Supplemental Materials

Table S1. Questions from CHAT-AH. *Bolded questions were used to assign participants to the AH+ group.*

CHAT-AH Questions
1. Have you ever thought you heard someone call your name, but then realized you must have been mistaken?
2. Have you ever heard your phone ringing, but then realized the phone hadn't actually rung?
3. Do you ever hear strange noises when you are falling asleep or waking up in the morning?
4. What about hearing music or other noises that other people around you did not seem to hear?
5. Have you ever had an experience where you heard things, such as loud noises, voices talking, or people whispering, that other people could not hear?
6. Have you ever been told that you are hearing things that are not real or are not really there?
7. Have you ever had an auditory hallucination?
8. Has a doctor or family member ever told you that you have had an auditory hallucination?

Table S2. Participants based on hallucination status, before demographic matching.

	AH-	AH+	p
n	239	289	
Age (mean (SD))	37.90 (10.80)	38.73 (13.58)	0.448
Total LSHS Score(mean (SD))	5.60 (5.94)	16.39 (9.14)	**<0.001**
Total PDI Score(mean (SD))	1.78 (2.55)	6.54 (4.08)	**<0.001**
Self-Reported Mental Illness n(%)	19 (9.4)	99 (35.0)	**<0.001**
Race n(%)			0.223
American Indian/Alaskan Native	5 (2.1)	2 (0.7)	
Asian	22 (9.2)	30 (10.4)	
Native Hawaiian or Other Pacific Islander	1 (0.4)	2 (0.7)	
Black or African American	6 (2.5)	13 (4.5)	
White	190 (79.5)	215 (74.4)	
More than one race	8 (3.3)	21 (7.3)	
Unknown/Prefer not to say	7 (2.9)	6 (2.1)	
Sex F n(%)	121 (50.6)	206 (71.3)	**<0.001**
Current Medication Use n(%)	10 (4.2)	68 (23.5)	**<0.001**
Self Report, Psychosis Spectrum Illness n(%)	1 (0.4)	32 (11.1)	**<0.001**
Total Raven Score (out of 9) (mean (SD))	6.58 (1.70)	6.02 (1.78)	**<0.001**

