# Headphone screening to facilitate web-based auditory experiments

Kevin J. P. Woods[1,2] · Max H. Siegel[1] · James Traer[1] · Josh H. McDermott[1,2]

Published online: 10 July 2017

**Abstract** Psychophysical experiments conducted remotely over the internet permit data collection from large numbers of participants but sacrifice control over sound presentation and therefore are not widely employed in hearing research. To help standardize online sound presentation, we introduce a brief psychophysical test for determining whether online experiment participants are wearing headphones. Listeners judge which of three pure tones is quietest, with one of the tones presented 180° out of phase across the stereo channels. This task is intended to be easy over headphones but difficult over loudspeakers due to phase-cancellation. We validated the test in the lab by testing listeners known to be wearing headphones or listening over loudspeakers. The screening test was effective and efficient, discriminating between the two modes of listening with a small number of trials. When run online, a bimodal distribution of scores was obtained, suggesting that some participants performed the task over loudspeakers despite instructions to use headphones. The ability to detect and screen out these participants mitigates concerns over sound quality for online experiments, a first step toward opening auditory perceptual research to the possibilities afforded by crowdsourcing.

**Keywords** Psychometrics/testing · Stimulus control · Audition

Online behavioral experiments allow investigators to gather data quickly from large numbers of participants. This makes behavioral research highly accessible and efficient, and the ability to obtain data from large samples or diverse populations allows new kinds of questions to be addressed. Crowdsourcing has become popular in a number of subfields within cognitive psychology (Buhrmester et al., 2011; Crump et al., 2013), including visual perception (Brady and Alvarez, 2011; Freeman et al., 2013; Shin and Ma, 2016), cognition (Frank and Goodman, 2012; Hartshorne and Germine, 2015), and linguistics (Sprouse, 2010; Gibson et al., 2011; Saunders et al., 2013). Experimenters in these fields have developed methods to maximize the quality of web-collected data (Meade and Bartholomew, 2012; Chandler et al., 2013). By contrast, auditory psychophysics has not adopted crowdsourcing to the same degree as other fields of psychology, presumably due in part to concerns about sound presentation. Interference from background noise, the poor fidelity of laptop speakers, and environmental reverberation could all reduce control over what a participant hears.

One simple way to improve the control of sound delivery online is to ensure that participants are wearing headphones or earphones (for brevity the term "headphones" will henceforth be used to refer to both). Headphones tend to attenuate external sources by partly obscuring the ear, and minimize the distance between eardrum and transducer, thus improving signal-to-noise ratios in favor of the sounds presented by the experimenter. Headphones also enable presentation of separate signals to the two ears (enabling binaural tests). Here we present methods to help ensure that participants are wearing headphones, along with validation of this method in the lab,

**Electronic supplementary material** The online version of this article (doi:10.3758/s13414-017-1361-2) contains supplementary material, which is available to authorized users.

✉ Kevin J. P. Woods
kwoods@mit.edu

[1] Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[2] Program in Speech and Hearing Bioscience and Technology, Harvard University, 260 Longwood Ave, Boston, MA 02115, USA

where we knew participants to be listening over headphones or over loudspeakers.

We checked that participants are wearing headphones using an intensity-discrimination task involving tones that sometimes have a phase difference of 180° between stereo channels—these *anti-phase* tones are heavily attenuated when played through loudspeakers, but are not attenuated over headphones. An example of this phenomenon is shown in Fig. 1, which displays the result of combining two sine waves with ≈ 180° phase difference (i.e., anti-phase), as would reach an ear at a location between the speakers. The phase relationship at the ear varies depending on the exact location of the listener but remains close to anti-phase for wavelengths that are large relative to the distance between speakers (as in Fig. 1). We used this property of sound to test whether an online participant is listening over loudspeakers rather than headphones: A simple "which tone is quietest?" task containing anti-phase tones—heard differently over headphones versus stereo loudspeakers—produces a different pattern of responses in participants who are not wearing headphones.

## Method

### Simulation and acoustic measurement of anti-phase attenuation

The relative phase of signals from stereo loudspeakers depends on the position of the listener and the two speakers. For a listener sitting at a computer, it was unclear a priori whether anti-phase attenuation would be sufficiently robust to variation in head position to be used effectively in online headphone screening. We simulated attenuation (power of an in-phase tone relative to an anti-phase tone) at frequencies from 200 Hz to 3.2 kHz and found that only at the lowest frequencies would attenuation hold over broad regions of space (Fig. 2A; see *Supplemental Materials*). These simulations suggested that the test tone should be as low in frequency as possible while remaining above the bass-range rolloff in frequency response seen in many commercial headphones, which tends to begin near 100 Hz (Gutierrez-Parera et al., 2015). Given these considerations, we settled on 200 Hz as the stimulus frequency and then measured the resulting attenuation empirically to confirm the simulation results. We used a head-and-torso simulator with in-ear microphones (KEMAR, G.R.A.S) and placed it at various locations relative to desktop and laptop speakers. Desktop speakers were placed with their centers 40-cm apart and set 40-cm back from the edge of the table (Fig. 2B); this setup and testing space was that used in Experiment 1 (which evaluated the screening test in the lab). Measurements with laptop speakers used a single laptop (Dell XPS 13) in one of the testing spaces from Experiment 2 (which evaluated the screening test at four locations in our departmental building). In each case, a 200-Hz tone (in-phase) was set to a level of 70 dB SPL at the central measurement position (using a level meter). Then, the head-and-torso simulator was used to measure the difference between in-phase and anti-phase tones (the anti-phase attenuation) at each of the locations depicted in Fig. 2B. Attenuation was similar at the left and right ears and was averaged together to yield the attenuation values plotted in Fig. 2B.

