHz cue tone. One of the two intervals (1 or 2) contained a 250-ms tone, which could either be a frequently presented signal (1000 Hz, at 75%) or a probe (800, 920, 1080, or 1200Hz, each 6.25%). Figure is drawn in the style of Anandan et al., 2021 for continuity and clarity.*

As practice before the first of three adaptive threshold staircase tracks, participants completed six trials with the signal presented at -13.8 dB SNR (i.e. the easiest level) and with performance feedback provided ("correct" or "wrong" shown for 1 sec on-screen after each response). The average performance of this practice block was 92.78% correct (SD = 13.85%) with 41 out of 60 participants (68%) making no mistakes.

Each of the subsequent three adaptive staircase threshold tracks consisted of 40 trials. Tracks were completed consecutively, with the opportunity for a short break between tracks. However, most participants did not take a break (mean break duration = 9.03 s, SD = 11.68 s).

To keep participants engaged throughout the procedure, progress was shown on the top left of the screen ("Progress: x/40", where *x* is the index of the current trial). Moreover, we awarded a bonus (maximum of £1.50) in addition to the base payment; after the 6 practice trials with feedback, participants were informed that if their accuracy surpassed 50%, they could earn a bonus of 50p per track. The accumulated bonus was shown at the end of each track, and all 60 participants got the full bonus of £1.50.

The threshold staircase procedure was achieved using in-house code [https://gitlab.pavlovia.org/sijiazhao/threshold_demo].

Probe Signal Paradigm

After the threshold was determined via the staircase procedure, participants completed a classic probe-signal task (Anandan et al., 2021; Botte, 1995; Dai et al., 1991; Greenberg & Larkin, 1968; Scharf et al., 1987; Tan, Robertson, & Hammond, 2008). Continuous broadband noise was present throughout all trials, as described for the thresholding procedure. As shown in Figure 4b, each trial began with a 1000-Hz, 250-ms cue tone followed by 500 ms of silence. At this point, the first of two listening intervals was indicated by a black "1" presented at central fixation on the white computer screen for 250 ms. The "1" disappeared during a 250-ms silent interval at which time a black "2" was presented at fixation to indicate a second listening interval.

A 250-ms tone was presented with equal probability in either the first or the second listening interval and participants reported which interval contained the tone with a keypress. *Signal* trials involved a tone that matched the 1000-Hz cue frequency; these trials comprised 75% of the total trials. Another four *probe* tones with 800, 920, 1080, and 1200 Hz frequencies

were presented with equal probability across the remaining 25% of trials (6.25% likelihood for each tone frequency).

In previous pilot experiments, we found that some participants' accuracy on the probe signal task was slightly higher than expected from their thresholds. To assure that the full sample of participants was not performing at ceiling, we adjusted each individual's probe-signal SNR threshold slightly, lowering it by one step size (0.75 dB SNR) from the threshold estimated in Expt 2a. The signal and probe tones were always presented at the adjusted threshold level; the preceding cue tone was suprathreshold, set at 14 dB SNR above the adjusted threshold SNR level.

Participants first completed five practice trials with suprathreshold signal and probe tones presented at -13.8 dB SNR. Immediately thereafter, another five practice trials involved signal and probe tones at the adjusted individual threshold. Performance feedback (“correct” or “wrong”) was provided on-screen for one second following each response to a practice trial.

Each of the subsequent 12 blocks consisted of 32 trials (384 trials total), with 24 signal trials (1000-Hz tone) and 2 probe trials at each of the other frequencies (8 probe trials total) in random order. Blocks were completed consecutively, with the opportunity for a short break between blocks (mean break duration = 10.44 s, SD = 22.44 s). There was no feedback for these trials.

Participants were informed that if their overall accuracy across the 12 blocks surpassed 65%, they would earn a bonus of £1.00 at the end of the experiment. In all, 63% of participants earned the bonus.

On average, participants spent 39.3 minutes (SD = 10.4) on the entire experiment, including both the Adaptive Staircase Threshold procedure (Expt 2a) and the Probe Signal procedure (Expt 2b).

Assessment of participant apathy, motivation, and fatigue

Individual differences in threshold levels might be influenced by participants' arousal, engagement, or fatigue (Bianco et al., 2021; Libera & Chelazzi, 2006; Shen & Chun, 2011). Therefore, we presented the Apathy Motivation Index (AMI) questionnaire before the experiment to measure lack of motivation (apathy). This 18-question survey is subdivided into three apathy subscales: emotional, behavioral and social apathy (Ang et al., 2017; see Supplemental Information for questions).

To track the dynamics of motivation and fatigue across the experiment, participants also rated their level of subjective motivation, fatigue and confidence before and after the threshold session. They were provided with three horizontal visual analogue scales, each with equally spaced tick marks along its axis, an accompanying question positioned centrally above, and labels at the extreme left and right of the scale. The questions and labels are available in Supplemental Information.

Responses were registered by a click on the appropriate position on each scale. After completing all three ratings, a “confirm” button appeared at the bottom of the screen, allowing participants to submit their ratings.

The questionnaire and ratings were added to the experiment on the second day of data collection, so of the 60 participants, 49 responded to both the AMI questionnaire and ratings of motivation and fatigue, 10 had the AMI questionnaire only, and one did not have the questionnaire or the ratings.

Results and Discussion

Online psychophysical staircase procedures yield reliable signal-to-noise thresholds for individual participants (Expt 2a)

First, we asked whether we could obtain good-quality and stable tone-in-noise thresholds online. As an initial qualitative approach, we examined the three 40-trial tracks for each participant, and found that they were generally well-behaved in terms of reaching a stable plateau with multiple reversals after the initial descent to the first error (All threshold tracks are available at https://github.com/sijiazhao/TPS_data). To estimate threshold distribution and reliability across tracks, we calculated the mean and range of thresholds for each participant, based on the last six reversals for each of the three tracks. The mean SNR threshold was -19.54 (SD = 1.39), with the distribution of mean thresholds slightly skewed toward lower SNR levels (see Figure 5a); the mean range of estimated thresholds across the three tracks was 1.71 dB (see Figure 5b); with a 10th and 90th percentile range of 0.39 - 3.41 dB SNR. In our experience this range of thresholds across three tracks was similar to that with university-recruited students in well-controlled laboratory settings.

Mode-derived thresholds are a robust alternative to reversal counting (Expt 2a)

We compared three different methods of deriving a threshold from psychophysical data collected in the 3-down/1-up adaptive staircase procedure. The goals were: 1) to determine whether reliable threshold estimates could be obtained using fewer trials; 2) to examine whether the statistical mode could be a viable alternative to the standard approach based on the mean of a number of reversals.

Here, our ‘gold standard' for thresholding is to average values at the last six reversals in each of three tracks, and to compute a grand mean threshold for each participant from these three-track means (orange bar in Figure 5c). We then calculated the statistical *mode* for all 40 trials in each of the three tracks per participant, and a grand mean was generated from these three modal values for each participant (purple bar in Figure 5c). The rationale for using the mode is that it can be thought of as a measure of the ‘dwell time’ -- how long a participant spends at a particular level in the adaptive staircase procedure. Finally, we computed the mode from the first 20 trials in each participant's first track; the 2nd and 3rd

tracks were not used so as to simulate the results from a single 20-trial thresholding experiment (green bar in Figure 5c).

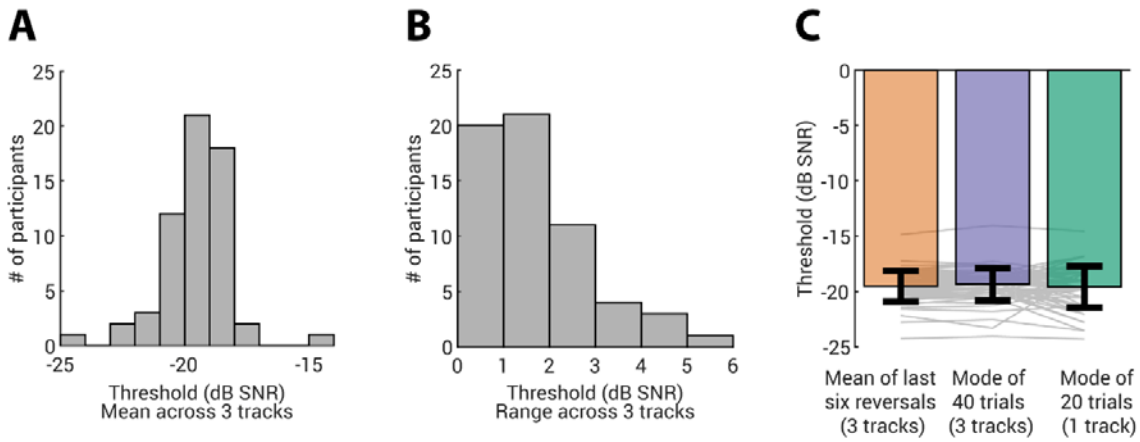


Figure 5. Results from Expt 2a. *(A) Frequency histogram of all participants' tone-in-noise thresholds in dB SNR based on the mean of the SNR values of the last six staircase reversals (count on y-axis). (B) Frequency histogram showing the distribution of the range of thresholds across the three thresholding tracks (in dB SNR). (C) The tone-in-noise detection thresholds averaged across all 60 participants as estimated using one of the three estimation methods. Error bars are one standard deviation from the mean. The grey lines indicate the individual data.*

We compared these three metrics using a Bayesian repeated measures ANOVA in JASP (JASP Team, 2020; Morey and Rouder, 2015; Rouder et al., 2012), which revealed a very low Bayes factor compared to the null hypothesis (BF_{10} = 0.132), as would be expected given the ≤ 0.2 dB SNR mean difference between any of the three metrics. This suggests that there is little if any significant bias in using either modal measure versus the reversal-based gold standard.

However, a potentially more consequential difference between obtaining a single 20-trial threshold track estimate and using the 'gold-standard' three-track 40-trial 6-reversal-based estimate would be unacceptably high variability in the former case. To quantify the degree of variability associated with the number of trials used to calculated the threshold, we compared the distributions of differences between the gold standard 3-track grand average and single-track thresholds calculated using the mode of 1) the first 20 trials; 2) or 30 trials; 3) all 40 trials; or 4) the mean of last six reversals. Each participant contributed 3 difference scores - one from each track - to each distribution. As Figure 6 shows, and as one would expect, dispersion decreases as more trials are used to calculate the threshold - but the magnitude of the difference is relatively small.

