An online headphone screening test based on dichotic pitch

Alice E. Milne*[1], Roberta Bianco*[1], Katarina C. Poole*[1], Sijia Zhao*[1,2], Alexander J. Billig[1] & Maria Chait[1]

*Equal contribution

[1] Ear Institute, University College London, London WC1X 8EE, UK

[2] Department of Experimental Psychology, University of Oxford, Oxford OX1 3PH, UK

Corresponding Author:

Alice Milne

a.milne@ucl.ac.uk

Ear Institute, University College London

332 Gray's Inn Road, London WC1X 8EE, UK

ABSTRACT

Online experimental platforms can be used as an alternative, or complement, to lab-based research. However, when conducting auditory experiments via online methods, the researcher has limited control over the participants' listening environment. We offer a new method to probe one aspect of that environment, headphone use. Headphones not only provide better control of sound presentation but can also "shield" the listener from background noise. Here we present a rapid (< 3 minute) headphone screening test based on Huggins Pitch (HP), a perceptual phenomenon that can only be detected when stimuli are presented dichotically. We validate this test using a cohort of "Trusted" online participants who completed the test using both headphones and loudspeakers. The same participants also trialled a widely used headphone test (AP test; Woods et al., 2017). We demonstrate that compared to the AP test, the HP test has a higher selectively for headphone users, rendering it as a compelling alternative to the existing screening method. Overall, the new HP test correctly detects 80% of headphone users and has a false positive rate of 20%. Moreover, we demonstrate that there is little overlap between participants who pass both HP and AP tests over loudspeakers. Therefore, combining the two tests can lower the false positive rate to ~7% (but at the expense of an increased false negative rate). This should be useful in situations where headphone use is particularly critical (e.g. dichotic or spatial manipulations). An implementation of the new test is available with JavaScript and through Gorilla (gorilla.sc).

Introduction

Online experimental platforms are increasingly used as an alternative, or complement, to in-lab work (Assaneo et al., 2019; Kell et al., 2018; Lavan, Knight, & McGettigan, 2019; Lavan, Knight, Hazan, et al., 2019; McPherson & McDermott, 2018; Slote & Strand, 2016; K. J. P. Woods & McDermott, 2018; Zhao et al., 2019). This process has been hastened in recent months by the COVID-19 pandemic.

A key challenge for those using online methods is maintaining data quality despite variability in participants' equipment and environment. Others have discussed this point at length, demonstrating that with appropriate motivation, exclusion criteria, and careful design, online experiments can not only produce high quality data in a short time, but also provide access to a more diverse subject pool than commonly used in lab-based investigations (Clifford & Jerit, 2014; Rodd, 2019; Woods et al., 2015).

A major leap of faith in moving experiments online involves relinquishing some of the control we normally have over participants' testing equipment. In studies that involve auditory stimuli, this means no command over remote participants' audio-delivery devices and listening environment. However, certain information can be gleaned through specially designed screening tests. Here we focus on procedures for determining whether participants are wearing headphones or instead listening through loudspeakers (note for brevity, we refer to both over-ear headphones and in-ear earphones as "headphones"). In many auditory experiments the use of headphones is preferred because they offer better control of sound presentation and "shield" the listener from other sound in their environment. However, how can we make sure that unknown online participants are indeed wearing them? This issue was addressed by Woods and colleagues (2017) who developed a now widely used test based on dichotic presentation, under the premise that participants listening over headphones, but not those listening over loudspeakers, will be able to correctly detect an acoustic target in an intensity-discrimination task (Figure 1a). In each trial the listener is presented with three consecutive 200Hz sinusoidal tones, and must determine which was perceptually the softest. Two of the tones are presented diotically: 1) the ”standard” and 2) the ”target" which is presented at -6dB relative to the standard. The third tone (a "foil") has the same amplitude as the standard but is presented dichotically such that the left and right signals have opposite polarity (anti-phase, 180°).

Woods and colleagues reasoned that over headphones, the standard and foil should have the same loudness, making the target clearly distinguishable as softer. In contrast, over loudspeakers the left and right signals interact destructively before reaching the listener's ears resulting in a weaker acoustic signal at both ears, and thus a weaker loudness sensation for the foil than the target, causing participants to respond incorrectly.

Woods et al. (2017) validated their test both in the lab and online. The test contained 6 trials and the threshold for passing the headphone screen was set at 5/6 correct responses. Using this threshold, in the lab 100% of participants wearing headphones passed the test, while only 18% passed it when using loudspeakers. In a large number of subjects recruited online (via Amazon Mechanical Turk), only 65% passed the test, suggesting that a third of the online listeners may have actually used loudspeakers, despite instructions to use headphones. The ability to detect those cases makes this test a valuable resource.

However, there are potential limitations with this test: Firstly, the destructive interaction over loudspeakers depends on the specific positions of the loudspeakers relative to the listener and may not generalize to all listening environments. Likewise, as Woods et al (2017) also point out, since the effect depends on stereo sound it will not be present in systems that broadcast a single channel or output the combined audio. This may allow some participants to pass the test even when listening over loudspeakers. Secondly, the antiphase foil stimulus causes inter-aural interactions that give rise to a particular binaural percept that is not present for the other two tones, and which may be more salient over headphones. To solve the loudness discrimination task, participants must thus ignore the binaural percept and focus on the loudness dimension. This introduces an element of confusion which might cause reduced performance among true headphone users. Here we present a different method for headphone screening that addresses these problems.