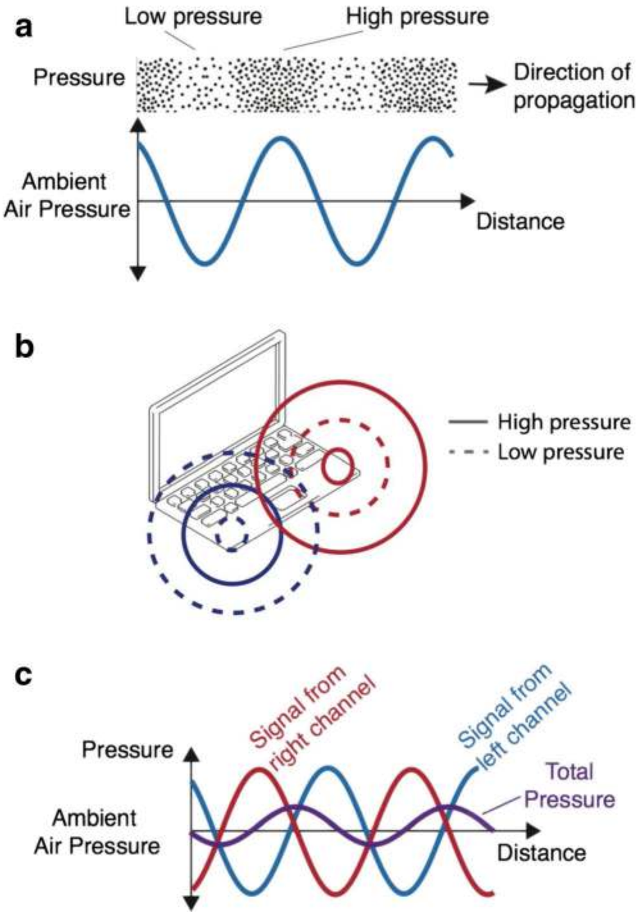

**Fig. 1** Attenuation of tones by phase-cancellation. (**a**) Schematic of sound pressure variation produced by a sinusoidal waveform. Sound is a pressure wave with neighboring regions of high (peaks) and low (troughs) of air pressure. (**b**) Schematic of a computer whose two speakers emit tones in anti-phase: the right speaker (red) emits a high pressure peak (solid lines) while the left speaker (blue) emits a low-pressure trough (dashed lines). Some distance away from the computer, the relationship is reversed. (**c**) Waveforms of two sinusoids close to anti-phase, and their superposition. Pressure from two waves sum linearly, such that when the waves are close to anti-phase (as is the case when the distance to the right and left speakers is similar) their sum is of lower power than either of the constituents

### Screening task

We used six trials of a 3-AFC "Which tone is quietest?" task: All three tones were 200-Hz pure tones with a duration of 1,000 ms, with 100 ms on- and off-ramps (produced by half