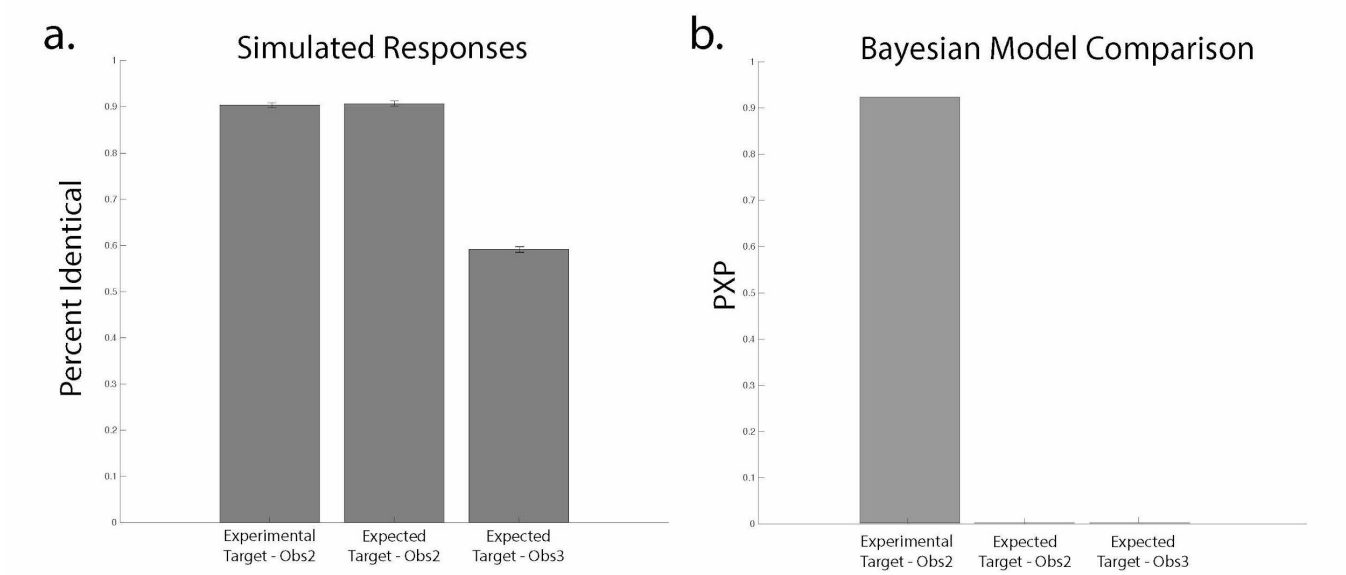

Figure S1. Model comparison for three HGF iterations. *Three versions of the HGF were tested for overall fit to the data. Two structural variants of the HGF were tested--the Obs2 variant (see* ***Fig. 3****) previously published and the Obs3 variant, specifically meant to identify if prior weighting may be dynamically linked to volatility estimates on any given trial. Stimulus strength data used for fitting were either the expected value or the empirically-determined mean response values for condition. Fit quality was determined by the number of identical responses produced by data simulation and model inversion (****a****) and Bayesian model selection (****b).*** *The Obs2 model using empirically-derived grand mean responses performed best on both metrics.*

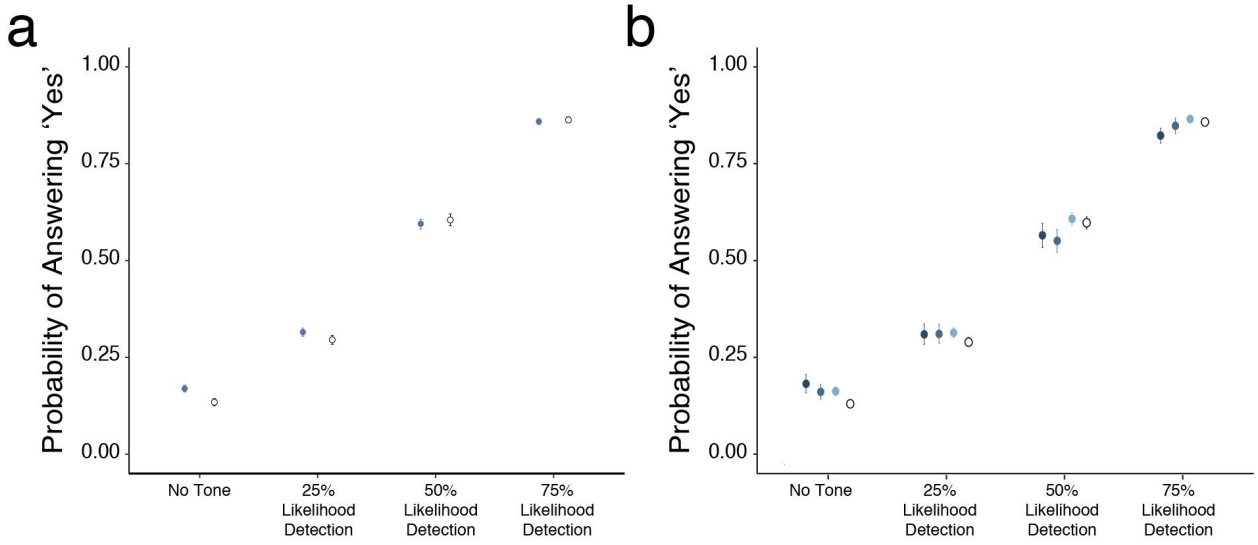

Figure S2. Mean scores of groups on each task condition. a. *AH+ (blue) and AH- group did not differ on any condition except the No Tone Condition (statistics as reported in Figure 2).* ***b.*** *Similarly, frequency groups (Daily, Weekly, Monthly or Less, Never, left to right, colors as noted in* ***Fig. 2****) did not differ on conditions other than the No-Tone condition.*