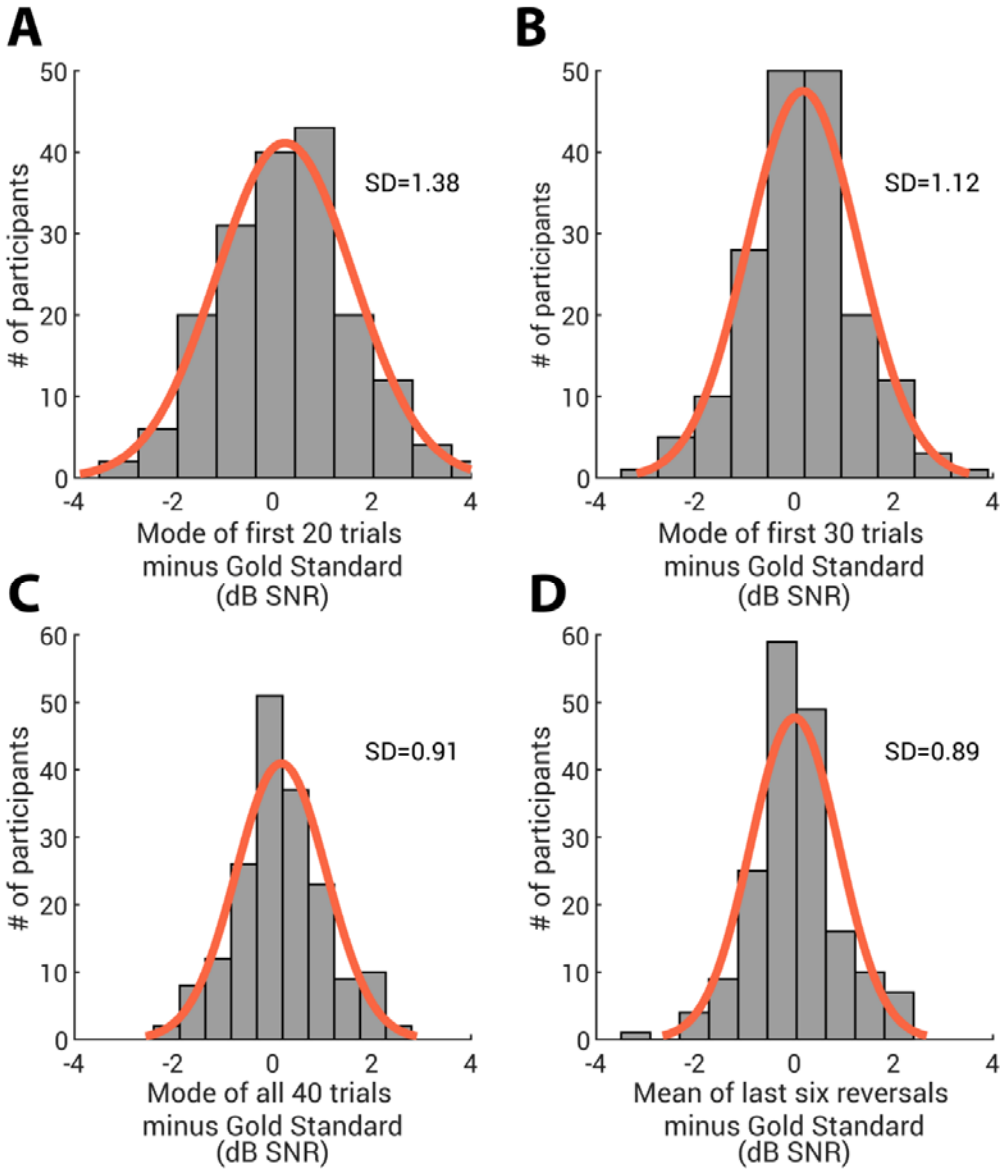


Figure 6. Difference in dB SNR of each participant's single tone-in-noise threshold tracks derived from the mode of the first 20, 30 and all 40 trials along with the mean of the last six reversals when compared to the 'gold standard' mean of three reversal-based thresholds. *Note that each participant contributes three datapoints (one from each track) to each distribution.*

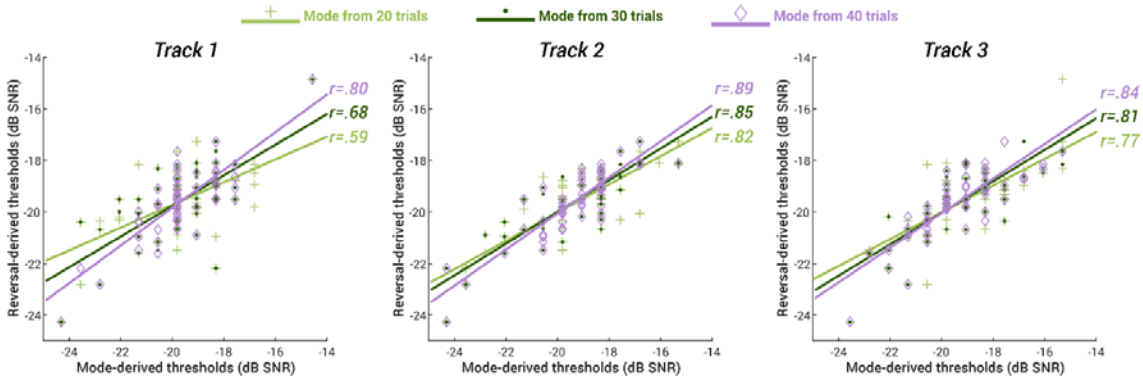


Figure 7. Correlations between the 3-track gold standard threshold (y-axis) and single-track threshold estimates (x-axis) based on the mode of the first 20, 30, or all 40 trials (track 1 (left), 2 (middle), and 3 (right panel)). *Light green crosses and lines refer to 20-trial mode estimates, dark green to the 30-trial estimates, and purple to the 40-trial estimates. Both axes show tone-to-noise dB SNR. Pearson's correlation coefficients for each estimate are shown on the right of the fits.*

We also assessed the adequacy of single-track mode-based threshold estimates using the initial 20, 30, or all 40 trials. To do so, we examined the relationship of each with the gold standard 3-track threshold across participants, and then statistically compared the difference in correlations. As tested using the r package *cocor* using the Hittner et al. and Zou tests (Diedenhofen & Musch, 2015; Hittner, May, & Silver, 2003; Zou, 2007), the fit between the gold standard and mode-based thresholds differed across tracks[1] (Figure 7, *values* shown in figure). Here, the correlations between each mode-derived threshold from first thresholding track and the gold standard threshold were all significantly ($p < 0.05$) lower compared to when the same measure used data from the second thresholding track; correlation differences between the first and third tracks were in the same direction, but all 'marginal' using the Hittner et al. tests ($p < 0.08$). The less-robust thresholds obtained in the first track suggest that at least some psychophysics-naive online participants had not quite acclimated to the thresholding procedure until later on in the track - a point we return to in the discussion.

Using the same *cocor*-based comparison method (and with the same statistical caveats), we also found that the relative reliability of mode-based thresholds derived from 20 or 30 versus 40 trials changed across tracks. In the first and second tracks, thresholds based on the first 20 trials were significantly less correlated with the gold standard than were those based on 40 trials ($p < 0.05$), but did not differ in the last track; correspondingly, first-track thresholds based on the first 30 trials were significantly less correlated with the gold standard than were those based on 40 trials ($p < 0.05$), but this difference was no longer significant in the second or third tracks. A *post hoc* explanation of this effect is that online participants acclimated to the thresholding procedure across the three tracks, and performance becomes more stable and consistent after a few minutes of practice. Nevertheless, as shown

[1] Note that this is a very weak form of statistical inference. To our knowledge, a formal test for an interaction between differences in correlations within and across levels of a repeated-measures design has yet to be developed, so the reader should not infer a significant interaction from these pairwise tests. We have also not applied any correction for multiple comparisons.

previously (Figure 6) even tone-in-noise thresholds based on the first 20 trials in the first track are reasonably accurate estimates of a participant's 'true' threshold.

In all, Expt 2a demonstrates that it is possible to quickly and reliably estimate a classic auditory psychophysics threshold online. Moreover, a very simple -- and easily automatized -- estimate of the level at which participants dwell for the most trials across the adaptive staircase procedure (the mode) was highly reliable and as robust at estimating threshold as traditional estimators based on staircase reversals. Regarding the number of tracks and trials to use, we outline potential usage cases in the Discussion.

Mode-derived thresholds relate to tone-in-noise detection (Expt 2a and Expt 2b)

We next asked how effective the online tone-in-noise threshold measurement was in setting the SNR level for an independent task. We examined this by comparing the mode-derived threshold (estimated with all 40 trials per track, averaged across the 3 tracks of Expt 2a) to the tone-in-noise detection accuracy in Expt 2b. The Expt 2a adaptive staircase procedure (3-down, 1-up) was designed to set the threshold to detect a 1000-Hz tone in noise at 79.4% accuracy. Thus, since Expt 2b tone amplitude was set according to individual thresholds, the Expt 2a threshold is a benchmark against which to compare the 1000-Hz signal tone accuracy in Expt 2b among the same participants. As noted in Section 3.1.5, we lowered the actual SNR level by 0.75 dB for each individual to retain additional 'head room' for accuracy in the probe signal task (see Section 3.1.5). Based on a pilot study mapping change in tone-in-noise SNR levels (in 0.75 dB steps) to 2AFC detection accuracy, we found that each increment in SNR corresponded to a detection accuracy change of 4.2%. Thus, we compared the distribution of detection accuracy scores for signal (1000 Hz) tones in the Expt 2b probe signal task to an adjusted target value of 75.2%. As shown in Figure 7a, average signal detection accuracy was 72.45% (SD = 8.86); thus actual signal accuracy was slightly (2.75%) yet significantly lower than the predicted accuracy ($t(59) = 62.67$, $p < 0.01$, BF > 10^{50}).

The mode-derived threshold from Expt 2a was also effective at minimizing variability in behavior in the Expt 2b probe-signal paradigm. The level of the tone in Expt 2b was set based on individual participants' mode-derived threshold estimated in Expt 2a. Therefore, if the mode-derived threshold adequately estimated thresholds for the 1000-Hz tone across participants, then the accuracy of detecting the 1000-Hz signal tone in Expt 2b should not be related to Expt 2a thresholds, since they are individually-adjusted to account for threshold differences. In other words, even if two participants have very different tone-in-noise thresholds, their accuracy on the Expt 2b 2AFC probe-signal task should be more or less equivalent. Indeed, Expt 2b signal accuracy was not correlated with the Expt 2a mode-derived threshold level (Spearman rho = -0.08, $p = .544$; Pearson $r = -.01$, $p = .929$). Thus, the online, mode-derived tone-in-noise threshold measurement was effective in setting the SNR level for an independent task.

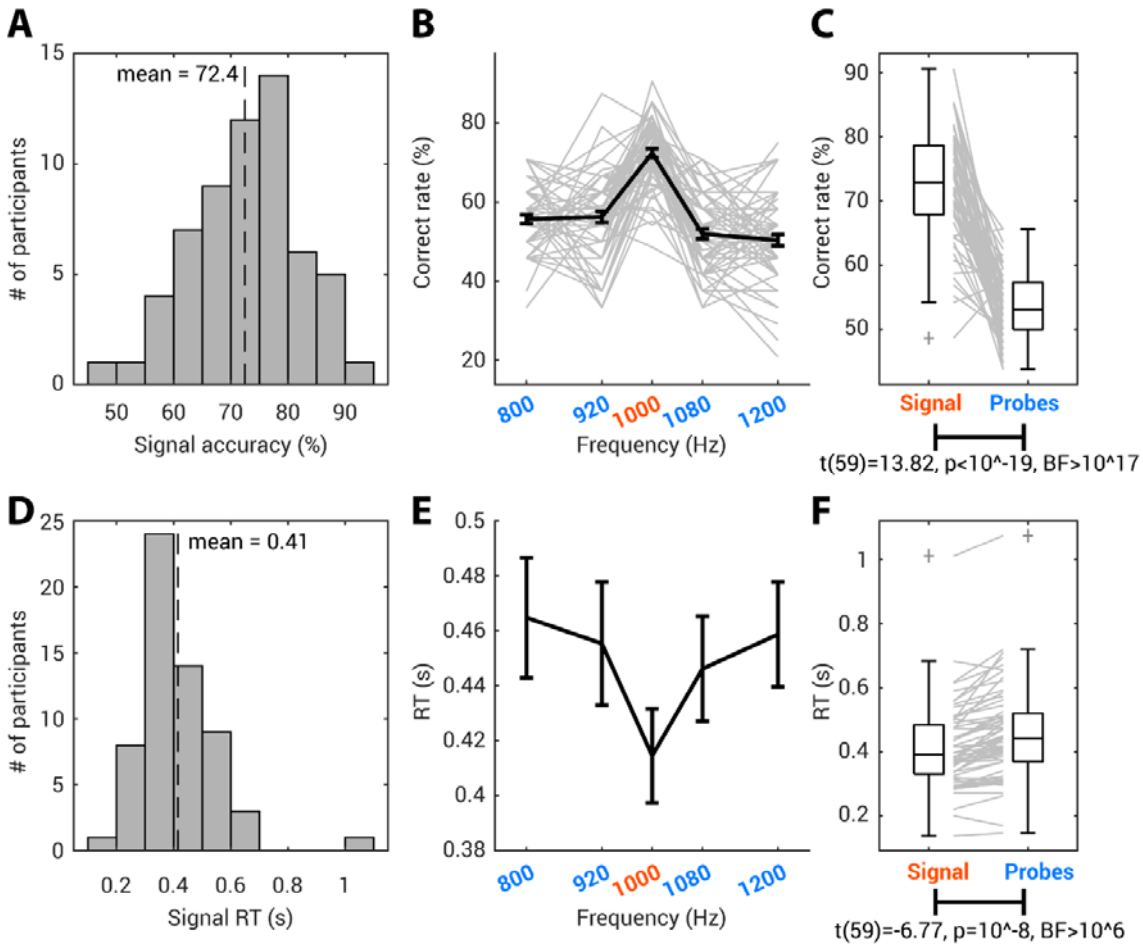


Figure 8. Experiment 2b: Probe-signal (n=60). *(A) Distribution of the signal accuracy using the mode-derived threshold. The population mean is labelled as a dash vertical line. (B) The percentage correct detection of 1000-Hz signal and each of the four probes (800, 920, 1080 and 1200 Hz). The thick black line presents the group mean, with error bar = 1 SEM. Each grey line indicates individual data. (C) The accuracy to detect signals (highly probable 1000Hz tones) was significantly higher than the average detection accuracy for probe tones (the less probable 800, 920, 1080, and 1200Hz tones) The population data is presented as a boxplot with the outliers marked as grey crosses. Each grey line indicates individual data, and paired t-test stats reported below the graph. The RT data is shown in the same manner below, D, E and F. For the visualization, the individual data are not presented in (E), but a summary of individual data is shown in (F).*

Online thresholds are reliable enough to support threshold-setting in a probe-signal paradigm (Expt 2b)

The probe-signal paradigm requires vigilance to carefully listen for a tone at low SNRs across many trials, and thus is the kind of psychophysical task that might be quite negatively impacted by a loosening of auditory hygiene in an online study. Foregoing the usual controls afforded by sound-attenuated booths, carefully calibrated equipment, and highly motivated participants would seem risky for such a task. Indeed, for these reasons, probe-signal

studies have sometimes employed expert listeners who respond to thousands (Borra et al., 2013; Dai et al., 1991; T. J. Green & McKeown, 2001; T. Green & McKeown, 2007; Greenberg, Bray, & Beasley, 1970; Howard, O'toole, Parasuraman, & Bennett, 1984; Macmillan & Schwartz, 1975; Mondor & Bregman, 1994; Moore et al., 1996; Tan et al., 2008; Wright & Dai, 1994), and even tens of thousands of trials in the laboratory (Dai et al., 1991; Greenberg et al., 1970; Greenberg & Larkin, 1968; Macmillan & Schwartz, 1975).

