Integration of a New Sensory Skill with Vision After Less Than 3 Hours of Training

James Negen*, Lisa Wen, Lore Thaler, and Marko Nardini

Durham University, Department of Psychology

* jnegen@gmail.com or james.negen@durham.ac.uk

Acknowledgements

We would like to acknowledge support through grant ES/N01846X/1 from the UK Economic and Social Research Council. Thanks to Hannah Roome, Alysha Chelliah and Claudia Bowman for help with piloting.

Humans are highly effective at dealing with noisy, probabilistic information[1,2]. One hallmark of this is Cue Combination: combining two independent noisy sensory estimates to increase precision beyond the best single estimate[3–6]. Surprisingly, this is not achieved until 10-12 years of age[7–12] despite other multisensory skills appearing in infancy[13–17]. It is unclear if this lack of integration before 10-12 years of age is due to maturation or experience with specific cues. The "experience" account predicts that adults learning new cues would fail to combine them for many years[18]. Here we show, in contrast, that adults rapidly combine a novel audio cue akin to human echolocation[19–21] with vision. Following two hours of training to judge distance using the novel echo cue, echo-naïve subjects were more precise given both cues together versus the best single cue. This ability also transferred to stimuli and reliability levels beyond those trained, showing learning beyond a simple decision rule with specific stimuli[22]. These results indicate that humans develop general-purpose cue combination abilities as they mature. The discovery of people's ability to immediately integrate new signals into their existing repertoire further suggests an optimistic outlook on substituting or augmenting human senses.

Using the information from our senses to make the best decisions can be surprisingly complex. Consider the simple act of comparing the distances of two free supermarket checkouts. There are potentially a dozen different methods of estimating visual distance[23], such as using perspective information, using the differences in the two eyes' views of the same scene, comparing the apparent sizes of counters and other nearby objects whose sizes are familiar, and so on. From a dozen different methods we might get a dozen different estimates. Since we can only act out one decision, one singular estimate must be decided upon. To arrive at a single estimate, different techniques can be pursued, some of which might be more prone to fail. For example, picking one estimation method at random could be extremely inaccurate if an unreliable method were selected. Averaging all estimates together might also result in a poor decision if a few very bad estimates and a few very good estimates were given equal weight. Even if we somehow knew that one estimation method is most reliable and decided to use it just by itself, we would be throwing away a lot of potentially useful information. How do we best handle all of this sensory input?

One good solution is to consider not just the individual estimates, but also their error distributions. If the different methods give approximately Gaussian error, the process for forming a single unified estimate with the lowest variance (uncertainty) is to take an average that is weighted by each estimate's precision (1/variance). This can be broadly termed *Bayesian Cue Combination* since it is considered optimal under Bayesian statistics[1,22]. This strategy is used explicitly by statisticians trained in meta-analysis and by engineers developing sensing and control systems. Surprisingly, human adults' behaviour suggests that their perceptual systems also carry out Bayesian computations when making perceptual decisions using multiple sensory inputs[1,2]. In many studies, human adults using sensory cues they are familiar with behave in a way that is statistically undistinguishable from a Bayesian algorithm[3–6]. Two central markers of Bayesian cue combination are (1) reducing variance below the best single estimate, accomplishing the basic goal of coordinating the different information sources to make better judgements, and (2) flexibly re-weighting (changing

reliance on) different cues when they change in reliability[22]. For example, in a classic study[3], participants quantitatively followed Bayesian predictions in reducing variance when comparing heights of blocks given vision and touch together vs the best single cue, and relied progressively less on vision as noise was added to the visual cue.

We wanted to know if adult humans' ability to combine cues in line with a Bayesian algorithm extends beyond the use of sensory cues they are familiar with, into the realm of sensory substitution and augmentation, where people are trained to use new devices or techniques to infer information about the environment in new ways[21,24]. Addressing how novel cues are integrated will also help us to answer an open question about perceptual development: why do children only reduce uncertainty by combining even familiar cues once they are 10-12 years old[7–12]? If adults fail to combine newly-learned cues, this will suggest that adult humans' impressive abilities to make efficient perceptual decisions under uncertainty[1,2] depend on significant experience with specific cues.

Our experimental task asked participants, wearing VR headsets and headphones, to estimate how far in front of them a virtual cartoon whale was hiding (Figure 1A). 12 healthy adults participated. Our augmented sense was inspired by human echolocation, a technique of listening to reflected sound in order to perceive surrounding objects and their spatial layout[19–21]. To make learning this new skill more manageable and predictable, we stripped the echo signal down to just its delay component: a longer time between the initial sound and its "echo" means that the target is farther away (Figure 1B). We confirmed in a separate experiment that naive participants did not show any meaningful ability to use these audio cues to judge distance (see Figure S2 in supplementary materials). Therefore, the perceptual skill we trained in the main experiment was genuinely new.

We also showed subjects how to use an independent noisy visual cue, a display of 'bubbles' in which a wider point in the display meant that the whale was more likely to be there (Figure 1C). To test for cue combination, we measured whether participants reduced variable error (uncertainty) when given the two cues together versus the best one alone, and whether they reweighted cues when their reliabilities changed[22].