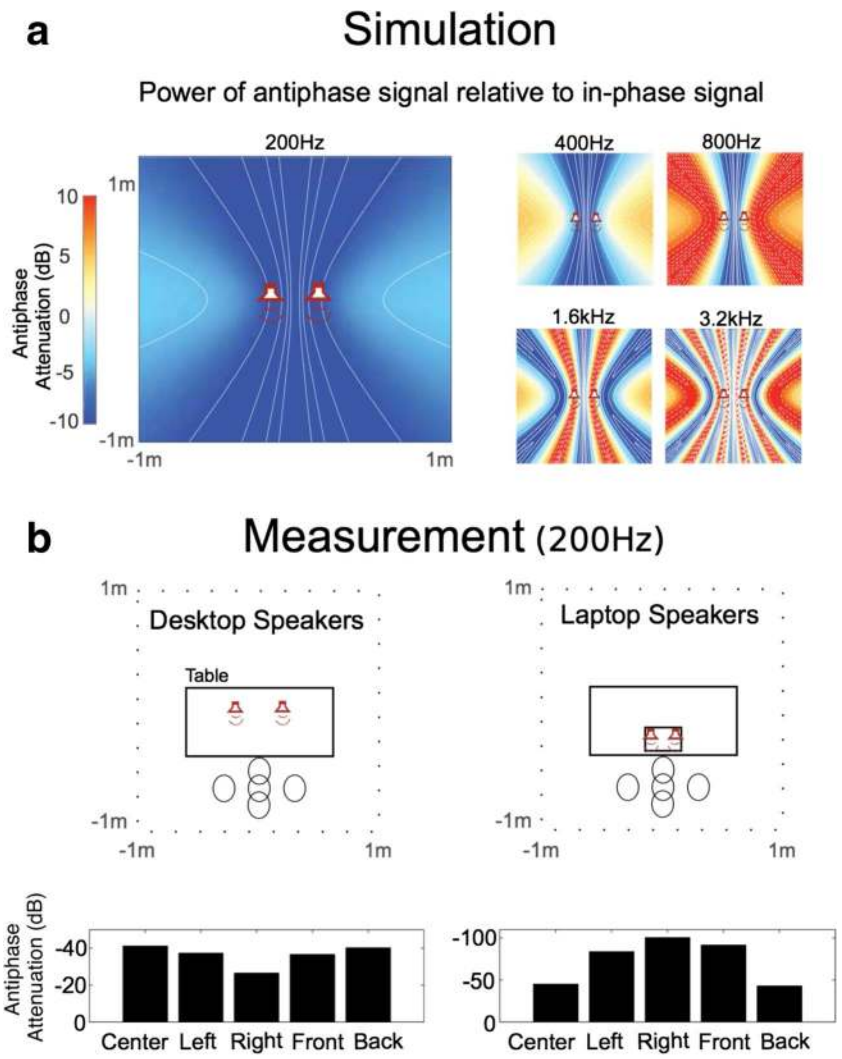

**Fig. 2** Simulation and measurement of attenuation over space. (**a**) Attenuation of anti-phase sinusoids simulated in free-field listening conditions with uniformly radiating speakers. We plot the computed attenuation over a 2-m x 2-m region centered on the speakers. In all subsequent attenuation plots, we show the same plane and use the same color scale. This color scale is truncated; values exceeding −10 dB are depicted at −10 dB. Solid contour lines indicate negative values, and dashed contour lines indicate positive values. The screening test works best at low frequencies, at which anti-phase signals are always attenuated. At higher frequencies, the anti-phase signal is amplified in some locations, and thus results could differ substantially depending on head position. (**b**) Measurement of anti-phase attenuation using desktop and laptop setups. For each setup a KEMAR head-and-torso simulator was placed in five locations as depicted. The laptop speakers were both closer together and closer to the listener. Attenuation was similar in the left and right ears; plots show attenuation averaged across the two ears

of a Hann window). A 3-AFC task (rather than 2-AFC) was chosen to reduce the probability of passing the screen by randomly guessing. A low-tone frequency (200 Hz) was chosen to produce a broad region of attenuation (Figure S1) intended to make the test robust to variation in head position. One of the tones had a level of −6 dB relative to the other two (which were of equal intensity). However, one of the two equal-intensity tones was phase reversed between the stereo channels; the other two tones had no phase difference between stereo channels (starting phases in the L/R channels were therefore 0°/0°, 0°/0°, and 0°/180° for the less intense and two more-intense tones respectively). On each trial, the three tones were presented in random order with an interstimulus interval of 500 ms. The listener was asked to pick the interval containing the quietest tone, by selecting one of three buttons labeled "FIRST sound is SOFTEST," "SECOND sound is SOFTEST," and "THIRD sound is SOFTEST."

### In-lab implementation

**Experiment 1** Participants completed the task on a Mac mini-computer in a quiet office environment using the same Mechanical Turk interface used by online participants. Half of participants (N = 20, 15 females, mean age = 27.6 years, standard deviation [SD] = 12.7) completed the task while listening to stimuli over Sennheiser HD 280 headphones.

The other half (N = 20, 11 females, mean age = 26.5 years, SD = 5.6) completed the task while listening over a pair of Harman/Kardon HK206 free-field speakers. The speakers were placed so that their centers were 40-cm apart and were set 40-cm back from the edge of the table at which the participant was seated (i.e., set at approximately ±30° relative to the listener). In both conditions, sound levels were calibrated to present tones at 70-dB SPL at the ear (using a Svantek sound meter connected either to a GRAS artificial ear or to a GRAS free-field microphone). In all other respects, the experiment was identical to the online experiment.

**Experiment 2** In a separate experiment, we invited participants to bring their own laptops into the lab (N = 22, 13 females, mean age = 27.3 years, SD = 10.6) and tested them over their laptop speakers in four different locations around the building (in random order). These testing spaces were selected to cover a range of room sizes and to offer different reflective surfaces nearby the listener. For example, in one room (Server room: Adverse) the laptop was surrounded by clutter including cardboard boxes and drinking glasses; in another room (Atrium), the laptop was placed alongside a wall in a very large reverberant space. Two of the spaces (Atrium and Ping-pong room) were open to use by others and had commensurate background noise. Participants were told to use the laptop as they normally would, without moving it from its predetermined location in the room.