Yet, Expt 2b demonstrates that, with the right auditory hygiene, even a task as potentially finicky as the probe-signal paradigm can yield clean data in an online testing scenario. Here, capitalizing on the volume setting procedure of Expt 1 and the approach to threshold estimates established in Expt 2a, Expt 2b has replicated the classic probe-signal results (Figure 8).

Figure 8b illustrates that online participants detect the high-probability 1000-Hz signal at levels that are approximately at threshold (72.45% (SD = 8.86), Figure 8a and 8c), whereas tones with less-probable frequencies are less accurately detected in noise (53.59% (SD = 5.36), Figure 8c). Figure 8c plots a direct comparison of what is visually apparent in Figure 8b. The signal tone was detected significantly more accurately than were probe tones ($t(59) = 13.82$, $p < .00001$, BF > 10^{17}; Figure 8c). This classic pattern of frequency-selective auditory attention is echoed in faster reaction times for the 1000-Hz signal tone compared to the probe tones ($t(59) = 6.77$, $p < .00001$, BF>10^{6}; Figure 8e, 8f). These results replicate the frequency-selective attention effects that have been documented in laboratory studies for decades (Anandan et al., 2021; Botte, 1995; Dai et al., 1991; T. J. Green & McKeown, 2001; Greenberg & Larkin, 1968; Moore et al., 1996; Scharf et al., 1987; Tan et al., 2008) using a naive online sample of participants who utilized variable consumer equipment in uncontrolled home environments. This effect was notably robust even at the individual participant level: 56 of the 60 participants (93.33%) showed at least a 5% detection advantage for signal versus probe frequencies.

Motivation and confidence do not predict tone-in-noise thresholds from Expt 2a, but fatigue does

As exciting as the results from psychophysical experiments can be, participation can be dull for subjects. But participant engagement matters. Lapses in attentive listening in repetitive and challenging tasks like the staircase thresholding procedures described above can dramatically impact experimental results. Thus, concern that anonymous, online participants may be less motivated to perform to the best of their abilities, as compared to more traditional in-person expert listeners has contributed to reticence in moving auditory investigation online.

Here, we asked whether estimated thresholds might in part reflect the personal motivation of online participants. To this end, we used a common self-report for a personality trait-like component of motivation among healthy populations (apathy in the AMI questionnaire, Ang et al., 2017). We also examined the dynamic change of motivation ratings across our task, measured before the first threshold track and again after the third threshold track.

Participants' tone-in-noise thresholds from Expt 2a did not correlate with any aspect of the motivation trait measured by the AMI questionnaire. Neither behavioral (rho = .055, p = .68), emotional (rho = .079, p = .55), nor social apathy (rho = .053, p = .69) dimensions were related to tone-in-noise thresholds. Self-reported motivation across the course of the staircase thresholding procedure also did not account for threshold level either before (rho = -0.15, p = 0.29) or after (rho = 0.007, p = 0.96) the threshold procedure.

To assure that this lack of correlation was not due to a faulty instrument, we tested whether there was a correlation between the trait motivation/apathy score and the in-experiment motivation ratings. Indeed, the behavioral dimension of the apathy questionnaire was associated with the post-experiment motivation level (rho = -0.37, p = 0.010) and this relationship remains significant after controlling for the threshold level (partial correlation, r = -0.39, p = 0.006). This indicates that more apathetic individuals reported feeling less motivated after the threshold session regardless of their behavioral performance, although no relationship was observed prior to the experiment.

The absence of a link between motivation and task performance was further confirmed by a repeated measures general linear model on the tone-in-noise threshold level with fixed effects of the total score of the apathy questionnaire, the pre-threshold and the post-threshold motivation ratings. The thresholds could not be predicted by apathy traits ($F(1,41)$ = 0.022, p = 0.88), or motivation ratings either pre-experiment ($F(1,41)$ = 0.93, p = 0.34) or post-experiment ($F(1,41)$ = 0.15, p = 0.71). Moreover, there were no three-way or two-way interactions (all p > 0.32). In sum, online participants' motivation did not contribute significantly to their tone-in-noise thresholds.

Threshold level was also not significantly related to confidence measured either before (rho = 0.074, p = 0.61) or after the experiment (rho = -0.12, p = 0.40), suggesting that participants showed quite limited metacognitive awareness of their performance.

Finally, we investigated the relation of self-reported fatigue to thresholds. Here, the threshold level *did* positively and moderately correlate with fatigue ratings both before (rho = 0.31, p = 0.027) and after (rho = 0.32, p = 0.025) the experiment, consistent with higher (poorer) thresholds among fatigued participants.

Motivation, but not fatigue, predicts the probe-signal effect in Expt 2b

During the 12-block probe-signal task, participants were instructed to rate how well they felt they performed, how motivated they were, and how tired they felt at the end of each block. This allowed us to examine how the probe-signal effect evolves along with individuals' dynamics of confidence, fatigue and motivation.

First, we looked at how the probe-signal effect may change as participants become more practiced over time. The classic probe-signal effect is a relative accuracy advantage for detecting highly probable 1000 Hz signal tones over infrequent probe tones of other frequencies, here computed as the difference between signal accuracy and average probe accuracy for each block (Figure 9a). A linear mixed-effect model (LMM) using block index as a fixed effect and participants as a random effect showed that the probe-signal effect diminished slightly as the task progressed ($F(1,718)$ = 7.87, p = .0052). This result was

mirrored in RTs (Figure 9b); although response times to both signal and probes decreased over time, the difference between the two was overall smaller at the end of the experiment (LMM, effect of block index: F(1,711) = 10.75, *p* = .0011).

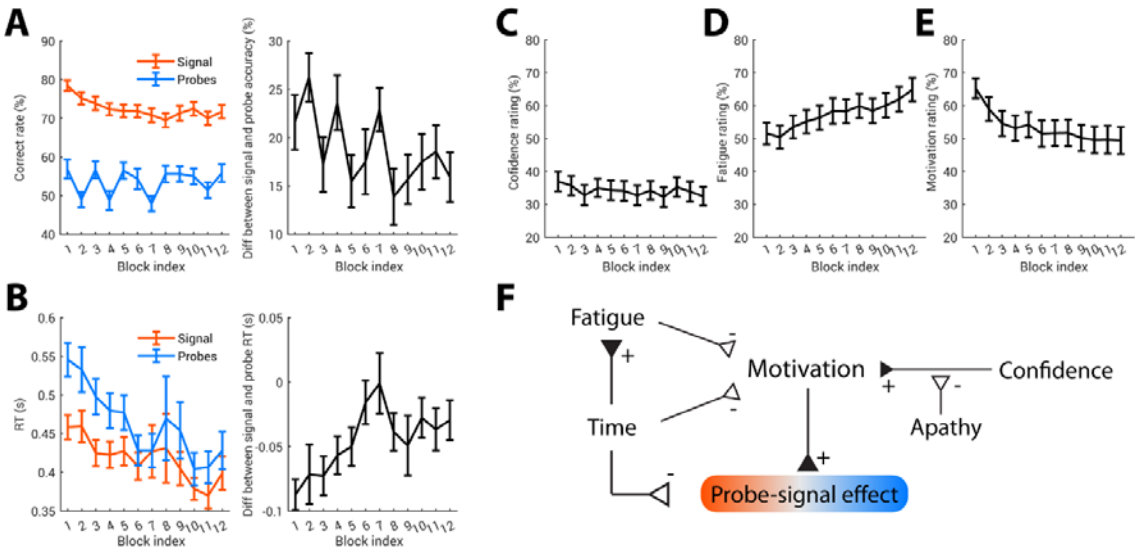


Figure 9. Dynamics of the probe-signal effect, motivation, fatigue and confidence over time. *In all plots, the error bar shows ±1 SEM. (A) Probe-signal effect in accuracy decreased over time. In the left panel, signal and probe accuracy are computed for each block and averaged across participants. The probe-signal effect is computed as Signal Accuracy - Average Probe Accuracy; the group average probe-signal effect is plotted in black in the right panel. (B) Probe-signal effect in RT is shown in the same manner; RT to signal and probe tones are plotted against the block index, with their difference is shown in the right panel. Note that since the probe-signal effect in RT is computed as RT-to-signal minus RT-to-probe, more negative values mean larger probe-signal effects. After each block, participants were instructed to rate how well they did in that block (C), how tired they felt (D), and how motivated they were (E). (F) Graphical interpretation of the findings.*

How did participants' confidence, fatigue and motivation evolve over the course of the probe-signal test? Unsurprisingly, due to the high difficulty of this task, confidence remained low throughout (Figure 9c). An LMM on confidence rating showed that as the task progressed, confidence decreased slightly but not significantly so (F(1,595) = 3.50, *p* = .062), with higher confidence associated with better overall accuracy (F(1,595) = 5.21, *p* = .023). With increasing time on task, fatigue accumulated (LMM on fatigue with block and accuracy, effect of block: F(1,595) = 47.01, $p < 10^{-10}$) and motivation diminished (LMM on motivation with block and accuracy, effect of block: F(1,595)=59.36, $p < 10^{-13}$). However, ratings of fatigue and motivation were not contingent on the probe signal performance of that block (LMM with block and accuracy, effect of overall accuracy on fatigue: F(1,595) = 0.32, p = 0.57; effect of overall accuracy on motivation: F(1,595) = 1.06, p = 0.30).

Finally, to investigate the effect of motivation and fatigue on the probe-signal accuracy effect (the relative accuracy advantage for detecting highly probable 1000Hz signal tones over

infrequent probe tones of other frequencies), we ran an LMM with block index, motivation rating and fatigue rating as fixed effects and participants as a random effect[2]. While fatigue did not show an influence on the probe-signal effect ($F(1,594) = 0.067$, $p = .80$), the probe-signal effect decreased over blocks ($F(1,594) = 5.54$, $p = .019$) and increased slightly with motivation ($F(1,594) = 5.61$, $p = .018$). This suggests that a larger probe-signal effect is predicted by high motivation, but not low fatigue. The finding is summarized in Figure 9f.