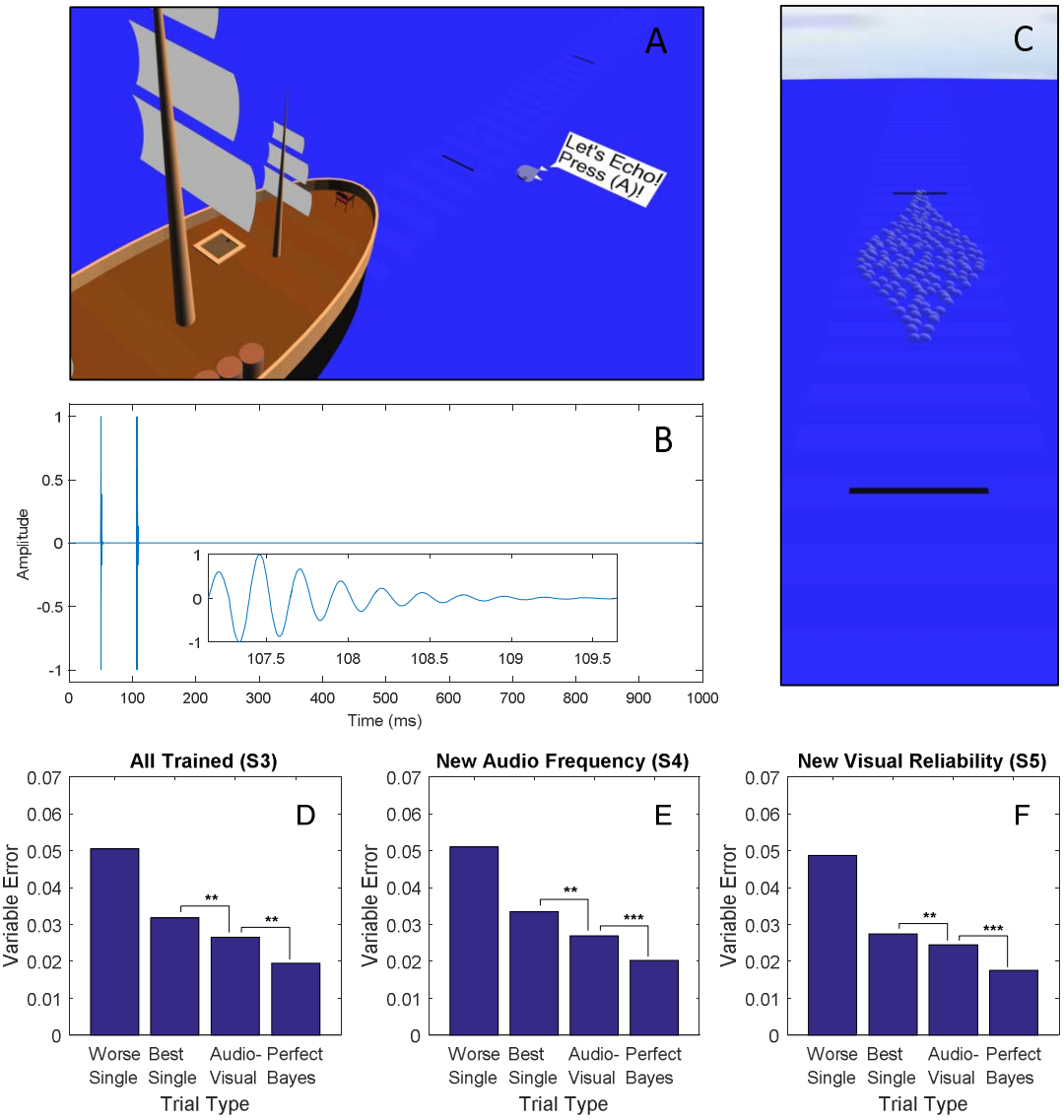

Figure 1. (A): The virtual environment (note that in the actual experiment participants saw the environment stereoscopically via a headset). Participants were asked to indicate how far along the line the whale was hiding based on audio and/or visual cues. (B): An example audio stimulus, with a zoom image of 2.5ms of the echo inlaid. In this case, the onset delay is about 57ms so the whale is about 10m away. (C): An example of the noisy visual stimulus, a display of bubbles where a wider patch meant that the whale was more likely to be there. (D-F): In each of sessions S3, S4, and S5, bimodal trials (where subjects had both cues) had lower variable error than trials with the best single cue in a two-tailed sign rank test; $*p < .05$; $**p < .01$; $***p < .001$. Variable Error is given on a log scale: 0.01 corresponds to a standard deviation of 10.5% of the subject-target distance, which translates to 1.1m at the near limit of 10m or 3.7m at 35m. Exact z-values and p-values are in the supplementary materials (Table S2 and S3).

Participants readily learned to estimate distances using the echo-like audio cue in two sessions (one hour each) of training with feedback. Targets and responses were transformed onto a log-scale to account for Weber's Law[25]. Each participant showed a significant correlation between target and response locations in the last half of the first session and also throughout the second session, all correlation coefficients > .80 (median r = 0.87), all p-values < .001 (individually displayed in Supplementary Figures S4-7). This is in line with previous findings that humans can quickly learn to use echoes and echo-like stimuli to make simple spatial judgements[26–28]. The following sessions tested participants' abilities to combine this newly-learned audio cue to distance with a visual one.

In Session 3, participants were given a mix of audio-only, visual-only, and simultaneous audio-visual trials, all with feedback. For each trial type, the log error was parsed into constant error (the mean of the residuals; displayed in Supplementary Figure S3) and variable error (the remaining variance, displayed in Figure 1D-F). In a sign-rank test, there was a significant decrease in variable error in audio-visual trials in Session 3 compared to the best single cue for that same participant (Figure 1D; exact statistics in Supplementary Table S2). Thus, participants benefited from having both cues available, meeting the first key criterion for Bayesian cue combination[1–6]. By the end of Session 3, there had been only 88 trials with both cues simultaneously. This effect therefore appeared very quickly. Up to that point, all trials had feedback – further sessions were used to examine what participants had learned and how they integrated the novel audio cue.

In Session 4, an untrained variation of the audio stimulus was introduced by altering the base frequency and removing feedback on all trials including this cue (feedback was still given on visual-only trials, to preserve motivation). If participants still integrate the novel audio cue in these conditions, it implies that their integration process goes beyond integration of specific stimuli learned by rote. The data show that even with these new, untrained audio-stimuli, participants still successfully lowered variable error below the best single cue on audio-visual trials in Session 4 (Figure 1E). The results imply that participants learned how the audio cue of time delay maps to distance. It leaves open the question whether they took cue reliabilities into account, like a Bayesian observer, or learned a simpler decision rule.

We addressed this question in Session 5 by changing the reliability of the visual stimulus and measuring the effect of this change on integration behaviour. Adapting to a change in reliability is a second key criterion for demonstrating Bayesian cue combination, since it implies an internal representation of own uncertainty[22]. We would expect people to rely less on vision when it is less reliable (i.e. the bubble distribution lengthened in Session 5 as compared to Sessions 1-4), and to rely more on vision when it is more reliable (i.e. the bubble distribution shortened in Session 5 as compared to Sessions 1-4). In Session 5, feedback was given on audio-only and visual-only trials, but not on audio-visual trials. Participants therefore had information about the change in visual reliability, but no feedback that could train them on how to combine the two cues. Participants' performance in Session 5 was in line with predictions based on Bayesian cue combination. Firstly, they still benefitted from having both cues available simultaneously (Figure 1F). Secondly, and more importantly, when vision was made less reliable in Session 5 as compared to previous Sessions 1-4 (Figure 2A, left) participants relied less on vision, and conversely, when vision was made more reliable in Session 5 as compared to previous Sessions 1-4 (Figure 2A, right), they

relied on it more. Overall, they lowered variable error regardless of which cue was more reliable (Figure 2B-C). This indicates flexible cue combination based on cue reliabilities, rather than use of a simple decision rule based on specific stimuli.