Specifically, we examine the efficacy of a headphone screening test based on a particular dichotic percept – Huggins Pitch – that should be audible via headphones but absent when listening over loudspeakers. The Huggins Pitch (HP) stimulus (Chait et al., 2006; Cramer & Huggins, 1958; Figure 1b)) is a type of illusory pitch phenomenon generated by presenting a white noise stimulus to one ear, and the same white noise—but with a phase shift of 180° over a narrow frequency band—to the other ear. This results in the perception of a faint tonal object (corresponding in pitch to the centre frequency of the phase-shifted band), embedded in noise. Importantly, the input to either ear *alone* lacks any spectral or temporal cues to pitch. The percept is only present when the two signals are dichotically combined over headphones, implicating a central mechanism that receives the inputs from the two ears, computes their invariance and differences, and translates these into a tonal object. Due to acoustic mixing effects, it is weak or absent when the stimuli are presented over loudspeakers.

Similarly to Woods et al. (2017) we created a 3AFC paradigm. Two intervals contain diotic white noise and the third ("target") contains the dichotic stimulus that evokes the HP percept. Participants are required to indicate the interval that contains the hidden tone. This paradigm is also attractive because it is based on a detection task (as opposed to the discrimination task in Woods et al. 2017) and, therefore, does not impose such a load on working memory or other cognitive resources. As a result, it should provide a cleaner screening tool.

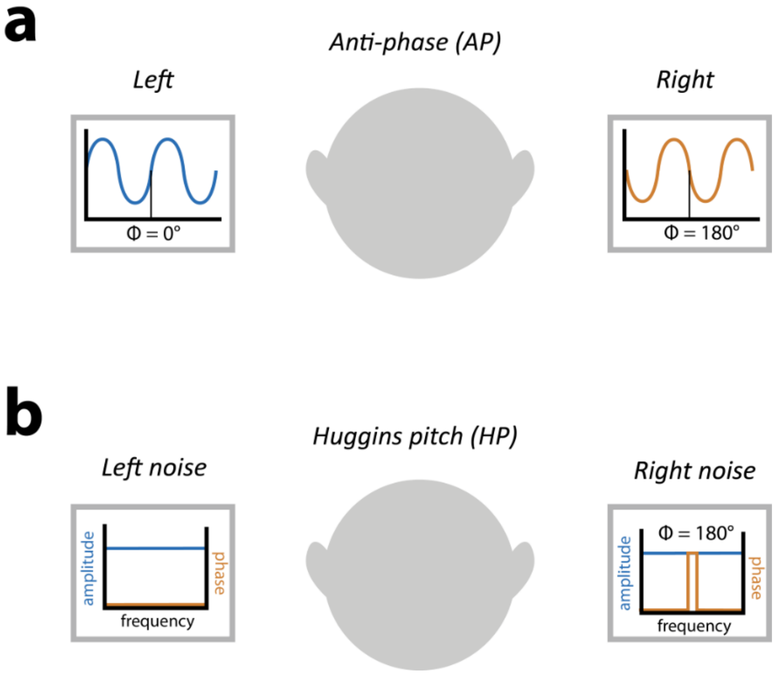

***Figure 1**: **Schematic of stimuli for each test**. (a) In the AP test a foil is created by presenting a 200Hz tone* dichotically *in anti-phase. When presented over loudspeakers this is expected to result in destructive acoustic interference, and thus reduced loudness, causing the foil to be mistaken for the target and the listener to fail the test. Over headphones there is no such interference, and thus no reduction of loudness and the listener should correctly detect the target, passing the test. However, the test is susceptible to certain loudspeaker configurations and the presence of binaural interaction which may reduce the effectiveness of the test (see text). (b) In the HP test, broadband noise is presented to one channel and the same broadband noise with a phase shift (anti-phase) over a narrow band (±6%) around 600Hz is presented to the other channel. Over headphones this results in a percept of pitch at this frequency that the listener will detect, allowing them to pass the screening. The percept depends on the left and right channels being independent, and thus tends to disappear over loudspeakers, preventing the listener from detecting the target and thus causing them to fail the test.*

To determine which test is more sensitive to headphone versus loudspeaker use, we sought to directly compare the two approaches (from here on we refer to the Woods et al. (2017) paradigm as the *Anti-Phase* (AP) test and the new paradigm as the *Huggins Pitch* (HP) test). In Experiment 1 we used the Gorilla online platform (Anwyl-Irvine et al., 2020) to obtain performance from 100 "Trusted" participants (colleagues from across the auditory research community), who completed both tests over loudspeakers and headphones. Importantly, each participant used their own computer setup and audio equipment, resulting in variability that is analogous to that expected for experiments conducted online (links to demos of each test are available in Materials and methods). In Experiment 2 we further tested the AP and HP screens using anonymous online participants. This group only used headphones to complete each test and we evaluated their performance using the profile of results we would expect for headphone and speaker use, based on Experiment 1.