Discussion

Here, we developed and tested new approaches to making auditory psychophysical methods viable for online studies with psychophysics-naive participants. We first showed that the problem of limiting the range of stimulus sound levels can be addressed by using each participant as a 'human sound level meter' and setting stimulus levels at a given dB RMS above their noise detection threshold. We then showed that online participants' perceptual tone-in-noise thresholds could be reliably estimated, not only by combining data from multiple tracks as is classically done, but also with a single short staircase track with a simple mode-based analysis that is easily implemented even by novice researchers. Individual differences in online participants' apathy, confidence, and motivation did not significantly influence their perceptual thresholds, although those who were more fatigued tended to show somewhat less sensitive thresholds. Online tone-in-noise thresholds also were reasonably reliable in setting the desired accuracy level for a new online version of the classic probe-signal task (Dai et al., 1991; Greenberg & Larkin, 1968; Moore et al., 1996; Scharf et al., 1987). Moreover, this abbreviated version of a somewhat persnickety task - with only a third of the trials of a recent and efficient in-lab version (Anandan et al., 2021) - showed a robust frequency-selective auditory attention effect overall, and in 93% of individual participants. This compares well with results from studies with few participants each undergoing thousands of trials. Indeed, the probe signal effect itself could be clearly detected at a group level from the first block of trials (Figure 9a). The magnitude of the attentional probe-signal effect decreased somewhat as the task proceeded; this decrease was slightly ameliorated by higher motivation, and was not significantly associated with overall participant fatigue, or with changes in fatigue ratings over time. In sum, these experiments show that basic 'auditory hygiene' measures coupled with gentle adaptations of traditional methods facilitate effective and efficient online auditory psychophysics.

A method for remotely setting stimulus amplitude levels

The human auditory system is capable of successful sensing signals across a remarkable range of acoustic intensity levels, and many perceptual and cognitive phenomena are robust

[2] An LMM was used to investigate the effects of the current block's fatigue rating, the previous block's confidence rating, and task progression on motivation loss. Unsurprisingly, longer time on the task ($\beta = -0.41$, $F(1,544) = 7.78$, $p = 0.0055$) and higher fatigue ($\beta = -0.29$, $F(1,544) = 66.08$, $p < 10^{-14}$) were associated with sharper motivation loss. Confidence, on the other hand, appeared to exert a restorative effect on motivation loss ($\beta = 0.249$, $F(1,544) = 31.31$, $p < 10^{-7}$). Adding the questionnaire-derived apathy index to the LMM revealed that apathy counteracted the restorative effect of confidence ($\beta = -0.0078$, $F(1,521) = 10.22$, $p = 0.0015$). That is, in motivated individuals' high confidence more strongly prevented motivation loss over time, while in apathetic people this effect was diminished.

to level changes (Moore, 2013). However, a lack of control over auditory presentation levels - as is often the case in online experiments - is far from desirable on several grounds. Hearing safety is of course a potential concern for online experiments, particularly when presenting punctate sounds for which onset times are considerably faster than the ear's mechanical protective mechanisms can respond. Sounds presented at different absolute levels evoke responses in distinct auditory nerve fibers, which can be selectively affected by pathological processes (Schaette, 2014; Verhulst, Altoè, & Vasilkov, 2018). As noted above, the frequency selectivity of subcortical and cortical auditory neurons can vary systematically as a function of sound pressure level (Moore, 2013; Schreiner et al., 2000). Of course, absolute sound pressure level is not the only factor to consider: individual participants with normal hearing will show thresholds with a range of up to 30 dB HL, and therefore a fixed absolute amplitude level can result in quite different perceptual experiences for participants who lie at one end or the other of this hearing range.

In Expts 1a-c, we found that community-recruited participants could very quickly set computer volume levels to estimate their hearing threshold using diotic pulsed white noise. The 20-25 dB SPL range accords well with that of normal hearing (Moore, 2013), and the extrapolated threshold levels are highly consistent across the outdoor settings of the three experiments. Expt 1c showed that participants' thresholds indoors and outdoors were highly correlated; this not only shows excellent reliability (albeit in a relatively small sample given the strictures of working during the COVID pandemic), but also demonstrates the robustness of this method to different acoustic environments. Given the ~30 dB difference in average ambient sound levels, we suspect that this is due in part to participant strategy in listening in gaps, averaging in time, or in-stream segregation of the pulsed white noise versus more variable background signals. We plan a larger-N follow-up when in-person studies in indoor environments are more feasible than at the time of writing.

Time-efficient and reliable estimation of tone-in-noise thresholds

We used the 'volume-setting' method above with all of the Expt 2 participants. Based on this, the online continuous white noise masker played during both parts of Expt 2 was set to an average of 66 dB SPL (SD = 4.3), and 40 dB above each participant's white noise detection threshold (as described in Section 3.1.4). Using a standard staircase technique to estimate tone-in-noise thresholds, we were able to obtain stable threshold estimates in online participants (Figure 5a, 5b), not only by using the traditional method of averaging the means of the last six reversals from 3 staircase threshold tracks, but also using an easy-to-calculate and outlier-robust mode of the SNR levels from the first 20, 30, or all 40 trials (Figure 6). We also found that it was possible to obtain a reliable threshold from a single track of 20 trials (Figure 7), entailing about a minute of online testing.

If a psychophysical task takes about 3 sec per trial, then a standard thresholding track of 40 trials would take two minutes, and three tracks would take 6 minutes excluding time between tracks. Using the same assumption, the mode-of-20-trials approach would take about one minute to generate a threshold. This is a significant reduction in the time it takes to derive a threshold. This streamlined thresholding approach may be very attractive for online testing settings, as the vigilance of participants might not be as high as it would be during in-person

testing, where participants can typically be expected to generate reliable data for 1.5 hours or more. This fact places a premium on time-to-threshold for online studies.

Although in more traditional psychophysical testing scenarios, this reduced thresholding time would not be worth the corresponding increased variability associated with the mode-based approaches described here, online testing easily offers larger sample size from a more diverse population than traditional in-person testing on the university campus. Thus, it is suggested that the streamlined thresholding approach described here, along with shortened online testing sessions, and increased sample sizes can yield better, more reliable outcomes when testing online.

There are other issues to consider in maximizing the efficacy of online testing using streamlined thresholding. For example, psychophysical tasks that require participants to work near or at their quiet thresholds are likely not suitable, because overall ambient sound level is less controlled in online studies (as described in Expt 1), which raises signal audibility as an issue. This adds more uncertainty by the participant to the task, which makes short, 20-trial tracks less reliable. In a similar way, tasks in which the required perceptual decision is based on subtle cue differences like those that are often categorized as timbral may not be good choices for online study, again because bad tracks are more likely. Generally, it is recommended to choose psychophysical tasks that are easy for novice listeners to understand and 'hear out,' and to implement a training regimen that is carefully designed to clarify the perceptual task for listeners to avoid bad tracks, which are more difficult to discern with streamlined thresholding procedures.

Finally, another potential advantage of the mode-based approaches might lie in their ease of computation. It is undeniable that online platforms such as Pavlovia.org and Gorilla.sc make psychophysical testing accessible to many, including students and other non-experts. These budding psychophysicists may have valid and interesting scientific questions. But, they may not have algorithms at-the-ready to estimate thresholds from staircase reversals using traditional approaches, a limitation that should never be a barrier to entry into the field.

Rapid and robust online auditory psychophysics of frequency-selective auditory attention in single participants

The probe-signal paradigm (Borra et al., 2013; Dai et al., 1991; T. J. Green & McKeown, 2001; T. Green & McKeown, 2007; Greenberg et al., 1970; Greenberg & Larkin, 1968; Macmillan & Schwartz, 1975; Moore et al., 1996) would not seem to be a promising target for online research. Classic and more recent psychophysical studies have both recruited highly experienced participants for multi-day experiments with extensive tone-in-threshold measurement, multiple practice sessions, and thousands to tens-of-thousands of trials in the primary experiment, all conducted with specialized equipment in acoustically isolated laboratory settings (Borra et al., 2013; Dai et al., 1991; T. J. Green & McKeown, 2001; T. Green & McKeown, 2007; Greenberg et al., 1970; Greenberg & Larkin, 1968; Howard et al., 1984; Macmillan & Schwartz, 1975; Mondor & Bregman, 1994; Moore et al., 1996; Tan et al., 2008; Wright & Dai, 1994). Here, Expt 2b violated each of these experimental *desiderata* in a single, brief online session with psychophysically naïve participants using their own computers and headphones in uncontrolled home environments. Nonetheless, the use of our volume setting procedure and accurate online threshold estimates allowed us to observe a

probe-signal effect in most participants, with a signal-to-probe accuracy advantage of about 20-25%, on par with the magnitude of frequency-selective attention observed in studies with tens of thousands of trials (Dai et al., 1991; Greenberg et al., 1970; Greenberg & Larkin, 1968; Macmillan & Schwartz, 1975). Despite the relatively uncontrolled online experimental setting, the probe-signal effect was apparent even in response time; participants were faster in noise at detecting the signal, as compared to probe tones.

Beyond the convenience of recruiting participants online, there is power in demonstrating psychophysical effects like the probe-signal effect in a diverse sample of psychophysically naïve participants. Rather than rely on highly expert listeners, or even naïve listeners sampled from the relatively homogeneity of a university campus, Expt 2b involved a world-wide sample. Behavioral science is increasingly recognizing that human behavior sampled for convenience only across university populations may be WEIRD (Western, Educated, Industrialized, Rich and Democratic; Henrich et al., 2010), and therefore not necessarily representative of populations at large. Although there are sound reasons to expect many psychophysics paradigms to generalize beyond ivory tower samples, this assumption has not often been tested (but see McDermott et al., 2016). The present results demonstrate that, with the right approach, it is indeed feasible to successfully conduct even fussy psychophysical paradigms dependent on thresholds online, and among inexpert participants. This substantially broadens the reach of psychophysics and opens the door to the possibility of large-scale psychophysics. Here, even with modest sample sizes (that nonetheless exceed typical probe-signal samples by an order of magnitude) Expt 2b demonstrated that it is possible to observe the evolution of frequency-selective attention via the probe signal effect from the first block onward, in both accuracy and RTs.

Motivation in online participants

Another concern with online experimentation is participants' motivation; low levels may result in high drop-out rates and poor task engagement and performance, in turn affecting the validity of the experimental results (Shen & Chun, 2011). Compared with online participants, those attending in person might be expected to be more motivated since they have already made the effort to visit the lab and social evaluative stress caused by the presence of the experimenter can motivate them to some degree (Bianco et al., 2021).

Meanwhile, online experiments are normally completely anonymous and without supervision, leading to a common worry that the online population might be more apathetic than in-lab participants. Because of these concerns, we expected that the estimated thresholds might, at least in part, reflect motivation level. However, surprisingly but reassuringly, the estimated thresholds in Expt 2a showed no relation with motivation, neither as expressed by the apathy index (a personality-trait-like component of motivation derived from a well-established apathy questionnaire, Ang et al., 2017), nor the motivation ratings before and after Expt 2a. Similarly, in Expt 2b, the self-reported questionnaire-derived apathy index, as well as its subdomains, could not explain the strong probe-signal effects observed. However, we did find a weak but significant effect of in-experiment motivation on the probe-signal effect: blocks in which listeners were more motivated generated a larger probe-signal effect. Interestingly, we also found that in motivated people, high confidence strongly prevented motivation loss over time, while in apathetic people this prophylactic effect was diminished.

It is interesting that motivation showed a small effect on the result in Expt 2b but not in Expt 2a. One explanation is that the effect of motivation on performance is sensitive to the length of the experiment; Expt 2b was longer (around 20 minutes) and came after Expt 2a (a length of around 10 minutes). This, with no observed effect of motivation in Expt 2a, implies an importance to keep online experiments short. In summary, any generalizations of the motivation-related findings here should be taken carefully.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by a grant from the National Institutes of Health [R01DC017734, to LLH and FD, and R21DC018408 to CAB]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions

Sijia Zhao: Conceptualization, Methodology, Software, Validation, Data Curation, Writing
Chris Brown: Conceptualization, Methodology, Validation, Writing
Frederic Dick: Conceptualization, Methodology, Validation, Writing, Funding Acquisition
Lori L. Holt: Conceptualization, Methodology, Validation, Writing, Funding Acquisition

ORCID iDs

Sijia Zhao: https://orcid.org/0000-0002-6246-0702
Christopher A. Brown: https://orcid.org/0000-0001-7550-3014
Lori L. Holt: https://orcid.org/0000-0002-8732-4977
Frederic Dick: https://orcid.org/0000-0002-2933-3912

References

Anandan, E. S., Husain, R., & Seluakumaran, K. (2021). Auditory attentional filter in the absence of masking noise. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-020-02210-z

Ang, Y.-S., Lockwood, P., Apps, M. A. J., Muhammed, K., & Husain, M. (2017). Distinct Subtypes of Apathy Revealed by the Apathy Motivation Index. *PLOS ONE*, *12*(1), e0169938. https://doi.org/10.1371/journal.pone.0169938

Bianco, R., Mills, G., de Kerangal, M., Rosen, S., & Chait, M. (2021). Reward Enhances Online Participants' Engagement With a Demanding Auditory Task. *Trends in Hearing*, *25*, 23312165211025940. https://doi.org/10.1177/23312165211025941

Bizley, J. K., Nodal, F. R., Nelken, I., & King, A. J. (2005). Functional organization of ferret auditory cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *15*(10), 1637–1653. https://doi.org/10.1093/cercor/bhi042

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. Computer Software. Available at http://www.praat.org/ (Version 6.0.17).