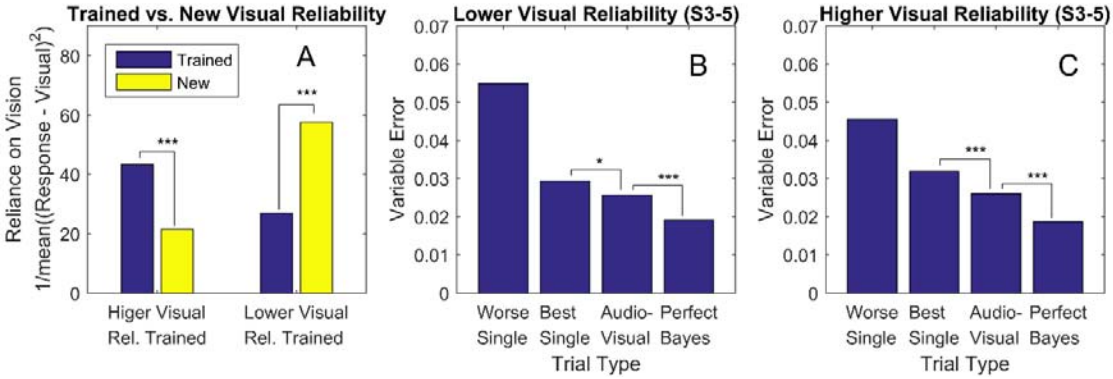

Figure 2. Measures of cue re-weighting. (A): The inverse of the mean squared deviation of responses from the center of the noisy visual cue, a measure of how much they relied on the visual aspect of the audio-visual stimuli, was decreased by participants for whom the cue became less reliable (bubble distribution lengthened; left) and increased by participants for whom the cue became more reliable (bubble distribution shortened; right). This graph and its sign-rank analyses only use the bimodal trials; * $p < 0.001$. (B-C): Participants reduced variable error regardless of which cue was more reliable.

However, participants achieved less than the optimal variance reduction predicted for a perfect Bayesian integrator (Figure 1D-F; 2B-C; all *p*-values < .01). One potential explanation for this is that perfect Bayesian integrators choose cue weights exactly proportionate to cue reliabilities, but that our participants did not know cue reliabilities as precisely, so that their integration differs for that reason. It is unknown how rapidly and accurately humans learn reliabilities of novel cues. Related studies on the acquisition of novel perceptual priors show that the correct reliabilities are not always learned accurately in limited training sessions (e.g. 1200 trials[29], similar to here). Another factor is that cue combination depends on inferring that two cues share a common cause[30]. This can limit combination with familiar or unfamiliar cues when these are perceived as discrepant, and with unfamiliar cues in particular, there may be less reason to assume that cues are related.

The results clearly show that adults rapidly learn to combine a novel audio cue with vision. This suggests that maturation, not experience, is the key factor in the development of cue combination. If years of cue-specific experience were necessary for learning to combine two cues, then we should have seen no benefit in the 3 hours that we gave adults to learn the new audio cue. Whilst it is the case that we did not see perfect Bayesian integration, even a modest overall benefit still surpasses the performance of children under 10 years[7–12]. Further, the integration we observed was flexible to accommodate stimulus changes in audio frequency and visual reliability, ruling out rote learning of specific stimulus associations and meeting key criteria of Bayesian integration. Our results are also counter to the predictions of current computational models of the development of cue combination that are free of

maturational factors[18,31]. Such models suggest that large numbers of cue-specific trials are needed to gather enough information for benefits from integration to emerge (approximately 1,000,000 trials in one paper[18]). Such models will need to incorporate maturational changes to account for the abilities of an adult system to rapidly learn and combine novel cues as we have shown here, in ways that a developing system (<10 years) cannot.

The specific maturational changes leading to flexible cue combination abilities are a crucial question for further research. Potential changes range from improvements in the (not yet well-understood) biological substrate for the operation of weighted averaging[32], to acquisition of unified multimodal representations (e.g. of distance), to slowing of physical growth so that sensory systems have reduced need for frequent cross-calibration[33], to accurate monitoring of each sensory system's uncertainty.

Another question that our results raise is the extent to which integration tapped into either low-level (sensory) or higher-level (decision) processes. It may be that, without more extensive practice, integration of an augmented sense requires focused effort on the part of users, which would have implications for its everyday use. Future studies should differentiate sensory vs decision-level integration, using neurorecording (e.g. "early" vs "late" signals via EEG) and/or behavioural manipulations (e.g. secondary tasks).

From an applied perspective, our results suggest that learning to combine a new sensory skill with a familiar sensory skill can happen on an extremely rapid timescale. In this case it was just a few hours, likely well before ceiling performance with the audio cue had been reached. This point is crucial because it means that new sensory skills need not replace familiar sensory skills. Humans whose senses are augmented stand to rapidly gain the Bayesian benefits of incorporating the new and standard sensory information, rather than having to choose between them. This discovery suggests that techniques like echolocation[19–21] and devices translating information from one modality into another[24] might not only hold promise for people with complete sensory loss (e.g. total blindness), but could in principle be a useful aid for less severe levels of impairment (e.g. moderate vision impairments, estimated to affect 214 million people alive today[34]). Our results also suggest an optimistic outlook on sensory augmentation more broadly, allowing people to use novel signals efficiently during specialist tasks – for example, auditory cues to position during brain surgery[35].

Cue combination is a ubiquitous phenomenon in adult perception, but not for children under 10-12 years old. Our results suggest that this difference does not rest on the accumulation of years-long experience with the individual cues. After learning how to use a novel sensory cue to distance, adults rapidly combined it with visual information to gain measurable precision, and also adapted to changes in both stimulus features and cue reliability. This is a step towards extending the human sensory repertoire – potentially overcoming sensory loss, as well as providing people with completely new kinds of signals.

Methods

Ethical approval was given by the Psychology Ethics Board at Durham University. Informed consent was acquired from all participants.