### Online implementation

We ran crowdsourced experiments with Amazon's Mechanical Turk, a service that allows remote participants to perform simple tasks in their web browser. Experiment scripts were written in html/Javascript. Participants were offered a base rate of $0.35 to complete the screening task, which takes about 3 minutes (this includes the time required to read task instructions and provide basic demographic information). Those who passed the screening task were offered a bonus payment to complete an additional task (not described here but typically the task of primary interest to the experimenter), with the amount of the bonus commensurate with the expected time and effort for that task. Mechanical Turk allows participants' results to be "rejected" without pay or "accepted" for pay. Participants could only run our tasks if more than 90% of their previous tasks had been "accepted" (Peer et al., 2013). Participants were restricted to the United States or Canada. Demographic information was collected, including age and hearing status ("Are you aware of any hearing loss?").

The online screening task began with the repeated presentation of a noise sample for loudness calibration. This was intended to help avoid presentation levels that would result in uncomfortably loud or inaudible stimuli during the main experiment (after screening), rather than being calibration for the screening task. As such, the calibration noise was spectrally matched to stimuli used in our experiments (it was a broadband, speech-shaped noise). Participants were asked to adjust their computer volume such that the noise sample was at a comfortable level. The rms of the stored noise sample waveform was 0.30; this is as high as possible subject to the constraint of avoiding clipping. Relative to this calibration noise, the levels of the test tones presented in the screening task were −6.5 dB (for the two more intense tones) and −12.5 dB (for the less intense tone). We expect that this screening task should be robust to different level settings as long as the (in-phase) test tones are audible. Nonetheless, if presentation level was set such that test tones were inaudible, we would expect listeners to perform at chance.

To pass the headphone screening, participants must correctly answer at least five of the six level discrimination trials. No feedback was provided. Responses are scored only if all trials are completed. Because we use a three-alternative task, correctly answering five or more of the six trials by guessing is unlikely (it should occur with a probability of 0.0178). Most participants who are not engaged with the task should be screened out. If a participant is engaged but is listening over speakers rather than over headphones, then the tone in anti-phase will be heavily attenuated due to cancellation and should be judged (incorrectly) as the least intense of the three tones. In such a situation, the participant is again unlikely to give the correct response on five of six trials and in fact should perform below the chance level of two correct trials.

**Experiment 3** The online screening task was run on 5,154 participants (2,590 females, mean age = 34.5 years, SD = 11.1). The 184 (3.6%) reporting hearing impairment were included in our general analysis (i.e., not analyzed separately). Listeners unable to hear the 200-Hz test tone due to hearing loss (or for any other reason) would likely be screened out.

**Experiment 4** A control task with all three tones in-phase (i.e., no anti-phase tones) was also run online, with 150 participants (75 females, mean age = 38.5 years, SD = 11.7). The three participants (2%) who reported hearing impairment were included in our general analysis.

## Results

### Simulation and acoustic measurement of anti-phase attenuation

The screening test relies on the attenuation of the anti-phase tone when played in free-field conditions. We thus first evaluated the extent of the attenuation produced by anti-phase tones. We used simulations to choose an appropriate test frequency and then made measurements to assess the degree of attenuation in realistic listening conditions.

Figure 2A shows the expected attenuation over space in ideal free-field conditions (see *Supplemental Materials*). In simulations, the test frequency used in the screening test (200 Hz) produces consistent attenuation over a broad region of space, making the attenuation effect robust to variations in head position. Higher frequencies produce attenuation that depends sensitively on head position and thus are not ideal for our screening task. Figure 2B shows measurements of attenuation of a 200-Hz anti-phase tone using a head-and-torso simulator placed at various locations relative to the speakers. Attenuation is greater than −20 dB in every case, substantially exceeding the −6 dB required for the screening test.

### In-lab experiments

To validate the task, we ran it in the lab, with participants either wearing headphones or listening over loudspeakers (Experiment 1; Fig. 3). Each participant completed six trials, as in the online experiment. The results show that our screening task was effective at distinguishing participants who were listening over headphones from those listening over loudspeakers: 20 of 20 participants wearing headphones passed the test, whereas 19 of 20 participants listening over loudspeakers did not. Critically, the task achieves good discrimination between headphone and laptop listening with just a small number of trials. The short duration of this screening task is intended to facilitate its use online, where it might be desirable to run relatively brief experiments.

To test our screening task in more arduous and varied conditions, we asked a second set of participants to use the speakers on their own laptops in several locations within the Brain and Cognitive Sciences Building at MIT (see *Methods*). Unlike the online task and the previous in-lab experiment, these participants ran the task four times in a row, rather than just once, to enable testing the robustness of the results across four different testing rooms. Because practice effects (due to, e.g., familiarity with the stimulus, or setting the volume differently) could have produced a performance advantage for this experiment relative to the experiment over desktop speakers, we examined the results for just the first run for each participant (Fig. 4A) in addition to that for all four runs combined (Fig. 4B). We additionally examined the mean score across all four runs for each participant (Fig. 4C) to get an indication of whether certain participants were consistently able to perform the task without headphones.