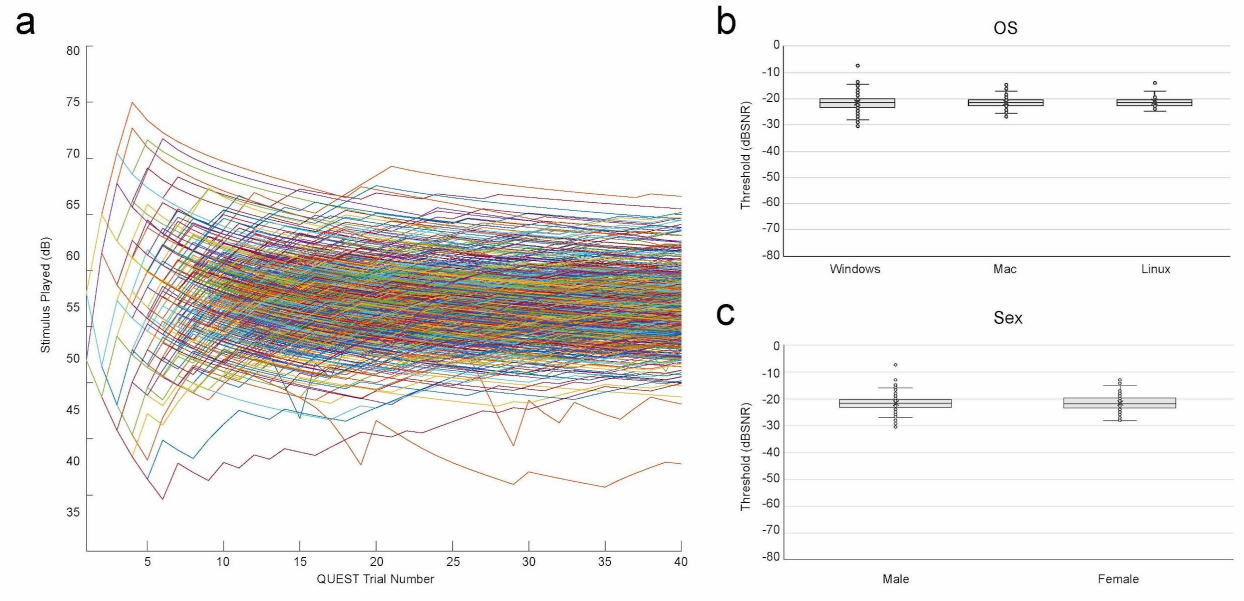

Figure S3. QUEST Performance. a. *QUEST converged appropriately on threshold values according to participant responses. Threshold was not affected by operating system (**b**) nor participant sex (**c**).*