Our results reveal that the HP test has a better diagnostic ability to classify between headphones and loudspeakers. Furthermore, we show that there is little overlap between the participants who pass the HP and AP tests over loudspeakers, suggesting that a combination of the two tests can provide a powerful tool for detecting non-headphone users.

Experiment 1 – “Trusted” participant group

Materials and methods

Participants

114 "Trusted" participants were tested. 14 of these were unable to hear an example target during the task explanation (see Procedure for more details) resulting in early termination. We report the results of the remaining 100 participants. Recruitment was conducted via email, inviting anyone who was over 18 and without known hearing problems to participate. The email was distributed to people we believed could be trusted to switch between headphones and loudspeakers when instructed to do so (e.g. mailing lists of colleagues/the scientific community). Participants were only informed of the general nature of the test which was to "assess remote participants' listening environments”, with no reference to specific stimulus manipulations. Individual ages were not collected to help protect the anonymity of the participants. Grouped ages are presented in Table 1. Experimental procedures were approved by the research ethics committee of University College London [Project ID Number: 14837/001] and informed consent was obtained from each participant.

Table 1. *Self-reported participant age range in Experiment 1 ("Trusted" group) and Experiment 2 ("Unknown" group, see below).*

Age Bracket (% cases)	"Trusted" group (Experiment 1)	"Unknown" group (Experiment 2)
18-25	24	72
26-35	36	22
36-45	22	4
46-55	13	2
56-65	3	-
Over 65	2	-

Stimuli and Procedure

Gorilla Experiment Builder (www.gorilla.sc) was used to create and host the experiment online (Anwyl-Irvine et al., 2020). Participants were informed prior to starting the test that they would need access to both loudspeakers (external or internal) and headphones. The main test consisted of four blocks. Two blocks were based on the Huggins Pitch (HP) test and two on the Anti-Phase (AP) test (Woods et al., 2017). Both HP and AP used a 3-alternative-forced-choice (3AFC) paradigm (Figure 2). At the start of each block, participants were told whether to use loudspeakers or to wear headphones for that block. The blocks were presented in a random order using a Latin square design. In total the study (including instructions) lasted ~10 minutes with each block (HP_headphone, HP_loudspeaker, AP_headphone, AP_loudspeaker) taking 1.5-2.5 minutes.

Volume Calibration

Every block began with a volume calibration to make sure that stimuli were presented at an appropriate level. For HP blocks a white noise stimulus was used; for AP blocks a 200 Hz tone was used. Participants were instructed to adjust the volume to as high a level as possible without it being uncomfortable.

HP Screening

The HP stimuli consisted of three intervals of white noise, each 1000 ms long. Two of the intervals contained diotically presented white noise (Figure 2). The third interval contained the HP stimulus. A centre frequency of 600 Hz was used (roughly in the middle of the frequency region where HP is salient). The signals were created by choosing Gaussian distributed numbers (sampling frequency 44.1 kHz, bandwidth 22.05 kHz). The HP signals were generated by introducing a constant phase shift of 180°in a frequency band (± 6%) surrounding 600 Hz within the noise sample delivered to the right ear, while the original sample was delivered to the left ear (Yost et al., 1987). Overall, 12 trials were pre-generated offline (each with different noise segments; the position of the target uniformly distributed). For each participant, in each block (HP loudspeaker / HP headphones) 6 trials were randomly drawn from the pool without replacement.

The participant was told that they will "hear three white noise sounds with silent gaps in-between. One of the noises has a faint tone within." They were then asked to decide which of the three noises contained the tone by clicking on the appropriate button (1, 2, or 3).

AP Screening

The AP stimuli were the same as in Wood et al. (2017). They consisted of three 200Hz tones (1000 ms duration and 100 ms ramp). Two of the tones were presented diotically: 1) the "standard", and 2) the "target" which was the same tone at -6dB relative to the standard. The third tone (the "foil") had the same amplitude as the standard but was presented such that the left and right signals were in anti-phase (180°) (Figure 2). Listeners were instructed that "Three tones in succession will be played, please select the tone (1, 2, or 3) that you thought was the quietest". As in the HP screening, for each participant, in each block (AP loudspeaker / AP headphones) 6 trials were randomly drawn from a pre-generated set of 12 trials.

Each screening test began with an example to familiarise the participants with the sound. The target in the example did not rely on dichotic processing but was simulated to sound the same as the target regardless of delivery device (for HP this was a pure tone embedded in noise; for AP two equal amplitude standards and a softer target were presented). Failure to hear the target in the example resulted in the participant being excluded from the experiment.

Following the example, each block consisted of 6 trials. No feedback was provided, and each trial began automatically.

The HP and AP test implementation can be accessed via Gorilla https://gorilla.sc/openmaterials/100917, or in JavaScript https://sijiazhao.github.io/headphonecheck/. The AP code implementation from Woods et al. (2017) can be accessed via http://mcdermottlab.mit.edu/downloads.html.