Administering the test over laptop speakers (Fig. 4) again produced substantially worse performance than when participants were wearing headphones (Fig. 3, in blue), although it elicited a different pattern of responses than our test with desktop-speakers (K-S test between distributions of Figs. 3 and 4B in red, $p < 0.05$, D = 0.37), with a greater proportion passing our threshold (>4 correct). The screening test thus failed to detect 4 of 22 participants using laptop speakers, a modest but nonnegligible subset of our sample. The distribution of participants' mean scores (Fig. 4C) indicates that some participants performed poorly in all rooms (mean scores in the range 0-1) while some performed well in all rooms (mean scores in the range 5-6). Examining scores obtained in each room (Figure S2) also suggests that the testing space had little impact on performance. Instead, the difference in performance could have arisen from variation in laptop speaker designs or variation in distance from the ears to the speakers due to user behavior (e.g., leaning in). Some participants (3/22) even reported using vibrations felt on the surface of the laptop to perform the task. Because 200 Hz is within the range of vibrotactile stimulation, and because phase-cancellation could also occur in surface vibrations, using touch instead of free-field hearing might not necessarily alter the expected results. However, this strategy could possibly improve performance if vibrations in the laptop-case fail to attenuate to the same degree they would in the air, for instance if a participant placed their hand close to a speaker.

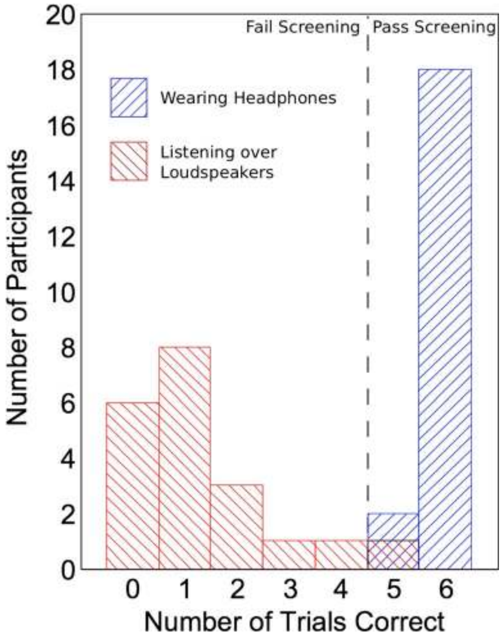

**Fig. 3** In-lab validation of headphone screening task (Experiment 1). Results from 40 participants run in the lab on 6 trials of the screening task; 20 listened over loudspeakers, 20 listened over headphones. The dashed line here and elsewhere indicates a recommended threshold for online headphone screening using a 6-trial task. Because the task is 3-AFC, chance performance would yield 2 trials correct on average

Figures 3 and 4 suggest that our screening task is more effective (i.e., produces lower scores absent headphones) when desktop speakers, rather than laptops, are used. This might be expected if desktop speakers generally sit farther from the listener, because anti-phase attenuation with low-frequency tones becomes more reliable as distance to the listener increases (Figure S1B).