Borra, T., Versnel, H., Kemner, C., van Opstal, A. J., & van Ee, R. (2013). Octave effect in auditory attention. *Proceedings of the National Academy of Sciences*, *110*(38), 15225–15230. https://doi.org/10.1073/pnas.1213756110

Botte, M. C. (1995). Auditory attentional bandwidth: Effect of level and frequency range. *The Journal of the Acoustical Society of America*, *98*(5 Pt 1), 2475–2485. https://doi.org/10.1121/1.414464

Bryan, M. E., Parbrook, H. D., & Tempest, W. (1965). A note on quiet threshold shift in the absence of noise. *Journal of Sound and Vibration*, *2*(2), 147–149. https://doi.org/10.1016/0022-460X(65)90092-1

Chait, M., Poeppel, D., & Simon, J. Z. (2006). Neural Response Correlates of Detection of Monaurally and Binaurally Created Pitches in Humans. *Cerebral Cortex*, *16*(6), 835–848. https://doi.org/10.1093/cercor/bhj027

Cramer, E. M., & Huggins, W. H. (1958). Creation of Pitch through Binaural Interaction. *The Journal of the Acoustical Society of America*, *30*(5), 413–417. https://doi.org/10.1121/1.1909628

Dai, H., Scharf, B., & Buus, S. (1991). Effective attenuation of signals in noise under focused attention. *The Journal of the Acoustical Society of America*, *89*(6), 2837–2842. https://doi.org/10.1121/1.400721

Damer, E., & Bradley, P. (2014). *Prolific academic. Computer Software. Available at https://www.prolific.ac/*.

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Franosch, J.-M. P., Kempter, R., Fastl, H., & van Hemmen, J. L. (2003). Zwicker tone illusion and noise reduction in the auditory system. *Physical Review Letters*, *90*(17), 178103. https://doi.org/10.1103/PhysRevLett.90.178103

Glasberg, B. R., & Moore, B. C. (2000). Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *The Journal of the Acoustical Society of America*, *108*(5 Pt 1), 2318–2328. https://doi.org/10.1121/1.1315291

Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *The Journal of the Acoustical Society of America*, *97*(6), 3749–3760. https://doi.org/10.1121/1.412390

Green, T. J., & McKeown, J. D. (2001). Capture of attention in selective frequency listening. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1197–1210. https://doi.org/10.1037/0096-1523.27.5.1197

Green, T., & McKeown, D. (2007). The role of auditory memory traces in attention to frequency. *Perception & Psychophysics*, *69*(6), 942–951. https://doi.org/10.3758/BF03193931

Greenberg, G. Z., Bray, N. W., & Beasley, D. S. (1970). Children's frequency-selective detection of signals in noise1. *Perception & Psychophysics*, *8*(3), 173–175. https://doi.org/10.3758/BF03210199

Greenberg, G. Z., & Larkin, W. D. (1968). Frequency-Response Characteristic of Auditory Observers Detecting Signals of a Single Frequency in Noise: The Probe-Signal Method. *The Journal of the Acoustical Society of America*, *44*(6), 1513–1523. https://doi.org/10.1121/1.1911290

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo Evaluation of Tests for Comparing Dependent Correlations. *The Journal of General Psychology*, *130*(2), 149–168. https://doi.org/10.1080/00221300309601282

Horst, J. W., McGee, J., & Walsh, E. J. (2018). Input-output curves of low and high spontaneous rate auditory nerve fibers are exponential near threshold. *Hearing Research*, *367*, 195–206. https://doi.org/10.1016/j.heares.2018.06.007

Howard, J. H., O'toole, A. J., Parasuraman, R., & Bennett, K. B. (1984). Pattern-directed attention in uncertain-frequency detection. *Perception & Psychophysics*, *35*(3), 256–264. https://doi.org/10.3758/BF03205939

Kopiez, R., & Platz, F. (2009). The Role of Listening Expertise, Attention, and Musical Style in the Perception of Clash of Keys. *Music Perception*, *26*(4), 321–334. https://doi.org/10.1525/mp.2009.26.4.321

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

Libera, C. D., & Chelazzi, L. (2006). Visual selective attention and the effects of monetary rewards. *Psychological Science*, *17*(3), 222–227. https://doi.org/10.1111/j.1467-9280.2006.01689.x

Macmillan, N. A., & Schwartz, M. (1975). Probe-signal investigation of uncertain-frequency detection. *The Journal of the Acoustical Society of America*, *58*(5), 1051–1058. https://doi.org/10.1121/1.380764

Manning, C., Jones, P. R., Dekker, T. M., & Pellicano, E. (2018). Psychophysics with children: Investigating the effects of attentional lapses on threshold estimates. *Attention, Perception & Psychophysics*, *80*(5), 1311–1324. https://doi.org/10.3758/s13414-018-1510-2

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, *advance online publication*. https://doi.org/10.1038/nature18635

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01514-0

Mok, B. A., Viswanathan, V., Borjigin, A., Singh, R., Kafi, H., & Bharadwaj, H. M. (2021). Web-based Psychoacoustics: Hearing Screening, Infrastructure, and Validation. *BioRxiv*, 2021.05.10.443520. https://doi.org/10.1101/2021.05.10.443520

Mondor, T. A., & Bregman, A. S. (1994). Allocating attention to frequency regions. *Perception & Psychophysics*, *56*(3), 268–276. https://doi.org/10.3758/BF03209761

Moore, B. (2013). An Introduction to the Psychology of Hearing: Sixth Edition. In *An Introduction to the Psychology of Hearing*. Brill. Retrieved from https://brill.com/view/title/24210

Moore, B., Hafter, E. R., & Glasberg, B. R. (1996). The probe-signal method and auditory-filter shape: Results from normal- and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, *99*(1), 542–552. https://doi.org/10.1121/1.414512

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pick, G. F. (1980). Level dependence of psychophysical frequency resolution and auditory filter shape. *The Journal of the Acoustical Society of America*, *68*(4), 1085–1095. https://doi.org/10.1121/1.384979

Rinderknecht, M. D., Ranzani, R., Popp, W. L., Lambercy, O., & Gassert, R. (2018). Algorithm for improving psychophysical threshold estimates by detecting sustained inattention in experiments using PEST. *Attention, Perception & Psychophysics*, *80*(6), 1629–1645. https://doi.org/10.3758/s13414-018-1521-z

Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences*, *108*(37), 15516–15521. https://doi.org/10.1073/pnas.1108912108

Schaette, R. (2014). Tinnitus in men, mice (as well as other rodents), and machines. *Hearing Research*, *311*, 63–71. https://doi.org/10.1016/j.heares.2013.12.004

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception & Psychophysics*, *42*(3), 215–223. https://doi.org/10.3758/BF03203073

Schreiner, C. E., Read, H. L., & Sutter, M. L. (2000). Modular Organization of Frequency Integration in Primary Auditory Cortex. *Annual Review of Neuroscience*, *23*(1), 501–529. https://doi.org/10.1146/annurev.neuro.23.1.501

Shen, Y. J., & Chun, M. M. (2011). Increases in rewards promote flexible behavior. *Attention, Perception, & Psychophysics*, *73*(3), 938–952. https://doi.org/10.3758/s13414-010-0065-7

Steed, O. T., & Martin, A. M. (1973). Studies on quiet threshold shift in the absence of noise. *Journal of Sound and Vibration*, *26*(3), 353–360. https://doi.org/10.1016/S0022-460X(73)80190-7

Taberner, A. M., & Liberman, M. C. (2005). Response Properties of Single Auditory Nerve Fibers in the Mouse. *Journal of Neurophysiology*, *93*(1), 557–569. https://doi.org/10.1152/jn.00574.2004

Tan, M. N., Robertson, D., & Hammond, G. R. (2008). Separate contributions of enhanced and suppressed sensitivity to the auditory attentional filter. *Hearing Research*, *241*(1–2), 18–25. https://doi.org/10.1016/j.heares.2008.04.003

Verhulst, S., Altoè, A., & Vasilkov, V. (2018). Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. *Hearing Research*, *360*, 55–75. https://doi.org/10.1016/j.heares.2017.12.018

Williams, K. N., Elfner, L. F., & Howse, W. R. (1978). Auditory temporal resolution: Effects of sensation level. *The Journal of Auditory Research*, *18*(4), 265–269.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Wright, B. A., & Dai, H. (1994). Detection of unexpected tones with short and long durations. *The Journal of the Acoustical Society of America*, *95*(2), 931–938. https://doi.org/10.1121/1.410010

Yost, W. A., & Watson, C. S. (Eds.). (1987). Complex spectral patterns with interaural differences: Dichotic pitch and the 'Central Spectrum'. In *Auditory Processing of Complex Sounds*. Routledge.

Zhao, S., Yum, N. W., Benjamin, L., Benhamou, E., Yoneya, M., Furukawa, S., … Chait, M. (2019). Rapid Ocular Responses Are Modulated by Bottom-up-Driven Auditory Salience. *Journal of Neuroscience*, *39*(39), 7703–7714. https://doi.org/10.1523/JNEUROSCI.0776-19.2019

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

Tables

Measure		**All (n=60)**
Age, years, mean (SD)		26.3(5.7)
Gender, female, n (%)		27(45%)
Played musical instruments for more than 1 hour per day for more than 2 years, n (%)		15(25%)
Apathy Motivation Index, mean(SD) *	Total	27.8(6.1)
	Behavioral	9.4(3.9)
	Emotional	6.7(3.5)

Table 1. Self-reported participant demographics. **One participant did not complete the Apathy Motivation Index questionnaire.*

Figure Captions

Figure 1. Sound pressure levels of the noise segment as a function of computer volume setting percentages. *Calibration was established in an anechoic chamber; the same stimulus, computer and headphones were used for the perceptual thresholding of Experiments 1a-1c.*

Figure 2. Perceptual thresholds set in Expt 1. *(A) Frequency histogram showing the number of Expt 1a participants who set their perceptual threshold at each volume setting/dB SPL level, as established in the anechoic calibration procedure. The top row of the x-axis shows estimated dB SPL level; the bottom row of the x-axis shows the range of the corresponding MacBook Pro volume settings. (B) Frequency histogram showing the number of Expt 1b participants who set their perceptual threshold at each dB SPL. The top row of the x-axis shows estimated dB SPL; the bottom row of the x-axis shows the range of the corresponding MacBook Pro volume settings. (C) Scatterplot showing Expt 1c data from the same participants, collected indoors in the anechoic chamber (x-axis), and outdoors in a Pittsburgh park (y-axis). The black line shows best linear fit; individual data points are slightly jittered to show all 20 individuals. (D) Frequency histogram showing the number of Expt 1c participants who set their perceptual threshold at each dB SPL, for indoor (anechoic chamber) and outdoor (park) settings.*

Figure 3. Perceptual thresholds in Expt 1. *(A) Scatterplot showing the relationship between Expt 1b and 1c participant age (on the x-axis) and estimated dB SPL threshold on the y-axis. The grey circles present the individual data pooled from both Expts 1b and 1c (N = 48 in total). Some circles are darker indicating overlapping data points. The black and grey lines show the best fit between age and dB SPL threshold when cohort (participants in Expt 1b or 1c) is included in the regression model. (B) Histogram of perceptual thresholds set by participants who were tested outdoors in all Expts 1a-1c (n=72) is shown, with the black bins indicating the number of participants who set their perceptual threshold at each dB SPL and the red solid curve indicating a normal distribution fit.*

Figure 4. Trial structure in the threshold staircase procedure (A) and the probe-signal task (B). *(A) In Expt 2a, only one of the three intervals (1, 2, or 3) contained the signal, a 250 ms, 1 kHz pure tone. Responses were collected by participants pressing the corresponding numerical key on their computer keyboards. (B) In Expt 2b, the trial always started with a 1 kHz cue tone. One of the two intervals (1 or 2) contained a 250-ms tone, which could either be a frequently presented signal (1000 Hz, at 75%) or a probe (800, 920, 1080, or 1200Hz, each 6.25%). Figure is drawn in the style of Anandan et al., 2021 for continuity and clarity.*