Versions of both tests can be previewed in a web browser using these URLs:
HP: https://sijiazhao.github.io/headphonecheck/headphonesTestHugginsPitch.html
AP: https://sijiazhao.github.io/headphonecheck/headphonesTestAntiPhase.html

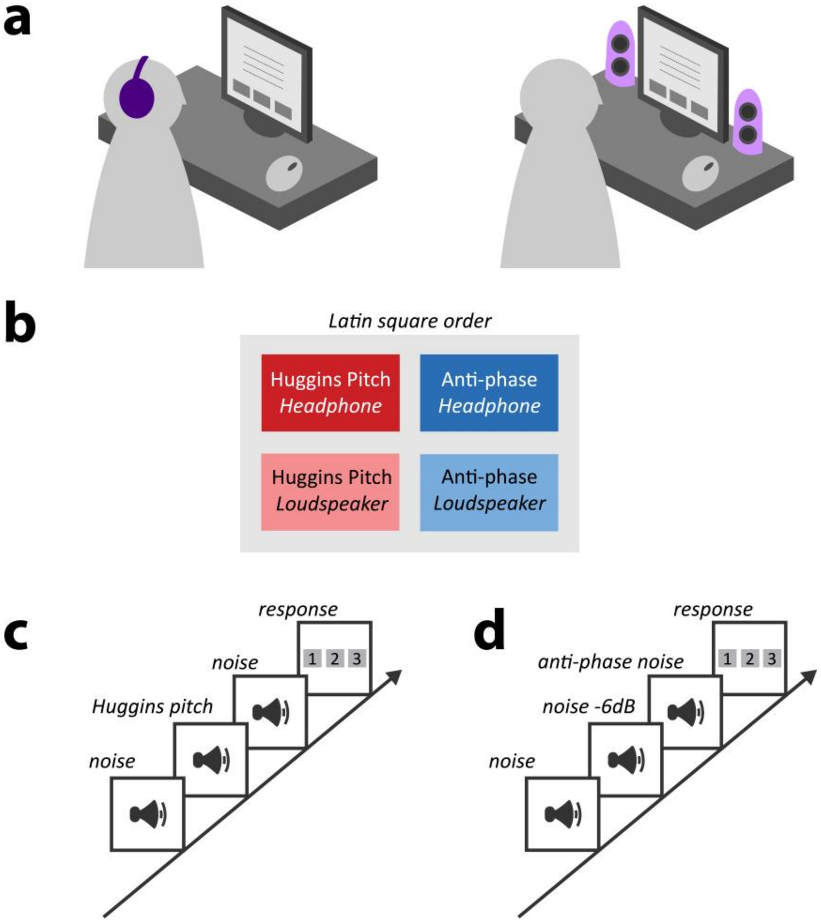

Figure 2: Schematic of the test design. *(a) At the beginning of each block, participants were informed whether the upcoming test was to be performed whilst wearing headphones or via loudspeakers. Participants responded using a graphic user interface and computer mouse. (b) The experiment was organised into four testing blocks: Huggins pitch (HP) test over headphones, HP test over loudspeakers, Anti-phase (AP) test over headphones, and AP test over loudspeakers. Test order was randomised across participants using a Latin square design. Both HP (c) and AP (d) tests used a 3AFC paradigm. For both tests, the example shows the target in the second position.*

Statistical Analysis

We used signal detection theory to ask how well the two test types (HP and AP) distinguished whether participants were using headphones or loudspeakers. Accepting a user (i.e. deciding that they passed the test) at a given threshold (minimum number of correct trials) when they were using headphones was considered a "hit", while passing that user at the same threshold when they were using loudspeakers was considered a "false alarm". We used these quantities to derive a receiver operating characteristic (ROC; Swets, 1986) for each test type, enabling a comparison in terms of their ability to distinguish headphone versus loudspeaker use. As well as calculating the area under the ROC curve (AUC) as an overall sensitivity measure, we also report the sensitivity (*d*') of the HP and AP tests at each of the thresholds separately. Note that "hits", "false alarms", and "sensitivity" here are properties of our tests (HP and AP) to detect equipment, not of the subjects taking those tests.

On the basis that a subject's performance above chance should be a minimum requirement for them to be accepted under any selection strategy, we considered only thresholds (number of correct responses required to pass) of 3, 4, 5, and 6 trials out of 6. This approach also side-stepped the issue that the AP test over loudspeakers can result in below-chance performance, as evident in Figure 3 (light blue line does not show a chance distribution).

We additionally considered whether a combined test that made use of responses both to HP and AP trials would be more sensitive than either condition alone. Under this **"Both"** approach, subjects passed only if they met the threshold both for HP and AP trials.