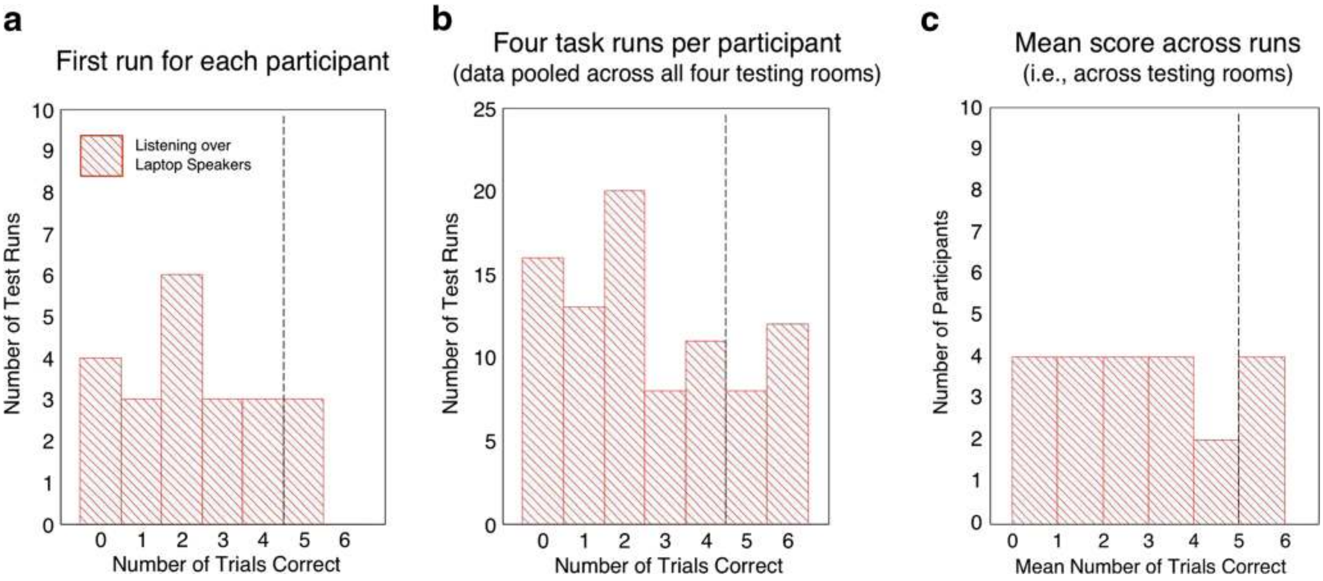

**Fig. 4** In-lab screening task run through loudspeakers on participants' own laptops (Experiment 2). Results from 22 participants run in the lab. Each participant performed the 6-trial screening task 4 times—once in each of 4 rooms, in random order. (**a**) Histogram of scores obtained on the first test run for each participant. (**b**) Pooled results showing 88 runs of the screening task, 4 runs per participant. (**c**) Participants' mean scores across the four rooms, binned from 0-1, 1-2,…5-6. Mean scores exactly equal to bin limits are placed in the higher-scoring bin (e.g., a participant with a mean score of 5.0 is placed in the "5-6" bin)

The dependence of test effectiveness on hardware raises the question of what sort of listening setup online participants will tend to have. To address this issue, for a portion of our online experiments (described below), we queried participants about this on our online demographics page. We found them split rather evenly between desktops and laptops. In the brief experiment run with this question added, 97 participants said they were using desktops while 107 said they were using laptops (45.8% and 50.5% respectively). The remaining 8 participants (3.6%) said they were using other devices (e.g., tablet, smartphone).

## Online experiments

The cumulative pass rate (with passing defined as at least 5 of 6 correct trials) for headphone screening tasks we have run online is 64.7% (3,335 of 5,154 participants). The distribution of scores for these participants (Fig. 5) contains modes at 0 and 6 trials correct; confidence intervals (95%) obtained by bootstrapping indicate that the mode at zero is reliable. Given that chance performance on this task produces two of six trials correct on average, the obtained distribution of scores is difficult to explain by merely supposing that some participants are unmotivated or guessing. Instead, the systematic below-chance performance suggests that some participants were not wearing headphones: participants attempting in earnest to perform the task over stereo loudspeakers might be expected to score below chance, because the sound heard as quietest under those conditions— the anti-phase tone—is always the incorrect response.

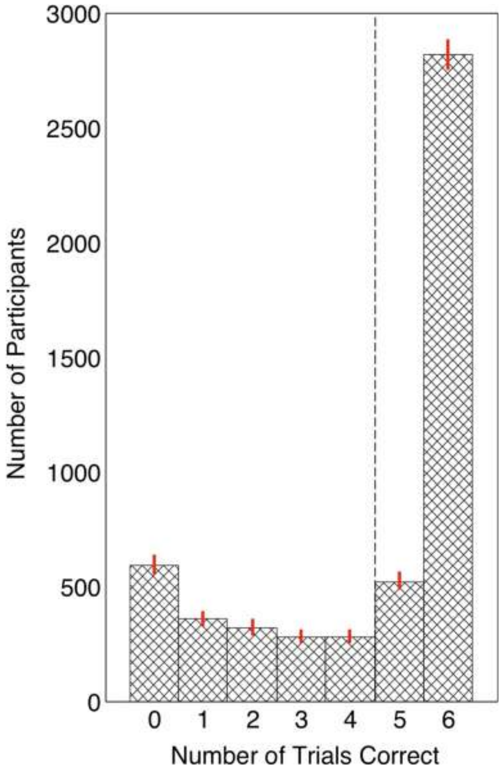

**Fig. 5** Online distribution of scores for the headphone screening task with anti-phase tones (Experiment 3). Results from 5,154 participants run online on the 6-trial screening task. Chance performance would yield 2 trials correct on average. Below-chance performance is expected if the task is attempted over stereo loudspeakers or if task instructions are unintentionally reversed by the participant. Error bars depict 95% confidence intervals from 1,000 bootstrapped samples

Another explanation for below-chance performance on the screening task is that participants tend to confuse the instructions in a way that leads to consistently incorrect responses (for example, attempting to select the loudest rather than softest of the 3 tones). To evaluate this possibility, we ran a control version of the screening task (conducted online) in which no tones were in anti-phase (i.e., all three tones had starting phases in the L/R channels of 0°/0°), such that listening over speakers should not produce below-chance performance if participants were otherwise following the instructions correctly. The screening task was otherwise identical to the previous experiments. Results from 150 online participants are shown in Fig. 6. As before, chance performance should yield two trials correct on average.