Design

Participants were trained and tested over 5 one-hour-long sessions, each containing up to 300 trials. On each trial, they heard an echo (Figure 1B), or saw a display of bubbles on the virtual sea (Figure 1C), or both. They then tried to indicate where the target, a cartoon whale (Figure 1A), was hiding under the sea, based on the stimuli they had received. The five different sessions varied in their cue parameters and feedback in order to test different aspects of learning (see Session Structure, below).

Participants

Participants were 12 healthy adults (10 female; age range 19.8-39.8 years; mean age 24.9 years; standard deviation 5.5 years) who did not have experience with echolocation. For half of participants the visual cue was set to be initially more reliable than the auditory, for half less; these reliabilities were reversed in Session 5 (see below for session structure, and supplemental materials for reliabilities). Furthermore, half of the participants were trained with audio-cues having a centre frequency of 2KHz, the other trained with 4KHz, swapping in Session 4 (see below for session structure, and 'stimuli' for the description of the audio stimulus). There were 3 people in each cell of this 2x2 design.

Apparatus

Participants used an Oculus Rift headset, a pair of AKG K271 MkII headphones and Soundblaster SB1240 soundcard, and an Xbox One controller. See Supplementary Materials for additional technical details.

Stimuli

The audio stimulus were two 'clicks' separated by an appropriate delay for the speed of sound in air and the distance to the target. Each click (illustrated in Figure 1B) was an amplitude modulated sinewave (either 2 or 4 kHz) (for details on amplitude modulation see Supplementary Materials), as used successfully by a previous project[36]. The length of the delay between the two clicks was the only auditory cue to distance. Our use of this single cue ensured that all participants were learning and using the same information.

The visual stimulus was a display of 256 'bubbles' arranged like a mirrored log-normal distribution (Figure 1C). The whale was equally likely to be beneath each of the bubbles – as if the whale chose a bubble at random and decided to hide directly under it. In absence of any other information, the optimal (error-minimising) strategy is to point to the centre, because points at the centre have the least average distance to all of the bubbles. However, because

there is considerable uncertainty about the whale's position, when a second (audio) cue to position is also available, the optimal strategy is to average the visual center with the auditory position estimate, based on their relative reliabilities[1–3]. Individual bubbles were easy to see (see Supplementary Materials for more details) and left visible on the sea while a response was entered. Using a stimulus with external uncertainty allowed us to perform the analyses in both Figure 1F (variable error reduction) and Figure 2A (cue re-weighting) on the same data collected immediately after the cue's reliability had changed.

Trial parameters for testing the reduction in variable error (Figure 1D-F) were generated in triplets. First, a bimodal trial was generated, based on four parameters: the target location, the audio frequency, the visual cue's center (where the bubbles were widest), and the visual cue's variance. Then a matched audio-only trial was formed by taking away the visual cue. Finally, a matched visual-only trial was formed by taking away the audio cue. This allowed us to perform matched sign-rank analyses between corresponding bimodal and unimodal trials. See Supplementary Materials for how these parameters were selected in different sessions.

Session Structure

Session 1 and 2 taught the audio cue to the participants. Session 1 (300 trials) started with two-alternative-forced choice (2AFC) for a point at the nearest vs farthest limit of the line, and progressed to 3AFC and 5AFC (see Supplementary Materials). The whale surfaced after every trial to give accurate feedback. Session 2 (300 trials), and all further sessions, began with a short warm-up of 2-, 3-, and 5-alternative-choice trials to remind participants how the audio cue works. In the second session, this proceeded to more audio-only trials with feedback, now with a continuous response (all positions along the line). For the continuous-response section, the targets were spaced evenly on a log scale.

Session 3 (299 trials; 83 matched triplets) tested for cue combination by showing people the visual cue only, the audio cue only, or both together. Feedback was given throughout.

Session 4 (298 trials; 62 matched triplets) tested for the ability to generalize the echo-like cue to a new emission, specifically a click with a different frequency of the amplitude modulated sine wave. The session was very similar to Session 3 except that all trials with the new sound were not given feedback.

Session 5 (300 trials; 83 matched triplets) tested for the ability to generalize to a new reliability level of the visual cue. It was much like Session 3 except that the log-normal distribution displayed by the visual cue changed in variance and there was no feedback given when the two cues were presented together.

Sessions were never on the same day, but were otherwise scheduled as close together as participants could accommodate. The first and last sessions were at most 9 days apart.

Formula for Variable Error

$$\left(\ln(\text{response}) - \ln(\text{target}) - \frac{\sum(\ln(\text{response}) - \ln(\text{target}))}{\text{Number of trials}}\right)^2 \qquad (1)$$

With the mean correction (i.e. constant error) calculated separately for each participant, session, and trial type combination (12 participants x 3 post-training sessions x 3 trial types = 108 corrections, displayed in Figure S3).

Control Experiment

This was like the audio-only continuous-response portion of Session 2 of the main experiment, but without any feedback or training.

Participants were 12 healthy adults (11 female; age range 18-20 years; mean age 19.4 years; standard deviation 0.66 years) who did not have experience with echolocation. The apparatus and audio stimulus was identical. Participants were played audio stimuli corresponding to targets along the line. The instructions were: “We want to see if you have any intuition about how echolocation works. Listen to the sound and try to point where I am hiding.”

Data Availability

Anonymized data are attached as a supplementary file.

Additional Details

The additional details needed to replicate the experiment, including stimulus parameters and a script of the cartoon whale’s feedback to participants, appear in the Supplementary Methods.

References

1. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27,** 712–719 (2004).

2. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16,** 1170–1178 (2013).

3. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415,** 429–433 (2002).

4. Hillis, J. M. *et al.* Slant from texture and disparity cues: Optimal cue combination. *J. Vis.* **4,** 1 (2004).

5. Knill, D. C. & Saunders, J. A. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res.* **43,** 2539–2558 (2003).

6. Alais, D. & Burr, D. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Curr. Biol.* **14,** 257–262 (2004).

7. Nardini, M., Jones, P., Bedford, R. & Braddick, O. Development of Cue Integration in Human Navigation. *Curr. Biol.* **18,** 689–693 (2008).

8. Gori, M., Del Viva, M., Sandini, G. & Burr, D. C. Young Children Do Not Integrate Visual and Haptic Form Information. *Curr. Biol.* **18,** 694–698 (2008).

9. Nardini, M., Bedford, R. & Mareschal, D. Fusion of visual cues is not mandatory in children. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 17041–6 (2010).

10. Gori, M., Sandini, G. & Burr, D. Development of visuo-auditory integration in space and time. *Front. Integr. Neurosci.* **6,** 77 (2012).