We assessed statistical significance of differences in sensitivity (AUC) in two ways. First, we determined reliability of the results through bootstrapped resampling over subjects. For each of 1,000,000 resamplings we randomly selected 100 subjects with replacement from the pool of 100 subjects (balanced) and obtained a distribution of differences in the AUC for HP versus AP tests. We then determined the proportion of resamples for which the difference exceeded zero (separately for each direction of difference, i.e. HP minus AP, then AP minus HP), and accepted the result as significant if this was greater than 97.5% in either direction (two tailed; $p < 0.05$). The other method we used to assess statistical significance of differences of interest was with respect to a null distribution obtained through relabelling and permutation testing. For each of 1,000,000 permutations we randomly relabelled the two headphone condition scores for each of the 100 subjects as HP or AP, and similarly for the two loudspeaker scores. We then calculated the AUC at each threshold for these permuted values. This generated a null distribution of AUC differences that would be expected by chance. We then determined the proportion of scores in these null distributions that exceeded the observed difference in either direction and accepted the result as significant if this was less than 2.5% in either direction (two tailed; $p < 0.05$). Identical procedures were used to test for differences between the "Both" approach and each of the HP and AP methods.

Results

Distribution of performance for each screening test

Figure 3 presents a distribution of performance across participants and test conditions. The x axis shows performance (ranging from a perfect score of 6/6 to 0/6). Chance performance (dashed black line) is at 2. The performance on the AP test with headphones (dark blue line) generally mirrored that reported in Woods et al. (2017), except that the pass rate in the headphones condition (70%) is substantially lower than in their controlled lab setting data (100%). This is likely due to the fact that the "Trusted" participants in the present experiment completed the test online, thereby introducing variability associated with specific computer/auditory equipment. Performance on the AP test with loudspeakers (light blue line) was also similar to that expected based on Woods et al. (2017). Some participants succeeded in the test over loudspeakers (30% at 6/6). Notably, and similarly to what was observed in Woods et al. (2017), the plot does not exhibit a peak near 2, as would be expected by chance performance in a 3AFC task, but instead a *trough,* consistent with participants mistaking the phase shifted “foil” for the “target”. For the HP test, a chance distribution is clearly observed in the loudspeaker data (peak at 2, light red line). There is an additional peak at 6, suggesting that some participants (20% at 6/6) can detect Huggins Pitch over loudspeakers. In contrast, performance using headphones for HP (dark red line) shows an “all-or-nothing” pattern with low numbers for performance levels below 6/6, consistent with HP being a robust percept over headphones.

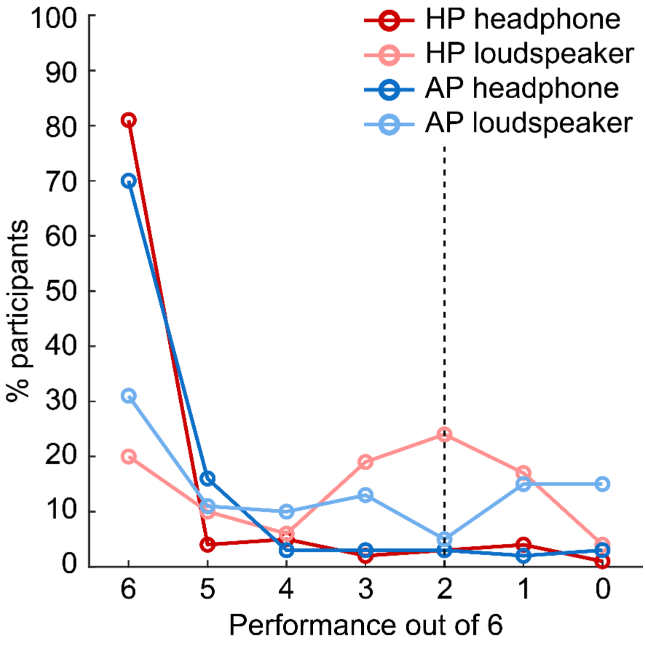

***Figure 3**: **Distribution of performance for each test condition in the “Trusted” group (Experiment 1; N=100)**. The x axis shows performance (ranging from a perfect score of 6/6 to 0/6 trials). The dashed black line indicates chance performance.*

Ability of each screening test to distinguish between headphone vs. loudspeaker use

We derived the receiver operating characteristic (ROC) for each test, plotting the percentage of participants who passed at each above-chance threshold while using headphones (“hits”, y-axis) or loudspeakers (“false alarms”, x-axis) (Figure 4a). The area under the curve (AUC) provides a measure of how well each test type distinguishes between headphone versus loudspeaker use. The AUC for HP (.821) was significantly larger than that for AP (.736) (bootstrap resampling: p = .022, permutation test: p= .018). This suggested that the HP test overall provides a better balance of hits to false positives (i.e., maximizing the headphones pass rate, while reducing the proportion of listeners who pass using loudspeakers). This is also illustrated in Figure 4b which plots d’ at each threshold. The maximum d’ reached is ~1.7, consistent with medium sensitivity at the highest threshold (6/6). At this threshold HP will correctly detect 81% of the true headphone users, but also pass 20% of loudspeaker users, AP will detect 70% of the headphone users, but also pass 31% of loudspeaker users; for threshold of 5/6 the values are 85%/30% for HP and 86%/42% for AP.