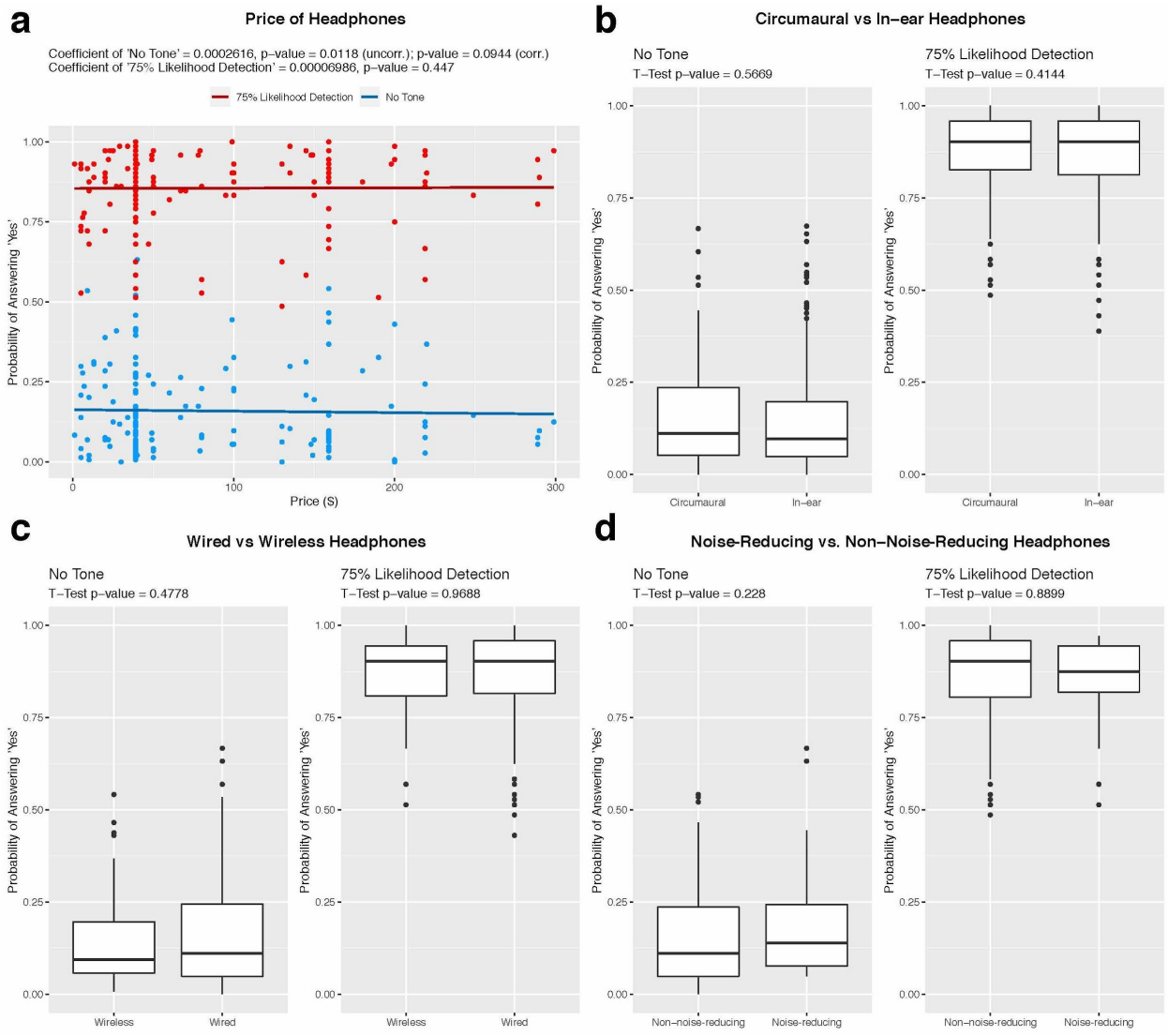

Figure S4. Participant responses and audio hardware. *Probability of reporting tone detection at the No-Tone and 75% Likelihood of Detection conditions did not differ by estimated headphone price (**a**), structure (**b**), communication type (**c**), or presence or absence of noise-reducing functionality (**d**).*