11. Petrini, K., Remark, A., Smith, L. & Nardini, M. When vision is not an option: children's integration of auditory and haptic information is suboptimal. *Dev. Sci.* **17,** 376–387 (2014).

12. Dekker, T. M. *et al.* Late Development of Cue Integration Is Linked to Sensory Fusion in Cortex. *Curr. Biol.* **25,** 2856–2861 (2015).

13. Spelke, E. Perceiving bimodally specified events in infancy. *Dev. Psychol.* (1979).

14. Bahrick, L. E. & Lickliter, R. Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Dev. Psychol.* **36,** 190–201 (2000).

15. Lewkowicz, D. & Turkewitz, G. Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Dev. Psychol.* (1980).

16. Gottfried, A., Rose, S. & Bridger, W. Cross-modal transfer in human infants. *Child Dev.* (1977).

17. Lewkowicz, D. The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychol. Bull.* (2000).

18. Daee, P., Mirian, M. S., Ahmadabadi, M. N., Brenner, E. & Tenenbaum, J. Reward Maximization Justifies the Transition from Sensory Selection at Childhood to Sensory Integration at Adulthood. *PLoS One* **9,** e103143 (2014).

19. Stroffregen, T. A. & Pittenger, J. B. Human Echolocation as a Basic Form of Perception and Action. *Ecol. Psychol.* **7,** 181–216 (1995).

20. Kolarik, A. J., Cirstea, S., Pardhan, S. & Moore, B. C. J. A summary of research investigating echolocation abilities of blind and sighted humans. *Hear. Res.* **310,** 60–68 (2014).

21. Thaler, L. & Goodale, M. A. Echolocation in humans: an overview. *Wiley Interdiscip. Rev. Cogn. Sci.* **7,** 382–393 (2016).

22. Maloney, L. T. & Mamassian, P. Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Vis. Neurosci.* **26,** 147 (2009).

23. Howard, I. P. & Rogers, B. J. *Seeing in Depth*. (Oxford University Press, 2008). doi:10.1093/acprof:oso/9780195367607.001.0001

24. Maidenbaum, S. & Abboud, S. Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. *Neurosci. Biobehav. Rev.* **41,** 3–15 (2014).

25. Getty, D. J. Discrimination of short temporal intervals: A comparison of two models. *Percept. Psychophys.* **18,** 1–8 (1975).

26. Teng, S. & Whitney, D. The acuity of echolocation: Spatial resolution in the sighted compared to expert performance. *J. Vis. Impair. Blind.* **105,** 20–32 (2011).

27. Tonelli, A., Brayda, L. & Gori, M. Investigate echolocation with non-disabled individuals. *J. Acoust. Soc. Am.* **141,** 3453–3453 (2017).

28. Thaler, L., Wilson, R. C. & Gee, B. K. Correlation between vividness of visual imagery and echolocation ability in sighted, echo-naïve people. *Exp. Brain Res.* **232,** 1915–1925 (2014).

29. Bejjanki, V. R., Knill, D. C. & Aslin, R. N. Learning and inference using complex generative models in a spatial localization task. *J. Vis.* **16,** 9 (2016).

30. Shams, L. & Beierholm, U. R. Causal inference in perception. *Trends Cogn. Sci.* **14,** 425–432 (2010).

31. Weisswange, T. H., Rothkopf, C. A., Rodemann, T. & Triesch, J. Bayesian Cue Integration as a Developmental Outcome of Reward Mediated Learning. *PLoS One* **6,** e21575 (2011).

32. Ohshiro, T., Angelaki, D. E. & DeAngelis, G. C. A normalization model of multisensory integration. *Nat. Neurosci.* **14,** 775–782 (2011).

33. Burr, D. & Gori, M. *Multisensory Integration Develops Late in Humans. The Neural Bases of Multisensory Processes* (CRC Press/Taylor & Francis, 2012).

34. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388,** 1545–1602 (2016).

35. Woerdeman, P. A., Willems, P. W. A., Noordmans, H. J. & van der Sprenkel, J. W. B. Auditory feedback during frameless image-guided surgery in a phantom model and initial clinical experience. *J. Neurosurg.* **110,** 257–262 (2009).

36. Thaler, L. & Castillo-Serrano, J. People's Ability to Detect Objects Using Click-Based Echolocation: A Direct Comparison between Mouth-Clicks and Clicks Made by a Loudspeaker. *PLoS One* **11,** e0154868 (2016).

Supplementary Methods (Main Experiment)

This section describes the full details of the study's method that are needed for replication.

Participants

Twelve participants were recruited through the Durham Psychology Participant Pool and through posters around Durham University. There were 2 males and a mean age of 24.9 years (age range 19.8-39.8 years; standard deviation 5.5 years). Participants were paid £8 per hour.

Apparatus

Virtual environment. A custom seascape was created in WorldViz Vizard 5 (Santa Barbara, CA, USA) and presented using an Oculus Rift headset (Menolo Park, CA, USA). This seascape contained a large flat blue sea, a 'pirate ship' with masts and other items, a virtual chair, and a friendly cartoon whale introduced with the name "Patchy" (see Figure 1 and Figure S2). Participants were seated 4.25 m above the sea. The response line (range of possible positions for the whale) stretched out from the bow of the ship, and was marked by periodic variations in the color of the sea. A pair of black bars marked the 10 m and 35 m points, which were the nearest and furthest possible responses. Distances along the line could be judged visually via perspective and height-in plane (see Figure S1) as well as, in theory, stereo disparity (although stereo information at the distances used is of limited use). Patchy gave written instructions via a white speech bubble (examples, Figure 1A and Figure S1D). To respond, participants set the position of a 3D arrow that touched the sea surface (Figure S1A). The sea surface remained still and the ship did not move.

The Oculus Rift headset has a refresh rate of 90 Hz, a resolution of 1080 x 1200 for each eye, and a diagonal field of view of 110 degrees. Participants were encouraged to sit still and look straight ahead during trials but did not have their head position fixed. The Rift's tracking camera and internal accelerometer and gyroscope accounted for any head movements in order to render an immersive experience.