We also plotted the ROC and sensitivity for a "Both” approach that required participants to reach the threshold both for the HP and AP tests. The AUC for Both was .844 and significantly higher than for AP (bootstrap resampling p < .001, permutation test: p = .014) but not for HP (bootstrap resampling: p = .279, permutation test: p = .979). Given the additional time that would be required compared to running HP alone, the lack of significant difference over HP suggests that the combined test is not generally a worthwhile screening approach. However, if the experiment is such that headphone use is critical then using the combined test will reduce the loudspeaker pass rate from 20% to 7% but at the expense of rejecting 40% of true headphone users. This is illustrated in Figure 4c, which plots the proportion of listeners who pass the AP and HP tests over loudspeakers (relative to the number of subjects who pass at least one test over loudspeakers). For each threshold, the proportion of listeners who pass the AP test over loudspeakers is larger than that for HP (Figure 4c). The proportion of listeners who pass both loudspeaker tests is very low, consistent with the fact that the conditions that promote passing the HP test over loudspeakers (listeners close to and exactly between the loudspeakers such that left and right ears receive primarily the left and right channels, respectively) are antithetical to those that yield better AP performance. Therefore, combining the two tests will substantially reduce the number of people who pass using loudspeakers. In contrast to the performance with loudspeakers, most participants passed both the HP and AP tests when using headphones (Figure 4d).

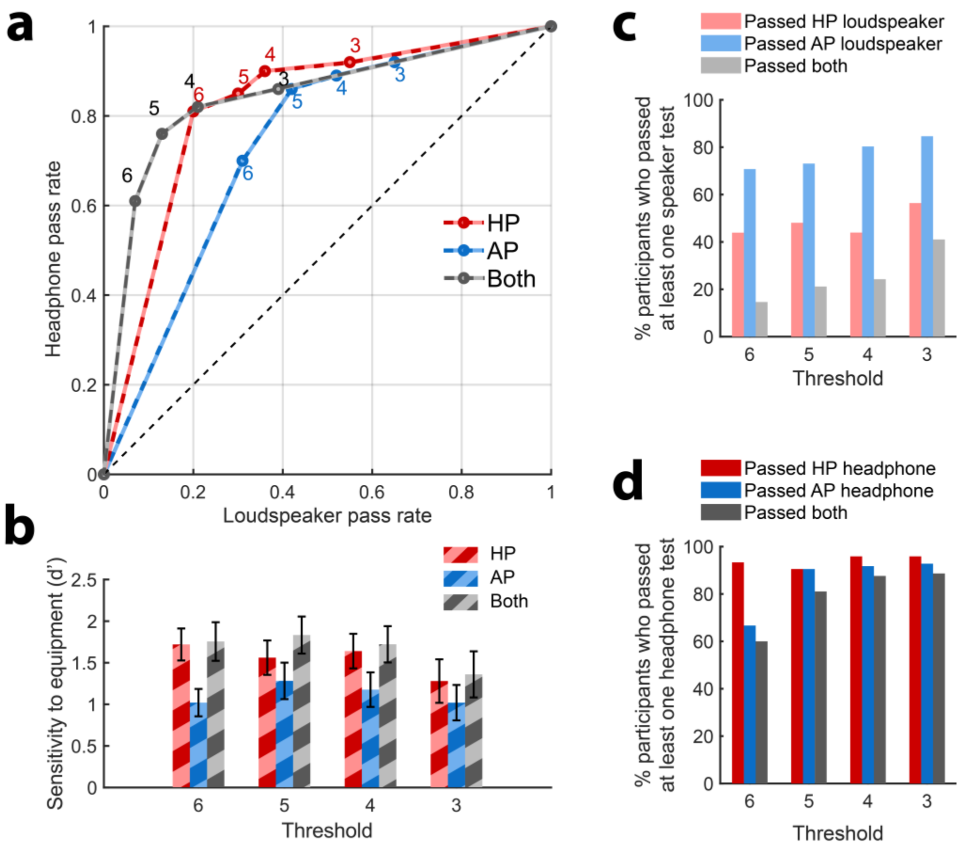

***Figure 4**: **Ability of HP, AP and a combined test ("Both") to distinguish between headphone and loudspeaker users (N = 100).** (a) ROC curves. The proportion of participants passing at each above-chance threshold (3,4,5,6 /6 labelled next to each data point) while using headphones ("hits", y-axis) or loudspeakers ("false alarms", x-axis) for HP, AP or a combined test ("Both"). (b) Sensitivity (d') at each threshold. Error bar = 1 std bootstrap with 10,000 iterations. (c) Pass rates with loudspeakers at each threshold, plotted relative to the total number of participants who passed at least one of the loudspeaker tests. (d) Pass rates with headphones. Whilst a large proportion of participants pass both AP and HP tests over headphones (dark grey bars in d), only a small proportion pass both tests over loudspeakers (light grey bars in c).*

Experiment 2 - “Unknown” online group

We probed performance on the AP and HP tests in a typical online population. This time, participants were unknown to us, recruited anonymously and paid for their time. We informed participants that headphones had to be worn for this study and sought to determine whether the pass rate would be similar to that in the "Trusted" cohort.