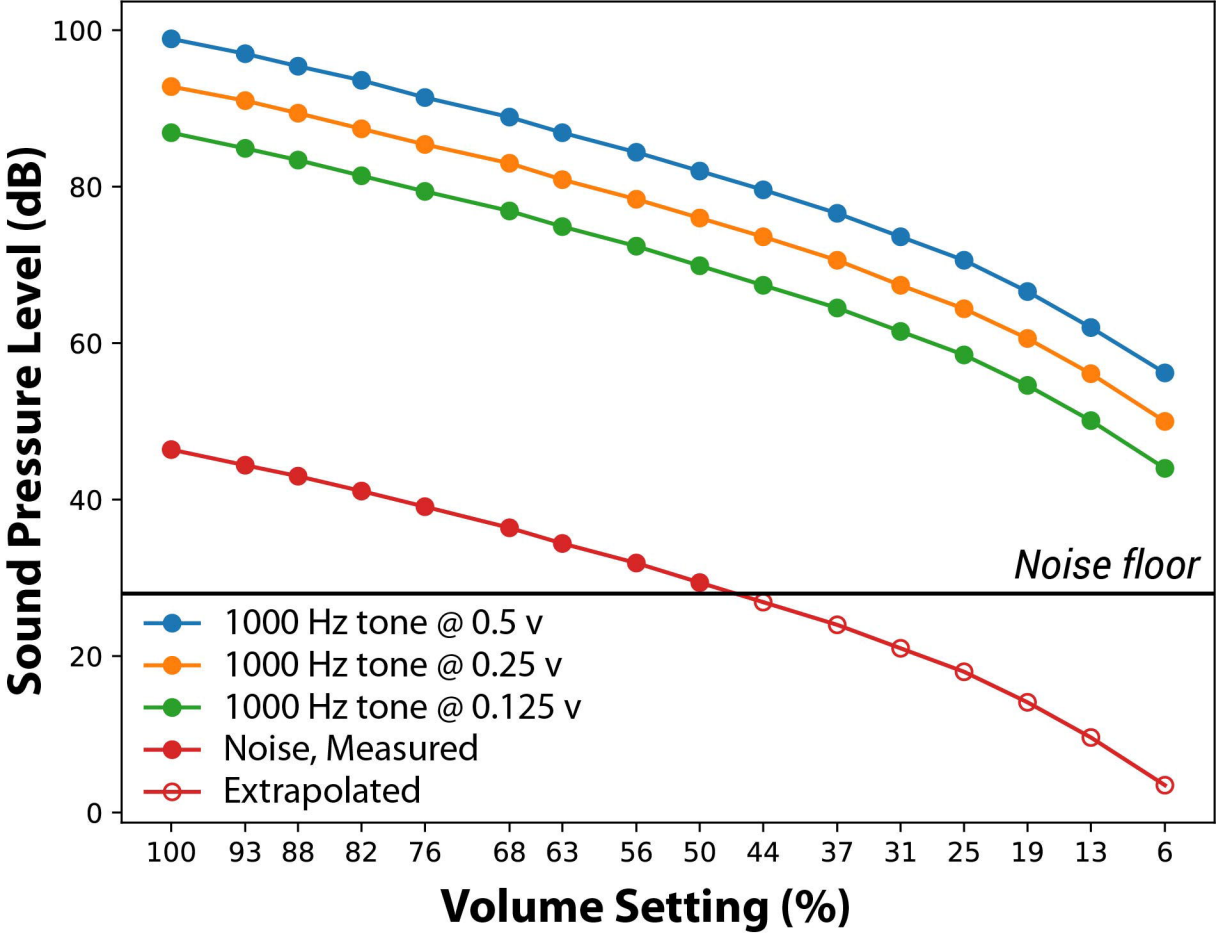
Figure 5. Results from Expt 2a. *(A) Frequency histogram of all participants' tone-in-noise thresholds in dB SNR based on the mean of the SNR values of the last six staircase reversals (count on y-axis). (B) Frequency histogram showing the distribution of the range of thresholds across the three thresholding tracks (in dB SNR). (C) The tone-in-noise detection thresholds averaged across all 60 participants as estimated using one of the three estimation methods. Error bars are one standard deviation from the mean. The grey lines indicate the individual data.*

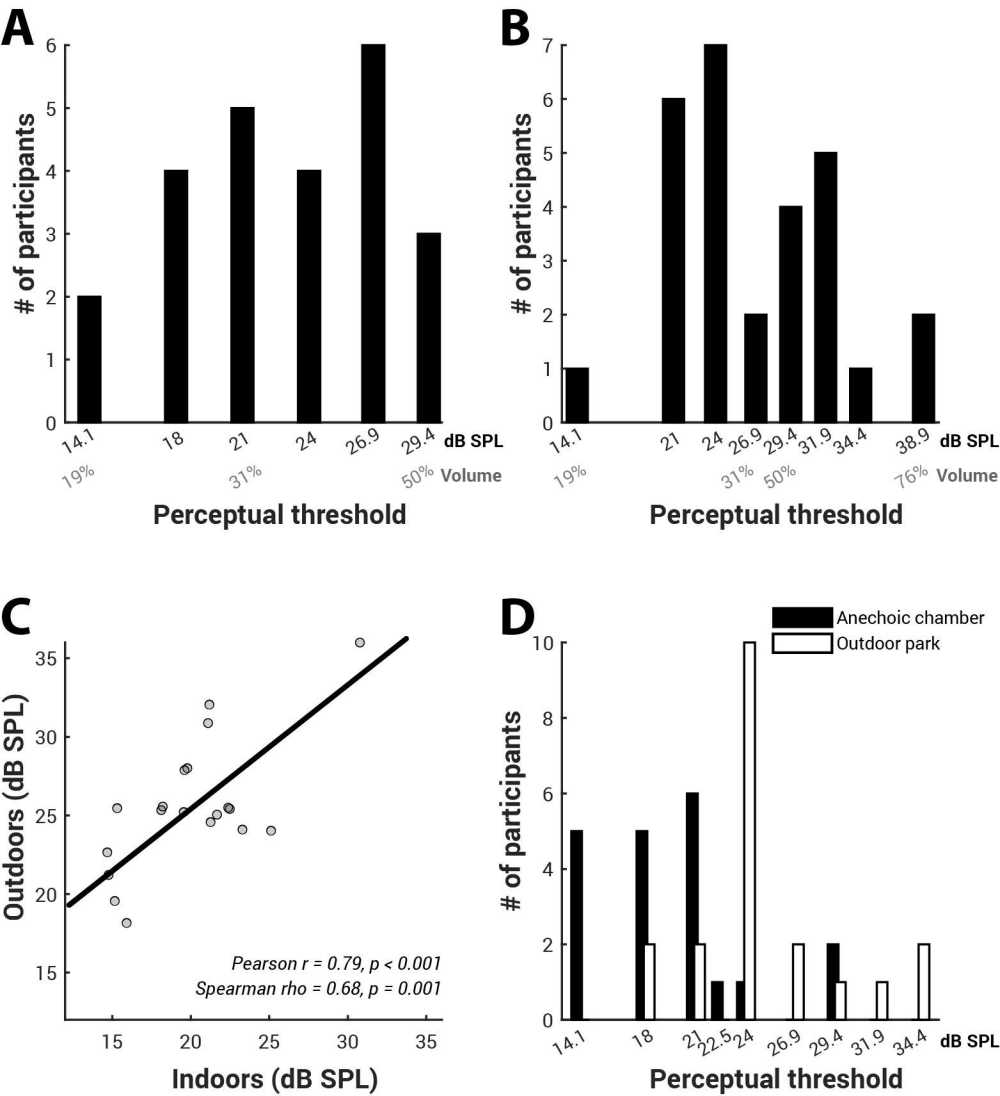
Figure 6. Difference in dB SNR of each participant's single tone-in-noise threshold tracks derived from the mode of the first 20, 30 and all 40 trials along with the mean of the last six reversals when compared to the 'gold standard' mean of three reversal-based thresholds. *Note that each participant contributes three datapoints (one from each track) to each distribution.*

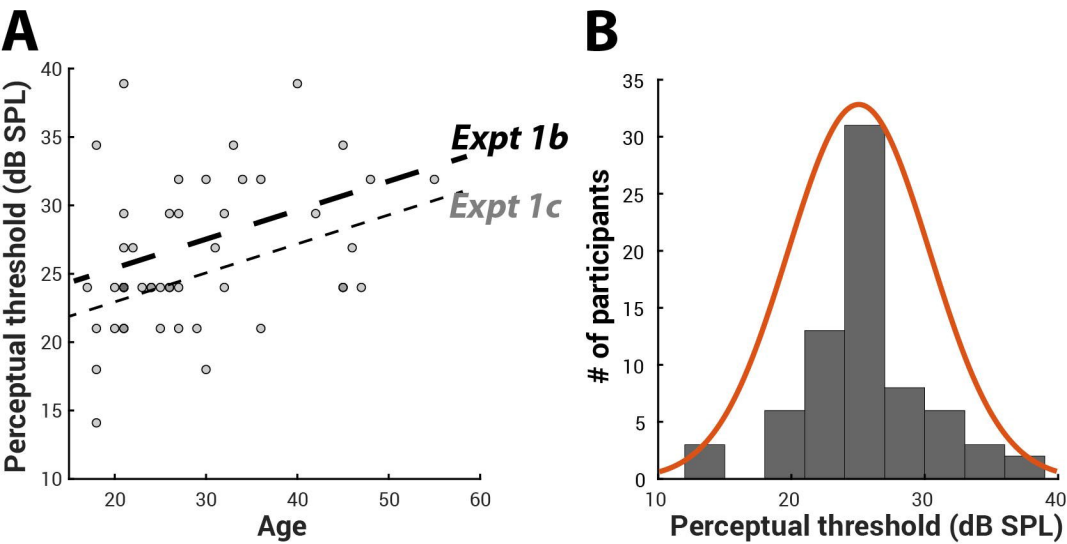
Figure 7. Correlations between the 3-track gold standard threshold (y-axis) and single-track threshold estimates (x-axis) based on the mode of the first 20, 30, or all 40 trials (track 1 (left), 2 (middle), and 3 (right panel)). *Light green crosses and lines refer to 20-trial mode estimates, dark green to the 30-trial estimates, and purple to the 40-trial estimates. Both axes show tone-to-noise dB SNR. Pearson's correlation coefficients for each estimate are shown on the right of the fits.*

Figure 8. Experiment 2b: Probe-signal (n=60). *(A) Distribution of the signal accuracy using the mode-derived threshold. The population mean is labelled as a dash vertical line. (B) The percentage correct detection of 1000-Hz signal and each of the four probes (800, 920, 1080 and 1200 Hz). The thick black line presents the group mean, with error bar = 1 SEM. Each grey line indicates individual data. (C) The accuracy to detect signals (highly probable 1000Hz tones) was significantly higher than the average detection accuracy for probe tones (the less probable 800, 920, 1080, and 1200Hz tones) The population data is presented as a boxplot with the outliers marked as grey crosses. Each grey line indicates individual data, and paired t-test stats reported below the graph. The RT data is shown in the same manner below, D, E and F. For the visualization, the individual data are not presented in (E), but a summary of individual data is shown in (F).*