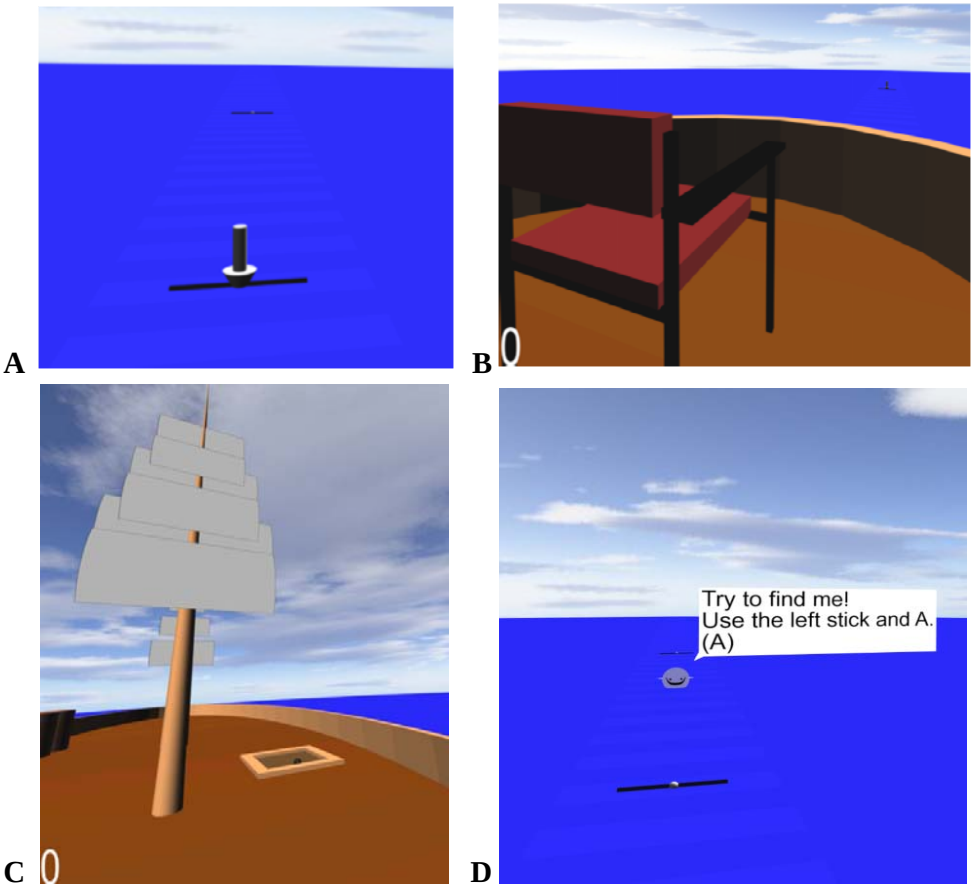

Figure S1. The virtual environment. The responding line and its limits, and the arrow participants used to indicate responses (A), the virtual chair they sat in (B), the ship they were on (C), and Patchy the whale (D) were the only salient objects in the environment.

Audio equipment and stimuli. Sound was generated and played using a MATLAB program with a bit depth of 24 and a sampling rate of 96 kHz. A USB sound card, (Creative SoundBlaster SB1240; Singapore), was attached to a pair of AKG K271 MkII headphones (Vienna, Austria) with an impedance of 55 ohms.

The audio stimuli were created by first generating a 5ms sine wave either 4000Hz or 2000Hz in frequency with an amplitude of 1. The first half-period of the wave was scaled down by a factor of 0.6. An exponential decay mask was created starting after 1.5 periods and ending at 5ms. The exponent was interpolated linearly between 0 and -10 over that period. This was all embedded in 1s of silence, with a 50ms delay before the sound appeared. An exact copy of the sound was added after an appropriate delay, calculating the distance to the target divided by the speed of sound (approximated at 350m/s), then times two (for the emission to go out, and also to come back). With a minimum distance of 10m, the two sounds (clicks) never overlapped (although it is possible that subjects experienced them as one sound). Real echoes contain more complex information, including reductions in amplitude with distance, but we chose to make delay the only relevant cue so that we could be certain

which information all participants were using. Our stimuli also allowed us to use range of distances at which real echoes are typically very faint, minimizing the scope for participants to have prior experience with them.

Visual cue. The visual cue was an array of 256 'bubbles' (translucent white spheres with a radius of .15 m and 50% opacity) arranged to show a mirrored log-normal distribution perpendicular from the line that the whale appeared on (Figure 1C). This was generated by first lining the outer edges with the bubbles and then filling in the interior area such that no two bubbles touched each other. The result looks like a violin plot of a log-normal distribution. This was arranged such that the target was always an actual draw from the distribution on display, i.e. the whale's true hiding place was under one of the bubbles.

We designed this visual cue to have external rather than internal noise. Participants could see exactly where the visual cue was on the sea, and it remained on display while they made their response. However, what they were shown was a distribution from which the target was chosen. This noise (uncertainty) makes it impossible to be exactly correct on every trial, but the nature of the cue also makes the relative probabilities of different regions clear. In contrast, the noise in the audio cue was internal rather than external – the audio stimulus was always played with exactly the correct delay, so any imprecision in its use is due to the perception of participants. The advantage of using a visual cue with external noise was to allow both the analyses in Figure 1 (variable error) and Figure 2 (reliance on the visual cue) in the same data collected soon after the visual reliability changed in Session 5. The variance of the visual cue was calculated for each participant based on their auditory-cue variance, and changed in the final session; see below.

Controller. Participants used an Xbox One controller (Redmond, WA, USA). They only used the left joystick and the A button. Pressing the other buttons did not have any effect on the experiment.

Procedure

Participants were told that that they were going to play a kind of 'hide and seek' game with Patchy. They were told to use the sounds and, later, the 'bubbles' to try and point as close as they can to Patchy on each trial. They were given a consent form and then helped into the equipment. After that, Patchy took over giving more specific instructions.

On each trial participants receiving the cues and then used the joystick to move an arrow to guess where Patchy was hiding. Participants pressed A to register their response. There was no time limit. Participants were not given any additional information while responding nor allowed to hear the audio stimulus again. The visual stimulus remained static on the sea while they were responding. Except where noted, after a response was entered Patchy appeared with a speech bubble indicating the error as a percentage of his true distance away from the ship (e.g. if he was 10 m away and they pointed 12 m away, they would see +20%). Participants pressed the A button again to move on to the next trial (or to view any new instructions that were being given). The structure of the different session is detailed below and summarised in Table S1. Sessions were never on the same day but were otherwise as close as possible as participants could accommodate (at most 9 days between Sessions 1 and 5).