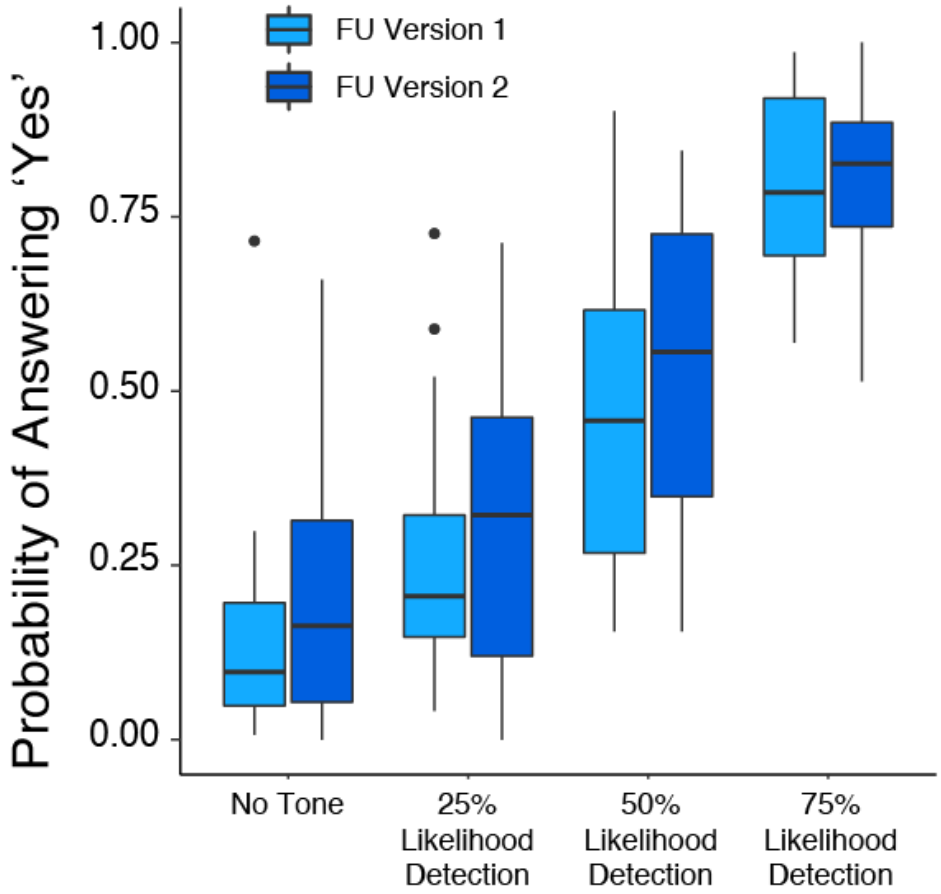

Figure S5. *No significant differences in likelihood of reporting the target tone existed between CH task versions.*

Supplemental Methods

Quality Control - Clinical and Demographic Data

Quality and accuracy of demographic and clinical data were accomplished via a series of automated checks. Participants were de-facto excluded for reported age over 65 (N=3) having a seizure-related disorder, and having a neurological disorder that would affect their cognitive abilities. Participants were excluded for inability to prove unique identity, which was detected via a combination of internet-protocol (IP) address tracking and short-message service (SMS)-based two-factor authentication (N=20).

Internal consistency of symptom endorsement and low likelihood of malingering were ensured using M-FAST consistency and malingering flags [65,66] as well as malingering checks built into the cbSASH [24,67]. Participants flagged for any reason were subject to one-on-one interviews with a clinician (author BQ) to ensure distinct identity, clarity of responses, and data integrity prior to compensation and data inclusion. Of those flagged (N=162), 101 were included, and 61 failed to comply with the required interview.

Quality Control - Task Performance

Online perceptual experiments may be impacted by differences in stimulus presentation across multiple different hardware and operating system configurations. Several measures were taken to ensure these differences were minimized to the greatest degree possible.

First, to minimize the impact of internet connectivity speed on stimulus presentation and response timing, the experiment was built for browser-based stimulus presentation and

response gathering. Thus, all timing-sensitive activities were executed client-side and then communicated back to the server at experiment completion.

Second, we took the position that differences between presentation configuration systems could be minimized by controlling auditory system configuration to the greatest degree possible and then ensuring that these configurations were operable using participant behavior. Participants were instructed to keep their screen brightness and system volumes at maximum levels throughout the experiments and to wear headphones. To ensure this, the participants were required to complete two qualifying tasks sensitive to these parameters. The qualifying tasks were created through labjs (https://lab.js.org/), hosted on the Powers laboratory server, and integrated with REDCap to link to the clinical and demographic data obtained. The auditory qualifying task asked participants to identify the quietest auditory stimulus from a sequence of three tones, and uses wave interference phenomena to make this task nearly impossible without the use of headphones [68]. The visual qualifying task was meant to ensure that participants had monitor brightness and contrast at optimal levels for performance (i.e., system maximum). This task asked participants to identify a shape that minimally differed in hue from the surrounding color, such that a high level of screen brightness was required to correctly identify the shape. Both tasks were considered successfully completed at 80% accuracy.

Third, the structure of the ACH task itself ensured target stimulus intensity would be based on participant response, thus controlling for differences in hardware that may have otherwise been confounding factors across auditory hardware configurations. Because the ACH task required individual thresholding, all stimulus strengths for the main experiment were defined by participant performance on their individual hardware configuration, again limiting heterogeneity that could contribute to differences in task performance. Initial QUEST parameters were derived

empirically from data acquired from several participants in person in the laboratory, across a number of hardware configurations and systems. Operating system and browser data were acquired automatically via the task web application, and additional details regarding the type of headphones, computer, and monitor being used were acquired by participant report. After completion of the experiment, in order to determine whether hardware differences impacted behavioral performance on the ACH task, we analyzed threshold and detection rates across operating systems and a range of hardware characteristics, including type of headphone (i.e., circumaural vs. in-ear) and price point. Analysis of threshold and reported detection at the 75% condition did not differ across any of the hardware or software configurations tested (**Fig. S3**).