Participants

We recruited online participants via the Prolific recruitment platform (prolific.co). 103 participants were tested, three were unable to hear one of the example sounds and left the study early, leaving a total of 100 participants. Participants were paid to complete the 5-7-minute study. We specified that they should not accept the study if they had any known hearing problems. Additional exclusion criteria were not applied to this sample in order to obtain a broad range of participants. Ages are provided in Table 1 (right panel). Experimental procedures were approved by the research ethics committee of University College London [Project ID Number: 14837/001] and informed consent was obtained from each participant.

Stimuli and procedure

Stimuli and procedure were the same as in Experiment 1 except participants only used headphones, thus completing each screening test, HP and AP, once. The instructions stressed that headphones must be worn for this experiment.

Results

Figure 5 plots the performance (black lines) observed for the “Unknown” online group. Overall, the performance patterns were different from the performance using headphones obtained from the “Trusted” group, suggesting that a proportion of listeners may not have heeded the instructions to use headphones, or used low quality equipment. In particular, there was a ~10% greater number of participants getting 6/6 with the AP vs. HP test, which is the reverse of what was seen in the “Trusted” group with headphones. This adds support to the results of Experiment 1 that suggest there is a higher false positive rate with the AP test.

To estimate the proportion of online participants that actually used headphones, we assumed that the distribution of online scores for each test type could be explained as a linear combination of the distributions of headphone and loudspeaker scores from the same test type in the “Trusted” group

(Experiment 1). We used a simple model with a single parameter, *propH*, for the proportion of headphone users. For values of *propH* varying from 0 to 1 in increments of .01 we multiplied the distribution of Experiment 1 HP headphone scores by *propH*, and summed these two values, giving a modelled distribution of Experiment 2 HP scores for each value of *propH*. We repeated the same process for AP scores. We then compared the modelled and observed distributions and selected the value of *propH* that minimised the sum of squared errors across both HP and AP scores. This analysis yielded an estimate that 40% of users in Experiment 2 likely did not use headphones (or had unsuitable equipment/completed the test in a noisy setting), demonstrating the importance of running an objective screen.

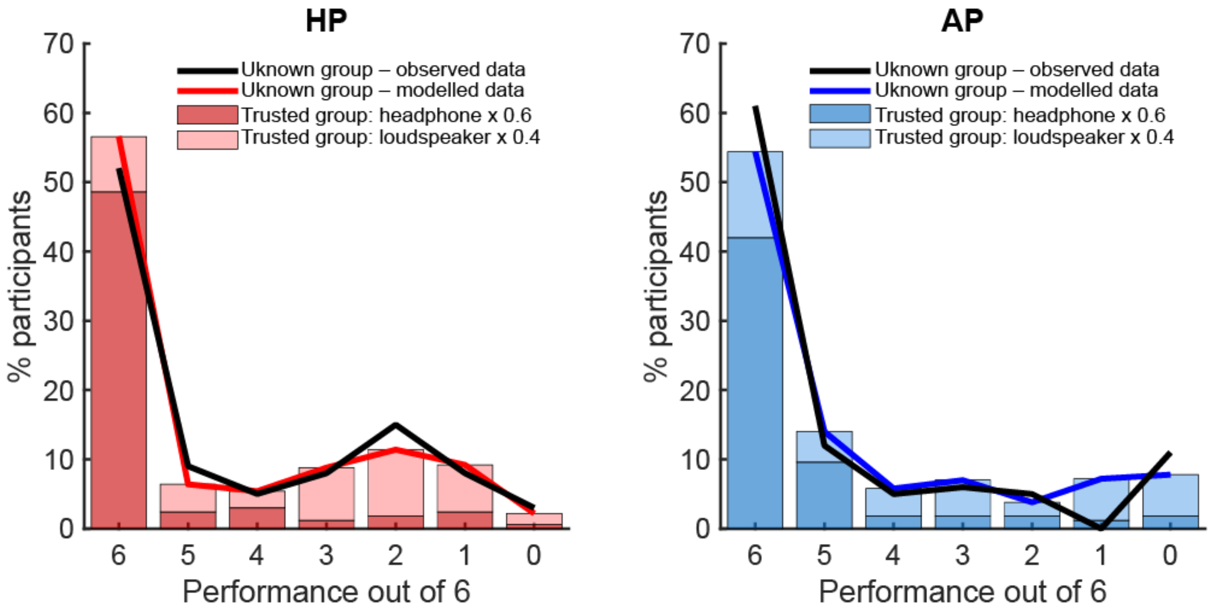

***Figure 5**. **Performance in the "Unknown" group (Experiment 2; N=100)**. The black solid line illustrates the performance observed in the "Unknown" group in the HP (left) and AP (right) tests. The red/blue lines and stacked bars illustrate the result of modelling to determine the likely proportion of subjects in the "Unknown" group who actually used headphones. The stacked bars indicate the product of the proportion of participants in the "Trusted" group at each performance level with the relevant coefficient from the best fitting model (0.6 for headphones and 0.4 for loudspeakers, fixed across HP and AP). The distribution of observed performance in the "Unknown" group matched the modelled data well for both HP and AP tests. This indicates that only roughly 60% of participants in the "Unknown" group showed performance that was consistent with headphone use.*

Discussion

We sought to develop an efficient headphone screening test for use with online auditory experiments that is easy to explain to listeners, quick to administer (< 3 mins) and which has a high selectivity for headphone users. We devised a new test (HP) based on a perceptual phenomenon that can only be detected when stimuli are presented dichotically. This detection test was contrasted with an existing test (AP). The analyses we reported demonstrate that HP has higher selectivity for headphone users than AP, rendering it a compelling alternative to the existing screening method. That it is based on a detection rather than a discrimination task, and therefore less dependent on working memory, further adds to its appeal.