Training session 1. This session was designed to introduce the echo cue and scaffold participants' learning towards its use. As with all sessions, participants were greeted with Patchy saying "Let's echo! Press A!" which remained until the participant pressed A. He then said, "First just listen to the nearest and furthest places I can hide." Participants then listened and watched passively while Patchy moved back and forth between 10 m and 35 m distance five times, playing the correct sound with each appearance. He then said, "Try to find me! Use the left stick and A." Participants were then played an audio stimulus indicating either 10 m or 35 m. They then used the left joystick to select between those two options. If they were correct, Patchy would appear under their arrow and play an animation like nodding yes (up and down). If they were incorrect, Patchy would appear at the correct target and play an animation like shaking his head no (left and right). There were a total of 50 of these two-alternative forced-choice (2AFC) trials, with 25 of the targets at 35 m and 25 and 10 m.

When those 50 trials were completed, they moved on to a 3AFC version. The targets were at 10 m, 22.5 m, and 35 m. Patchy said, "Good job! Now listen to the middle too." They were then played all three sounds, cycling from nearest to farthest and coinciding with Patchy's appearance at those distances, four times. He then said, "Now try to find me!" and 100 trials of 3AFC were run. Feedback was given in the same way as the 2AFC. The distribution of targets was even among the three possible places.

When those 100 trials were completed, they moved on to a 5AFC version. The targets were at 10 m, 16.25 m, 22.5 m, 28.75 m, and 35 m. Patchy said, "Good job! Now listen to two more." They were then played all five sounds in three cycles. He again said "Now try to find me!" and 150 trials of 5AFC were run. Feedback was the same again, and the target distribution was again even. No introduction of the next session was given. As with the ending of all sessions, Patchy said, "That's all for today! Bye!" and the participant was helped out of the equipment.

Warmup block. Sessions 2 through 5 began with a warmup block that was a shortened version of Training Session 1. It was all the same except there were only 8 trials of 2AFC, 12 of 3AFC, and 20 of 5AFC, for a total of 40. This was done to remind participants of how the echo cues work.

Training session 2. This session was designed to help participants move from using the echoes for identifying a constrained set of possible responses over to making a response anywhere along a continuum. The session began with the warmup block for 40 trials and then asked participants to find Patchy based on the audio cue alone, with feedback, for 250 trials. Target locations were spaced evenly from 10 m to 35 m on a logarithmic scale, shuffled into a random order. Every 50 trials, Patchy appeared briefly to tell participants their average error in percent over that period - this was in addition to being given the percent error after every trial. This was phrased as, "For echoes, you were off by X% on average over the last 50 trials."

After completing the 290th trial of the session, two things happened. First, the standard deviation of participants' errors (response minus target) over the last 100 trials was calculated after converting both targets and responses onto a log scale. A log conversion as used to account for the expected effects of Weber's Law on time interval judgements[25]. This estimate of auditory-only performance was used to generate the trial parameters for the rest of the experiment as detailed below.

Second, the next session was previewed. The visual cue was explained to participants. Patchy said, "You're almost done today. Let me show you what you'll do next time. Let's try something you can see. I like hiding by bubbles. More bubbles nearby means I'm more likely to be there. If you only see bubbles, the best you can do is point where there are the most of them." No further instructions about the log-normal distribution were given. Participants were then given five opportunities to find Patchy with just the visual cue. They were given feedback on each trial. Then Patchy said, "From now on, sometimes you get both the echoes AND the bubbles!" Participants were then given five opportunities to find Patchy with both the audio and visual cues, again with feedback. Visual stimuli (bubble distributions) had a standard deviation on a log scale that was either 75% or 125% of the estimate of each participant's audio-cue standard deviation (see above), with half of participants assigned to each of these conditions (see below). The centers of the bubble distributions were placed evenly on the response line on a log scale, in random order. The targets' deviations were generated as 10 values from -2 to +2 standard deviations and then randomly assigned to the distributions, with the actual target locations truncated at 10m and 35m. This completed session 2.

Stimulus generation for sessions 3 through 5. Sessions 3 through 5 used a common scheme for generating trials that are audio-only, visual-only, or audio-visual. First we took the number of available trials, divided by three, and rounded down. We then generated this number of distinct audio-visual trials. For these, we selected centers for the visual stimuli that were even from 10 m to 35 m on a logarithmic scale. We calculated a bank of cumulative probabilities spaced evenly from 2% to 98%. Each visual center was paired randomly with a single cumulative probability. The actual target (Patchy's location) was then placed at the appropriate place on the visual cue's distribution according to its associated cumulative probability. For half of participants, on all of the trials in Session 3, and all but the last 10 trials of Session 4, the standard deviation on a log scale was 75% of the estimate of their audio standard deviation on a log scale as generated on the 290th trial of Session 2. For the rest of the trials, it was 125%. For the other half of participants, this was switched, with the higher standard deviation going first. Finally, an appropriately-delayed sound corresponding to the actual Patchy location was generated and paired to the target. These audio-visual trials were placed randomly in the available trial slots.

For each of these audio-visual trials, we then created a matching visual-only and a matching audio-only trial. This was done just by removing the audio or the visual component as appropriate. These trials were likewise placed randomly in the available trial slots. This matching procedure means that within each session, trials existed in triplets where (a) the visual cue in the visual-only trial and the audio-visual trial were exactly the same, (b) the audio cue in the audio-only trial and the audio-visual trial were exactly the same, and (c) the target was in the same place. This means that when we compare the single-cue trials to the dual-cue trials within each triplet, there was nothing different except the other unused cue in the single-cue trials. This method also guaranteed that the visual cue was valid over the session i.e. there were an appropriate number of trials near the center mass and out on the edges so it could be interpreted directly as a likelihood function of a distribution, making good on the promise that "more bubbles means I'm more likely to be there".

Main testing session (3). This session was intended to assess cue combination with the newly-learnt cue. It began with the warmup block for 40 trials and then had 249 trials that were a mix of audio-only, visual-only, and both. The trial parameters were generated as described above. Feedback was given after every trial and aggregate feedback was given every 50 trials, as described above. Patchy appeared briefly just before the first visual-only trial to say, "Remember those bubbles I like?"

For the last 10 trials, the next session was previewed. Patchy appeared and said, "Almost done today. Let's look at what you'll do next time. Let's try an echo with a different sound going out. The way you use it is the same. More time between sounds means further out. For these, you won't see me pop up." They were then given 10 trials, evenly spaced on a log scale, with audio stimuli using the untrained frequency. When they entered a response, the arrow bobbed down into the sea for a quarter of a second, but they were not given any feedback (nor on any other trial involving the untrained frequency). This completed session 3.