Lastly, we ensured quality of data after participant completion, using participant behavior itself. If participants are able to hear the target tone, understand the task instructions, and are attending to its performance, all behavior should conform to certain patterns: 1) detection rates at the threshold condition should be greater than chance; and 2) detection rates should increase as stimulus intensities increase. Failure to exhibit these features could be caused by poor thresholding, hardware malfunction, or participant inattention. We administered quality control tests to the data collected, ensuring that participants: 1) detected the estimated threshold (75%) stimulus at least 55% (greater than chance) in the first block of the main experiment; and 2) exhibited a positive relationship between stimulus intensity and likelihood of stimulus detection across the experiment by linear regression, corresponding to the relationship predicted by initial QUEST-based modeling of individual psychometric curves. These criteria were selected to ensure threshold estimates were accurate (criterion #1) and responses corresponded to stimulus strength and were not random (criterion #2). They also had the added feature of being insensitive to the main performance metric of interest (i.e., reported detection at the no-tone condition). If these criteria were not met, participants were asked to repeat the qualifying tasks

and the ACH task in their entirety (n=18). Out of the 617 total participants who completed the ACH task online and were included after identity and consistency checks, 583 participants passed first-pass criteria for successful task completion, including those who were asked to re-do the tasks after not passing the first time.

From this sample, prior to group-level analysis, reported overall detection rates outliers (criterion < Q1 - 1.5 * IQR or > Q3 + 1.5 * IQR) (N=8) and linear regression coefficient outliers (criterion > 2 SDs from mean) (N=41). During HGF analysis, participants with extremely small changes in the X_3 trajectories in the HGF model were also removed (N=6).

Hierarchical Gaussian Filter and Model Comparisons

The HGF incorporates information on incoming sensory evidence and an agent's implicit state to model an agent's understanding of uncertainty in a dynamic environment. The model applies hierarchical Bayesian computation on consecutive decision-based inputs to develop three states of the world (X1-X3). In this study, the model was used to predict the probability of detecting a tone (i.e. responding yes) on a given trial after applying sigmoid which represents decision noise:

$$P(\text{“yes”}|belief) = sigmoid(belief).$$

Here, *belief* is defined as the agent's posterior probability of a tone being present given the subject's prior expectation of a tone being presented and intensity of the tone stimulus. This is formally represented as:

$$belief = prior + [1/1(1 + v)] (observation - prior)$$

Here *observation* is based on the 25, 50, and 75% detection rates for a given trial. *Prior* reflects the strength of the association between the visual and auditory stimuli by the agent. *Observation* and *prior* are used to derive v, which is the weighting of *prior* over *observation*.

The final logistic sigmoid for the HGF model is:

$$f(x) = 1 / e^{-\beta*(V1-V0)}$$

V1 and V0 correspond to the agent choosing “yes” and “no”, respectively. β determines the shape of the sigmoid and corresponds to the likelihood of choosing yes.

We tested three implementations of the HGF, 1) using target detection rates for the 25,50,and 75 conditions, 2) with empirically-determined detection rates, and 3) a novel HGF iteration where prior weighting is directly tied to volatility estimates. All implementations were inverted to produce synthetic response data based on fitted model parameters. Because randomness is built into the model, fitted participant models were inverted 10,000 times each and mean accuracy of simulated responses calculated by comparison against observed responses. We compared the simulated responses were then compared to observed behavioral responses, and found that the 2nd iteration of the model performed better at predicting observed data (stats). This conclusion was also supported by Bayesian Model Comparison of the two model iterations (stats) (S2).

Simulated responses did not differ significantly from observed responses across any task conditions, indicating that model parameters were capable of recapitulating the behavioral data observed.