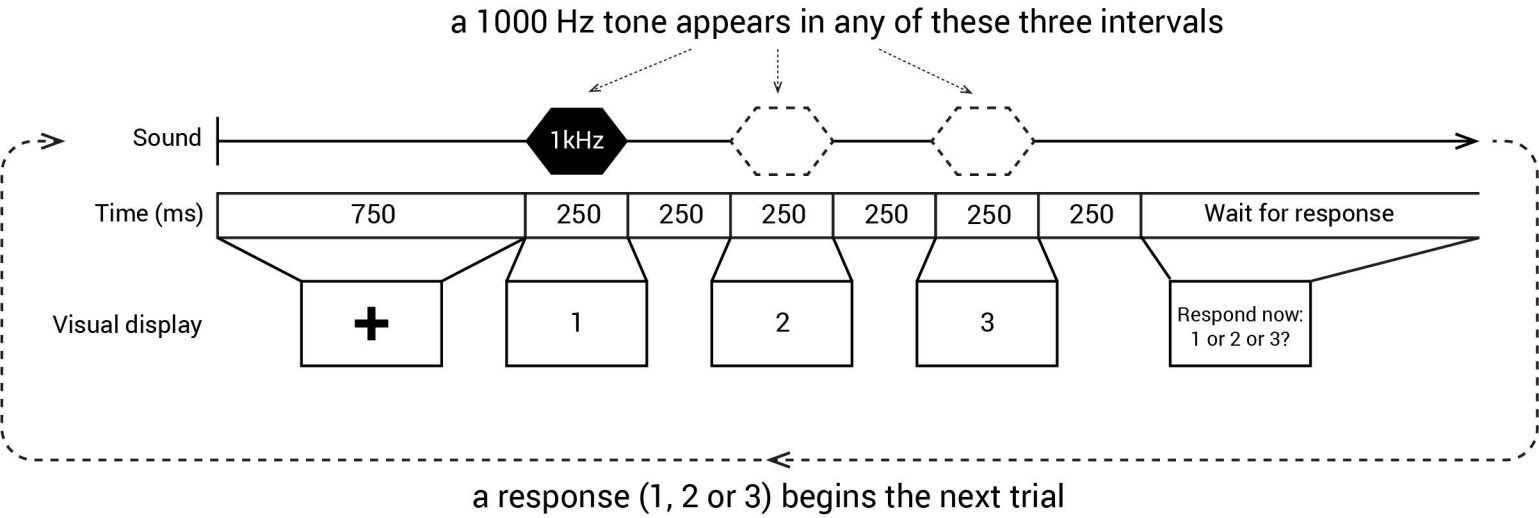
Figure 9. Dynamics of the probe-signal effect, motivation, fatigue and confidence over time. *In all plots, the error bar shows ±1 SEM. (A) Probe-signal effect in accuracy decreased over time. In the left panel, signal and probe accuracy are computed for each block and averaged across participants. The probe-signal effect is computed as Signal Accuracy - Average Probe Accuracy; the group average probe-signal effect is plotted in black in the right panel. (B) Probe-signal effect in RT is shown in the same manner; RT to signal and probe tones are plotted against the block index, with their difference is shown in the right panel. Note that since the probe-signal effect in RT is computed as RT-to-signal minus RT-to-probe, more negative values mean larger probe-signal effects. After each block, participants were instructed to rate how well they did in that block (C), how tired they felt (D), and how motivated they were (E). (F) Graphical interpretation of the findings.*


Sound Pressure Level (dB)
100
80
60
40
20
0
Noise floor
1000 Hz tone @ 0.5 v
1000 Hz tone @ 0.25 v
1000 Hz tone @ 0.125 v
Noise, Measured
Extrapolated
100 93 88 82 76 68 63 56 50 44 37 31 25 19 13 6
Volume Setting (%)

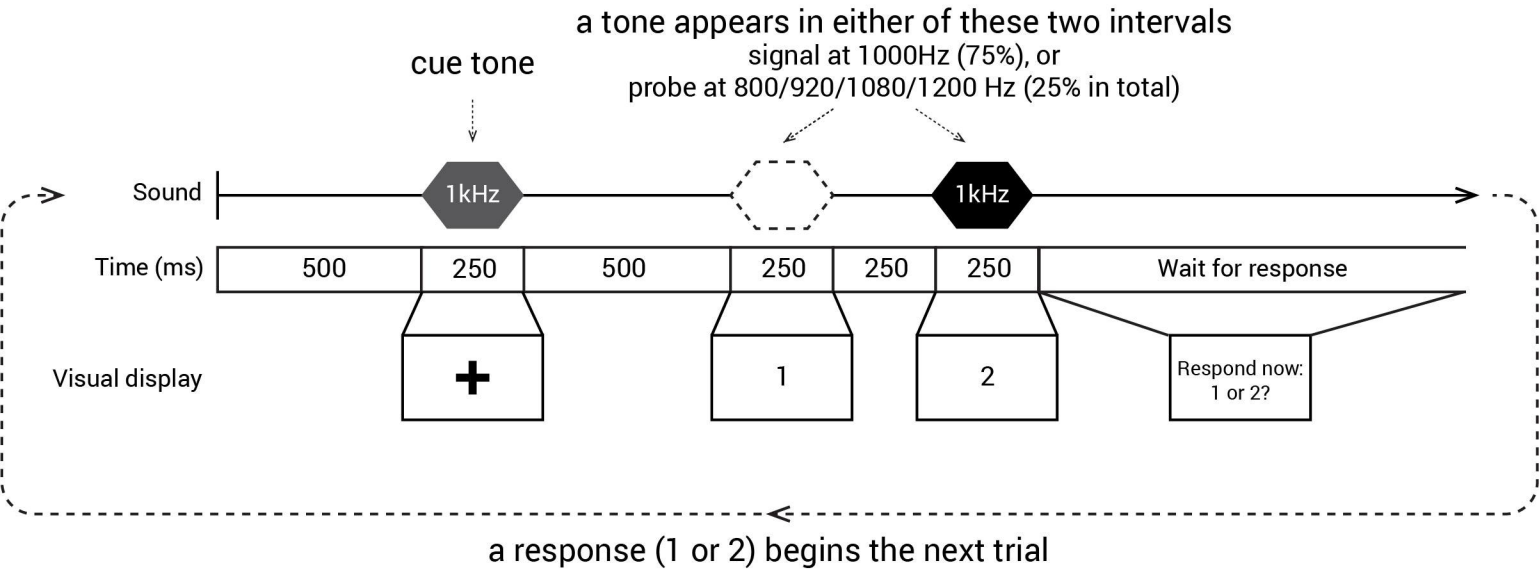

A
# of participants
0
1
2
3
4
5
6
14.1
18
21
24
26.9
29.4
dB SPL
19%
31%
50%
Volume
Perceptual threshold
B
# of participants
0
1
2
3
4
5
6
7
14.1
21
24
26.9
29.4
31.9
34.4
38.9
dB SPL
19%
31%
50%
76%
Volume
Perceptual threshold
C
Outdoors (dB SPL)
15
20
25
30
35
Pearson r = 0.79, p < 0.001
Spearman rho = 0.68, p = 0.001
Indoors (dB SPL)
D
Anechoic chamber
Outdoor park
# of participants
0
2
4
6
8
10
14.1
18
21
22.5
24
26.9
29.4
31.9
34.4
dB SPL
Perceptual threshold

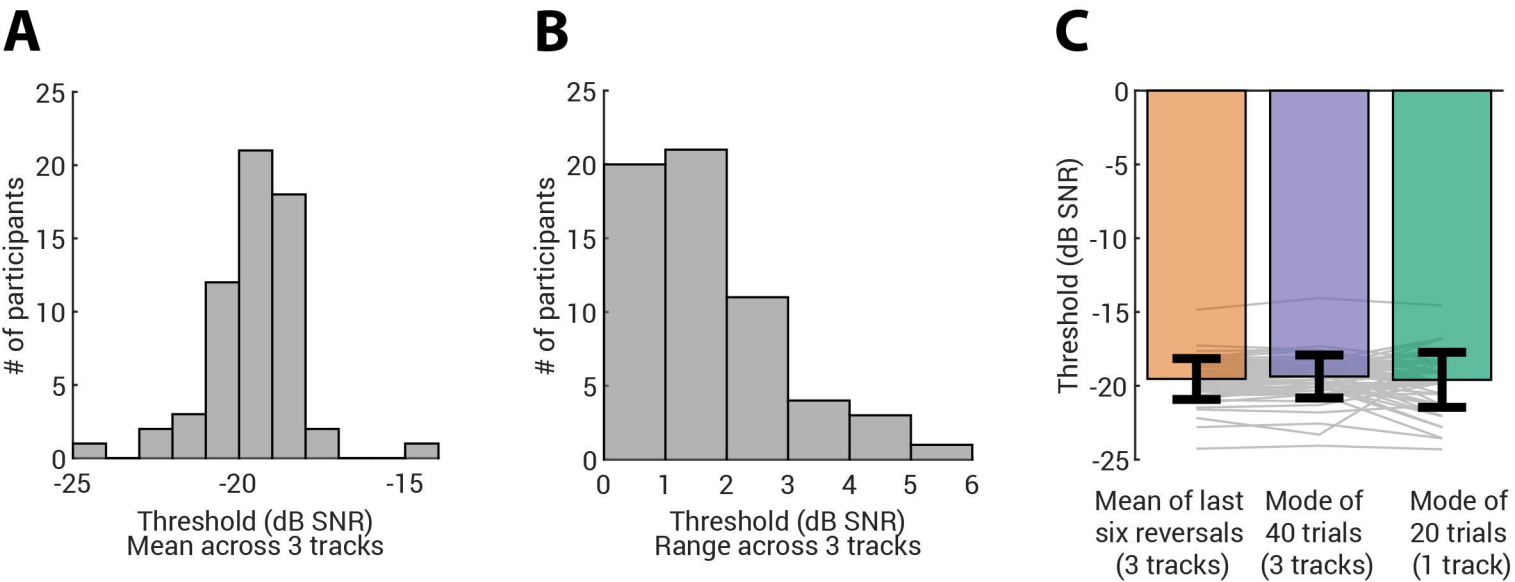

A
Perceptual threshold (dB SPL)
40
35
30
25
20
15
10
20
30
40
50
60
Age
Expt 1b
Expt 1c
B
# of participants
35
30
25
20
15
10
5
0
10
20
30
40
Perceptual threshold (dB SPL)