The scores obtained from this control version of the screening task are distributed differently from the scores from our standard task (K-S test, $p < 0.0001$, D = 0.24). In particular, there are far fewer below-chance scores. This result suggests that the preponderance of below-chance scores observed in the standard task (i.e., when anti-phase tones are used; Fig. 4) is not due to confusion of instructions. The control task results also reveal that some proportion of online participants are screened out for poor performance even without anti-phase tones—given a pass threshold of 5 or more trials correct, 18 of 150 participants (12.0%) in this control task would have failed to pass screening (35.3% fail in the standard task with anti-phase tones). In contrast, none of the 20 participants who performed the task in the lab over headphones would have been screened out (Fig. 3). Our procedure appears to act as a screen for a subset of online participants that perform poorly (e.g., due to low motivation, difficulty understanding the instructions, or adverse environmental listening conditions), in addition to screening out those attempting the task over loudspeakers.

## Discussion

We developed a headphone screening task by exploiting phase-cancellation in free-field conditions coupled with dichotic headphone presentation. The screening consisted of six trials of a 3-AFC intensity discrimination task. In the lab, participants with headphones performed very well, whereas participants listening over loudspeakers performed very poorly. When run online (where we cannot definitively verify the listening modality), a distribution of scores was obtained that suggests some participants were indeed listening over loudspeakers despite being instructed to wear headphones and can be screened out effectively with our task.

The effectiveness of our screening task can be considered in terms of two kinds of screening errors: screening out participants who are in fact wearing headphones, or passing participants who are not wearing headphones. The first type of error (excluding participants despite headphone use) can result from poor performance independent of the listening device, because participants unable to perform well on a simple 3-AFC task are screened out. This seems desirable, and the cost of such failures is minimal since participants excluded in this way are easily replaced (especially in online testing). The second type of screening error (including participants who are not wearing headphones) is potentially more concerning since it permits acquisition of data from listeners whose sound presentation may be suspect. The relative rates of each kind of error could be altered depending on the needs of the experimenter by changing the threshold required to pass the screening task. For example, requiring >5 correct instead of >4 correct would result in a screen that is more stringent, and would be expected to increase errors of the first kind while reducing errors of the second kind.

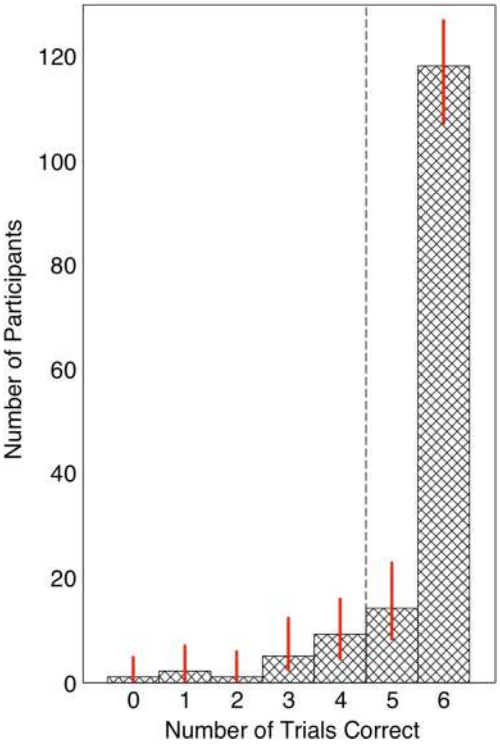

**Fig. 6** Online performance for a version of the headphone screening task *without* anti-phase tones (Experiment 4). All tones had the same phase across stereo channels, removing the manipulation that differentiates listening modes (headphones vs. speakers). This is intended to control for the possibility that the below-chance performance observed in Fig. 5 was due to confusion of task instructions. Results from 150 participants run online on the control screening task. Error bars depict 95% confidence intervals from 1,000 bootstrapped samples. Chance performance would yield 2 trials correct on average

### Differences between in-lab and online experiments

We found that online participants were much more likely to fail the headphone check than in-lab participants who were wearing headphones (failure rates were 35.3% vs. 0%, respectively). What accounts for the relatively low pass rate of this task online? As argued above, the tendency for below-chance performance suggests that some participants were not in fact wearing headphones despite the task instructions, but this might not be the only difference. Hearing impairment in online participants seems unlikely to have substantially contributed to the online pass rate, because just 3.6% reported any impairment. It is perhaps more likely that some participants wore headphones but did not understand the task instructions. Prior studies using crowdsourcing have observed that a significant number of participants fail to follow instructions, potentially reflecting differences in motivation or compliance between online and in-lab participants. As such, it is standard for experiments to contain catch trials (Crump et al., 2013). Our screening task may thus serve both to screen out participants who ignored the instructions to use headphones as well as participants who are unwilling or unable to follow the task instructions. Both these functions likely help to improve data quality.