We note that all our estimates are based on the "Trusted" participant group. However, this cohort (primarily from a network of colleagues and our scientific community) may not be fully representative of the general online participant population. For instance, it is conceivable that they possess higher quality equipment or were more motivated than the average online participant. In general, it is prudent to treat the "Trusted" group data as reflecting the best-case scenario with actual performance probably somewhat lower in the general population. Importantly, the test is designed to distinguish between participants who are using stereo headphones (i.e. where the left and right channels are independently delivered to the left and right ear, respectively) from those listening without headphones (where typically the left and right channels will interact in some way before reaching the listeners' ears). Though the screen is not designed to be sensitive to other aspects of the listener's environment per se, headphone users may nonetheless fail the test if the quality of the equipment is very low or if their environment is particularly noisy.

Overall, we conclude that the HP test is a powerful tool to screen for headphone use in online experiments. We have made our implementation openly available and ready for use via JavaScript and Gorilla (gorilla.sc). The test consists of 6 trials and, based on the ROC analysis, our recommendation is to use a threshold of 6/6. Lower thresholds will result in a similar *d'* but will pass a larger proportion of loudspeaker users.

It must be stressed that the test is not perfect, in that it passes only 80% of genuine headphone users and fails to reject 20% of loudspeaker users. Failing the test over headphones could be attributable to poor quality equipment (e.g. left and right channels not kept separate), background noise, or hearing loss. Conversely, those subjects who pass with loudspeakers might be optimally spatially positioned (e.g. equally between the two loudspeakers for HP). In situations where it is important to reach a high level of certainty that the participant is using headphones (e.g., where stimuli involve a dichotic presentation or a spatial manipulation) the HP and AP tests can be combined. This will yield a false positive rate of ~7% but at the expense of rejecting 40% of true headphone users.

Overall, the rapid test we have validated here can effectively aid researchers in confirming the listening environment of their participants thereby reaping the benefits of using online experimental platforms whilst controlling (at least certain aspects of) data quality.

Funding

This work was funded by BBSRC grant [BB/P003745/1] to MC, a Wellcome Trust grant [213686/Z/18/Z] to AEM and supported by NIHR UCLH BRC Deafness and Hearing Problems Theme. AJB is supported by Wellcome Trust grant [091681]. KCP is supported by an ERC grant [771550-SOUNDSCENE].

Acknowledgments

We are grateful to Gordon Mills and Fred Dick for comments and advice and to our colleagues across the international auditory research community for participating in the online experiments and providing feedback.

Competing Interests

The authors declare no competing interests.

Data and Code Availability

All data and the analysis code to reproduce the figures are available at [https://github.com/sijiazhao/headphonecheck_analysis].

Code implementing both headphone screening tests in JavaScript can be downloaded from the project website [https://sijiazhao.github.io/headphonecheck/].

Gorilla (gorilla.sc) versions of both screening tests are available through Gorilla Open Material https://gorilla.sc/openmaterials/100917.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01237-x

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience*. https://doi.org/10.1038/s41593-019-0353-z

Chait, M., Poeppel, D., & Simon, J. Z. (2006). Neural response correlates of detection of monaurally and binaurally created pitches in humans. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhj027

Clifford, S., & Jerit, J. (2014). Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory. *Cambridge.Org*. https://doi.org/10.1017/xps.2014.5

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

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. v., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*. https://doi.org/10.1016/j.neuron.2018.03.044

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*. https://doi.org/10.1016/j.cognition.2019.104026

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*. https://doi.org/10.1038/s41467-019-10295-w

McPherson, M. J., & McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-017-0261-8

Rodd, J. (2019). How to Maintain Data Quality When You Can't See Your Participants. *APS Observer*, *32*(3). https://www.psychologicalscience.org/observer/how-to-maintain-%09data-quality-when-you-cant-see-your-participants

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*. https://doi.org/10.3758/s13428-015-0599-7

Swets, J. A. (1986). Indices of Discrimination or Diagnostic Accuracy. Their ROCs and Implied Models. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.99.1.100

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. In *PeerJ*. https://doi.org/10.7717/peerj.1058

Woods, K. J. P., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1801614115

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

Yost, W. A., Harder, P. J., & Dye, R. H. (1987). Complex spectral patterns with interaural differences: dichotic pitch and the ‘central spectrum’. In: Auditory processing of complex sounds No Title. In *Auditory processing of complex sounds* (pp. 190–201).

Zhao, S., Yum, N. W., Benjamin, L., Benhamou, E., Yoneya, M., Furukawa, S., Dick, F., Slaney, M., & Chait, M. (2019). Rapid Ocular Responses Are Modulated by Bottom-up-Driven Auditory Salience. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0776-19.2019