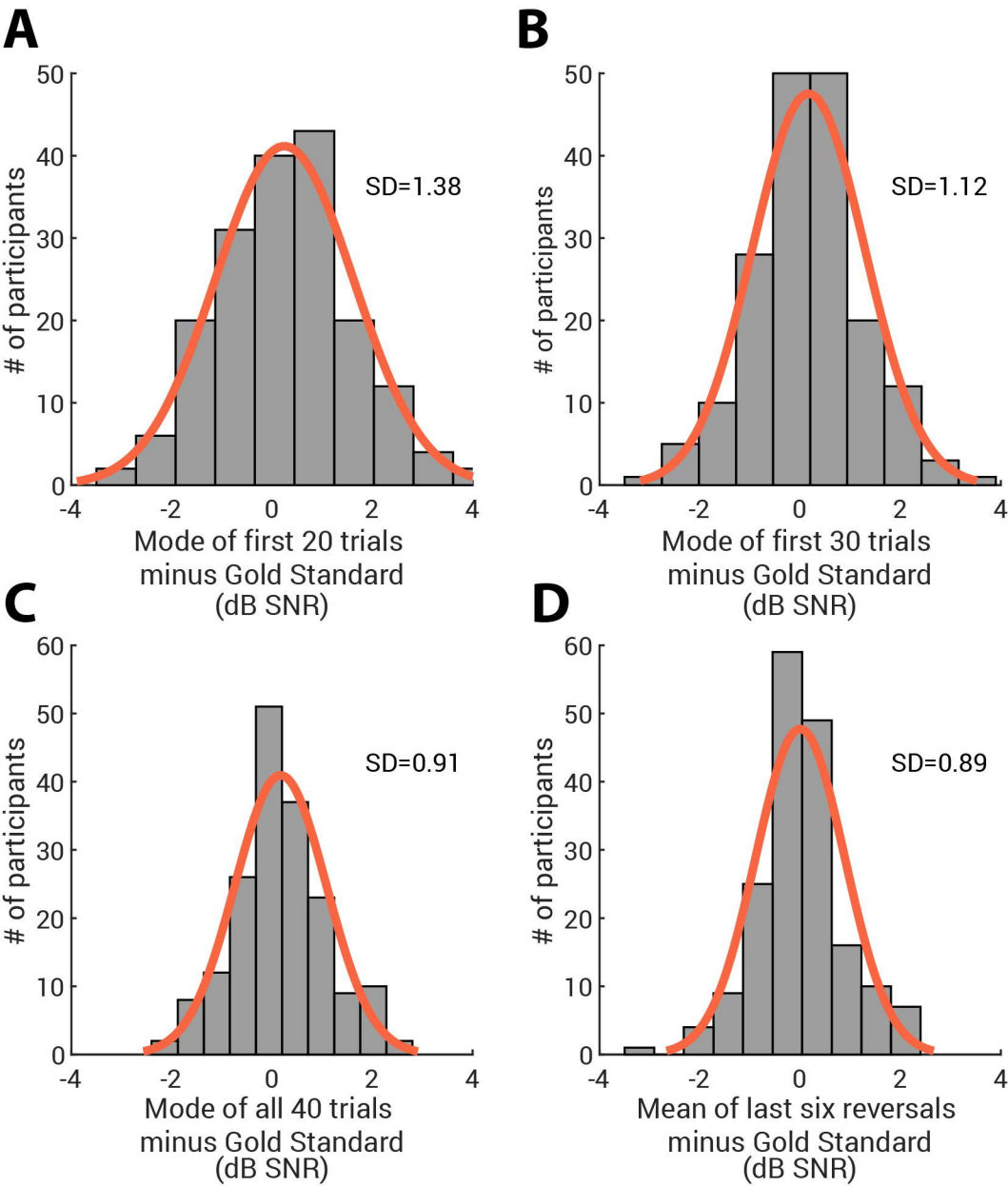
A Experiment 2a : Threshold

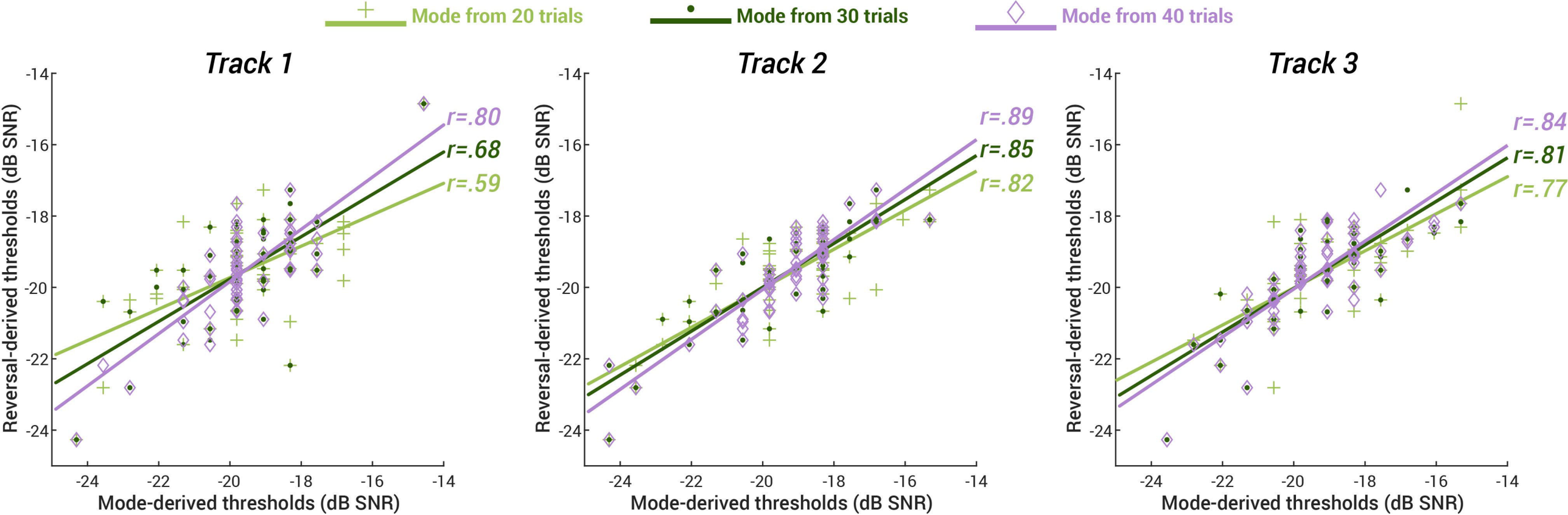


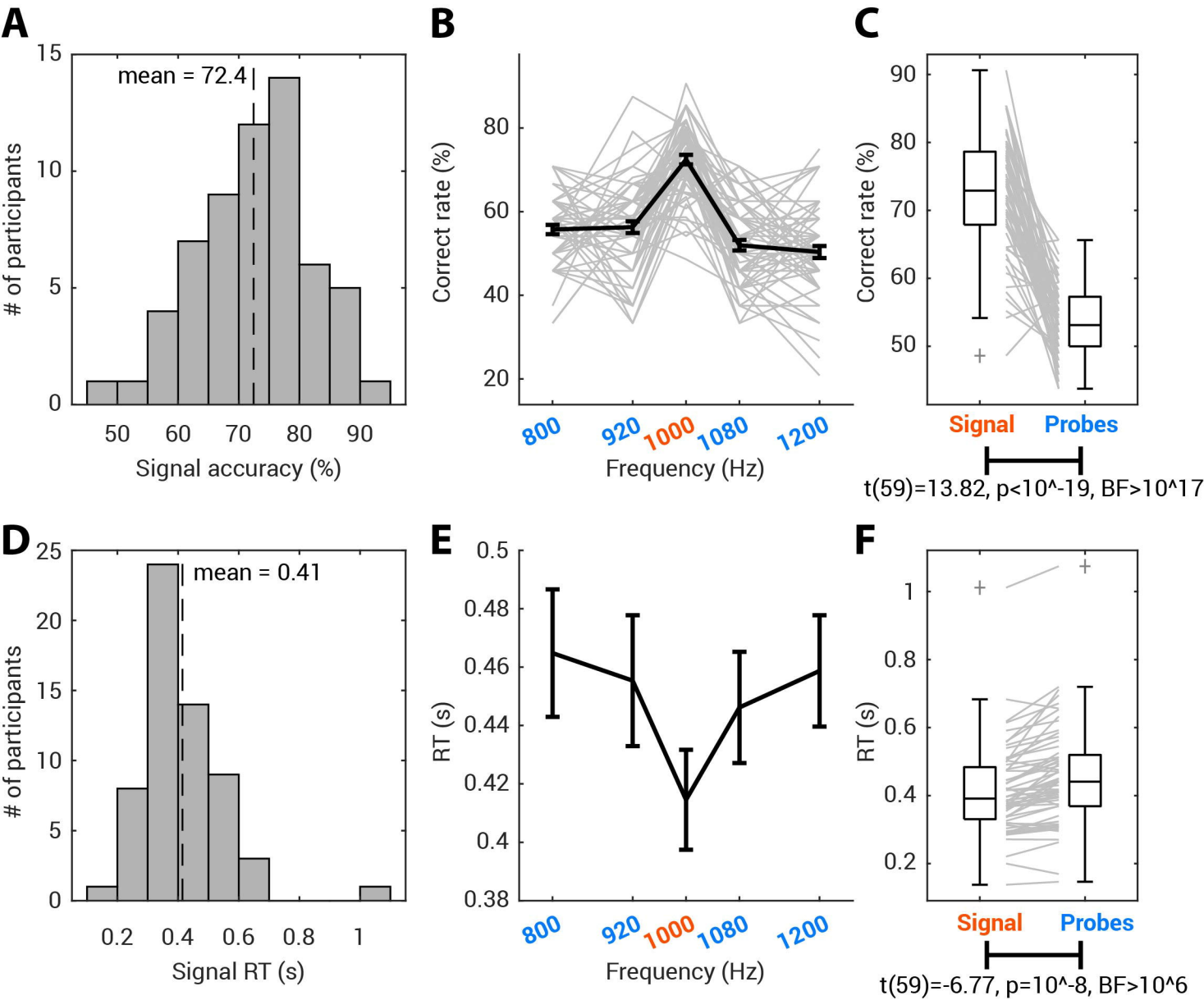
B Experiment 2b : Probe-signal

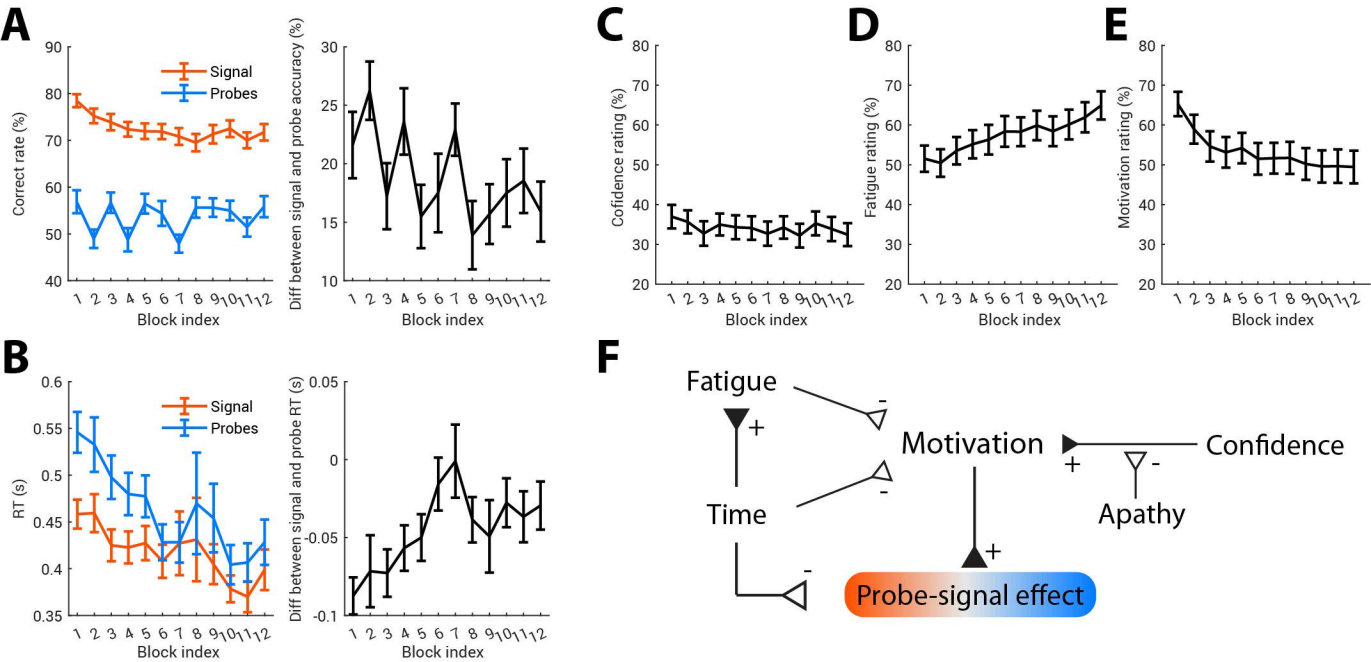



A
25
20
15
10
5
0
# of participants
-25
-20
-15
Threshold (dB SNR)
Mean across 3 tracks
B
25
20
15
10
5
0
# of participants
0
1
2
3
4
5
6
Threshold (dB SNR)
Range across 3 tracks
C
0
-5
-10
-15
-20
-25
Threshold (dB SNR)
Mean of last six reversals (3 tracks)
Mode of 40 trials (3 tracks)
Mode of 20 trials (1 track)


A
B
C
D
# of participants
SD=1.38
SD=1.12
SD=0.91
SD=0.89
Mode of first 20 trials
minus Gold Standard
(dB SNR)
Mode of first 30 trials
minus Gold Standard
(dB SNR)
Mode of all 40 trials
minus Gold Standard
(dB SNR)
Mean of last six reversals
minus Gold Standard
(dB SNR)


Mode from 20 trials
Mode from 30 trials
Mode from 40 trials
Track 1
Track 2
Track 3
r=.80
r=.68
r=.59
r=.89
r=.85
r=.82
r=.84
r=.81
r=.77
Reversal-derived thresholds (dB SNR)
Mode-derived thresholds (dB SNR)
-14
-16
-18
-20
-22
-24


A
mean = 72.4
# of participants
Signal accuracy (%)
B
Correct rate (%)
800
920
1000
1080
1200
Frequency (Hz)
C
Correct rate (%)
Signal
Probes
t(59)=13.82, p<10^-19, BF>10^17
D
mean = 0.41
# of participants
Signal RT (s)
E
RT (s)
800
920
1000
1080
1200
Frequency (Hz)
F
RT (s)
Signal
Probes
t(59)=-6.77, p=10^-8, BF>10^6


A
Signal
Probes
Correct rate (%)
Block index
Diff between signal and probe accuracy (%)
Block index
B
Signal
Probes
RT (s)
Block index
Diff between signal and probe RT (s)
Block index
C
Cofidence rating (%)
Block index
D
Fatigue rating (%)
Block index
E
Motivation rating (%)
Block index
F
Fatigue
Motivation
Confidence
Time
Apathy
Probe-signal effect