### Limitations and possibilities in crowdsourced auditory psychophysics

Although our methods can help to better control sound presentation in online experiments, crowdsourcing obviously cannot replace in-lab auditory psychophysics. Commercially available headphones vary in their frequency response and how tightly they couple to the ear, thus neither the exact spectra of the stimulus nor the degree of external sound attenuation can be known. This precludes the option of testing a participant's hearing with an audiogram, for instance. In addition, soundcards and input devices may have small, unknown time delays, making precise measurement of reaction times difficult. Because environmental noise is likely to remain audible in many situations despite attenuation by headphones, online testing is inappropriate for experiments with stimuli near absolute threshold and may be of limited use when comparing performance across individuals (whose surroundings likely vary). Microphone access could in principle allow experimenters to screen for environmental noise (or even for headphone use), but this may not be possible on some computer setups, and even when possible may be precluded by concerns over participants' privacy. We have also noted cases in which our screening method could be affected by uncommon loudspeaker setups: for example, subwoofer speakers that broadcast only one audio channel (as may occur in some desktop speaker setups, as well as high-end "gaming" laptops and recent models of the Macbook Pro), setups that combine stereo channels prior to output (as may occur in devices with just one speaker), or speakers with poor low-frequency response that render the test tones inaudible. In many of these cases participants would be screened out as well, but the mechanism by which the screening operates would not be as intended.

The limitations of online experiments are less restrictive for some areas of research than others. In many situations, precise control of stimulus level and spectrum may not be critical. For instance, experiments from our own lab on attention-driven streaming (Woods and McDermott, 2015) and melody recognition (McDermott et al., 2008) have been successfully replicated online.

Crowdsourcing has the potential to be broadly useful in hearing research because it allows one to ask questions that are difficult to approach with conventional lab-based experiments for practical reasons. For example, some experiments require large numbers of participants (Kidd et al., 2007; McDermott et al., 2010; Hartshorne and Germine, 2015, Teki et al., 2016) and are much more efficiently conducted online, where hundreds of participants can be run per day. Experiments may also require recruiting participants from disparate cultural backgrounds (Curtis and Bharucha, 2009; Henrich et al., 2010) that are more readily recruited online than in person. Alternatively, it may be desirable to run only a small number of trials on each participant, or even just a single critical trial (Simons and Chabris, 1999; Shin and Ma, 2016), after which the participant may become aware of the experiment's purpose. In all of these cases recruiting adequate sets of participants in the lab might be prohibitively difficult, and online experiments facilitated by a headphone check could be a useful addition to a psychoacoustician's toolbox.

**Acknowledgments** This work was supported by an NSF CAREER award and NIH grant 1R01DC014739-01A1 to J.H.M. The authors thank Malinda McPherson, Alex Kell, and Erica Shook for sharing data from Mechanical Turk experiments, Dorit Kliemann for help recruiting subjects for in-lab validation experiments, and Ray Gonzalez and Kelsey R. Allen for organizing code for distribution.

Code implementing the headphone screening task can be downloaded from the McDermott lab website (http://mcdermottlab.mit.edu/downloads.html).

## References

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science, 22,* 384–392.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods, 46,* 112–130.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE, 8,* e57410.

Curtis, M. E., & Bharucha, J. J. (2009). Memory and musical expectation for tones in cultural context. *Music Perception: An Interdisciplinary Journal, 26,* 365–375.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336,* 998.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16,* 974–981.

Gardner, W. G. (2002). Reverberation algorithms. In *Applications of digital signal processing to audio and acoustics* (pp. 85–131). Springer US.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language Linguistics Compass, 5,* 509–524.

Gutierrez-Parera, P., Lopez, J. J., & Aguilera, E. (2015). On the influence of headphone quality in the spatial immersion produced by Binaural Recordings. In *Audio Engineering Society Convention 138*. Audio Engineering Society.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science, 26*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466,* 29.

Jensen, F. B., Kuperman, W. A., Porter, M. B., & Schmidt, H. (2000). *Computational ocean acoustics*. Springer Science & Business Media.

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America, 122,* 418–435.

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychological Science, 19,* 1263–1271.

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual differences reveal the basis of consonance. *Current Biology, 20,* 1035–1041.

Meade, A. W., & Bartholomew, S. (2012). Identifying careless responses in survey data. *Psychological Methods, 17,* 437–455.

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavioral Research Methods, 46,* 1023–1031.

Saunders, D. R., Bex, P. J., & Woods, R. L. (2013). Crowdsourcing a normative natural language dataset: A comparison of Amazon Mechanical Turk and in-lab data collection. *Journal of Medical Internet Research, 15,* e100.

Shin, H., & Ma, W. J. (2016). Crowdsourced single-trial probes of visual working memory for irrelevant features. *Journal of Vision, 16,* 10.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28,* 1059–1074.

Sprouse, J. (2010). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavioral Research Methods, 43,* 155–167.

Teki, S., Kumar, S., & Griffiths, T. D. (2016). Large-scale analysis of auditory segregation behavior crowdsourced via a smartphone app. *PloS one, 11*(4), e0153916.

Traer, J. A., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences, 113*(48), E7856–E7865.

Woods, K. J. P., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology, 25,* 2238–2246.