Frequency generalization session (4). This session was designed to see if participants' learning in the first session was specific to the emission that they learned or if it would generalize to a new frequency. It again started with the warmup block for 40 trials. Then there were 248 trials that went in the order of (i) a reinforced trial with the trained frequency, (ii) an unreinforced trial with just the audio at a new frequency, (iii) a reinforced trial with just a visual stimulus, and (iv) an unreinforced trial with the new frequency and the visual stimulus, repeating 62 times. The trials with the new frequency were generated in triplets as described above. This sequence was selected to keep the session from becoming too discouraging and also to give participants some feedback to help keep them calibrated to the mapping of delays. Patchy appeared every 50 trials to give aggregate feedback as before, including the new frequency trials in the report.

For the last 10 trials, the next session was previewed. Patchy appeared and said, "The bubbles are going to be a little [more/less] spread out from now on." Then there were 10 trials with the visual stimulus with a changed standard deviation, 75% or 125% of the estimate from above. The actual targets were evenly spaced from -2 standard deviations to 2 standard deviations, in random order, truncated at 10m and 35m. Feedback was given. This completed session 4.

Reliability generalization session (5). This session was designed to see if participants could adapt to a change in the reliability of one of the stimuli. First there was the warmup block of 40 trials. Then there were 10 trials, with feedback, with the visual stimulus and the changed reliability. This was done to make sure that participants could notice the change. A single audio-only trial was inserted to remind them that they needed to listen for it, with a target in the middle of the response line and feedback. Then there were 249 trials that were a mix of audio-only (with the trained frequency), visual-only with lower reliability, and both. Feedback was given on the visual-only trials and the audio-only trials, but not on trials with both cues. This was done to prevent more rote methods of adaptation. The trial parameters were generated in triplets as described above. Aggregate feedback was given every 50 trials for the audio-only trials. This completed the final session.

Table S1. Session organization.

Session	Stimulus	Response	No. of Trials	Feedback
1	Audio only	2AFC	50	Yes
1	Audio only	3AFC	100	Yes
1	Audio only	5AFC	150	Yes
2	Audio only	2AFC	8	Yes
2	Audio only	3AFC	12	Yes
2	Audio only	5AFC	20	Yes
2	Audio only	Continuous	250	Yes
2	Visual only	Continuous	5	Yes
2	Both	Continuous	5	Yes
3	Audio only	2AFC	8	Yes
3	Audio only	3AFC	12	Yes
3	Audio only	5AFC	20	Yes
3	Mixed: Audio, Visual, Both (83 each)	Continuous	249	Yes
3	Audio w/ new emission	Continuous	10	No
4	Audio only	2AFC	8	Yes
4	Audio only	3AFC	12	Yes
4	Audio only	5AFC	20	Yes
4	Mixed: Audio w/ trained emission, Audio w/ new emission, Visual, Both w/ new emission (62 each)	Continuous	248	Yes for audio w/ trained emission, Yes for visual, No for others
4	Visual w/ changed reliability	Continuous	10	Yes
5	Audio only	2AFC	8	Yes
5	Audio only	3AFC	12	Yes
5	Audio only	5AFC	20	Yes
5	Visual only w/ changed reliability	Continuous	10	Yes
5	Audio only	Continuous	1	Yes
5	Mixed: Audio, Visual w/ changed reliability, Both w/ changed visual reliability (83 each)	Continuous	249	Yes for audio only or visual only, No for both

Supplementary Methods, Results, and Discussion for the Control Experiment

In the control experiment, 12 echo-naïve participants were asked to estimate distances based on the same 250 audio stimuli in the same distribution as the continuous-response trials of Session 2 of the main experiment, except without feedback or prior training. Our main conclusion is that these participants did not have any useful knowledge of how the echoes mapped onto distances at the beginning of the experimental session, possibly not at the end either. From this we can conclude that participants in the main experiment were indeed naïve to the new cue we chose to train them with, and that typical participants have no meaningful amount of prior experience or ability to guess how the cue works.

The analysis again proceeded by taking the natural logarithm of the targets and responses (Figure S2A-B). Just from visual inspection, it is clear that the relation between the two is weak for both the first 125 trials ($R^2 = 0.093$) and the last 125 trials ($R^2 = 0.172$), especially compared to the results in the main experiment (Figure S2D-E), though it is improving in the second half. It is possible that participants had some intuition about the correct structure: a longer delay means further away. But they may not have the correct specific mapping: a delay of X seconds indicates that the target is 175 * X meters away. The distances we tested were much greater than those over which echoes can usually be heard. Participants might have tried to make this mapping during the experiment by trying to remember the shortest delay they experienced, mapping this to the nearest response point, and the same with the longest/furthest.

We reasoned that a good standard for evaluating participants' use of the audio cue is to compare it to how they would have performed with a degenerate strategy where they just pointed at the center of the response line on a log scale on each trial. If participants could not perform better than this, then they were no better off listening to the sounds than they would be using a guessing strategy.

Participants' performance in the control experiment was compared to predicted performance with this degenerate strategy. This was quantified as the observed error on each trial, squared to make them all positive, minus the distance between the target and the center, also squared. This means that positive numbers represent worse performance than the degenerate strategy and negative numbers represent better performance. This measure was positive and significant overall for the control experiment, with a mean of 0.026, $t(2999) = 6.275$, $p < .001$. That is, overall, these participants performed significantly worse than predicted by guessing the middle location on every trial. To see if there was any improvement over the course of the task, we also regressed this measure onto trial number (Figures S2C). This line was above 0 for all trial numbers, and its 95% CI was above 0 until trial 189, never falling entirely below 0. In comparison, the same line and 95% CI is below 0 for all trial numbers in the main experiment (Figure S2F). This suggests that training allowed participants to outperform the degenerate strategy for this entire block of trials, but untrained people were worse for at least 189 trials. Mere exposure to the range of the stimuli may be responsible for a slow improvement that could potentially lead to some useful use of the cue over a longer time-scale, but crucially subjects with neither feedback nor this exposure – the left-hand (trial 0) estimate for the regression line – do not show this.

We conclude that these data are inconsistent with the hypothesis that echo-naïve participants have any meaningful cue-specific prior experience with this task.

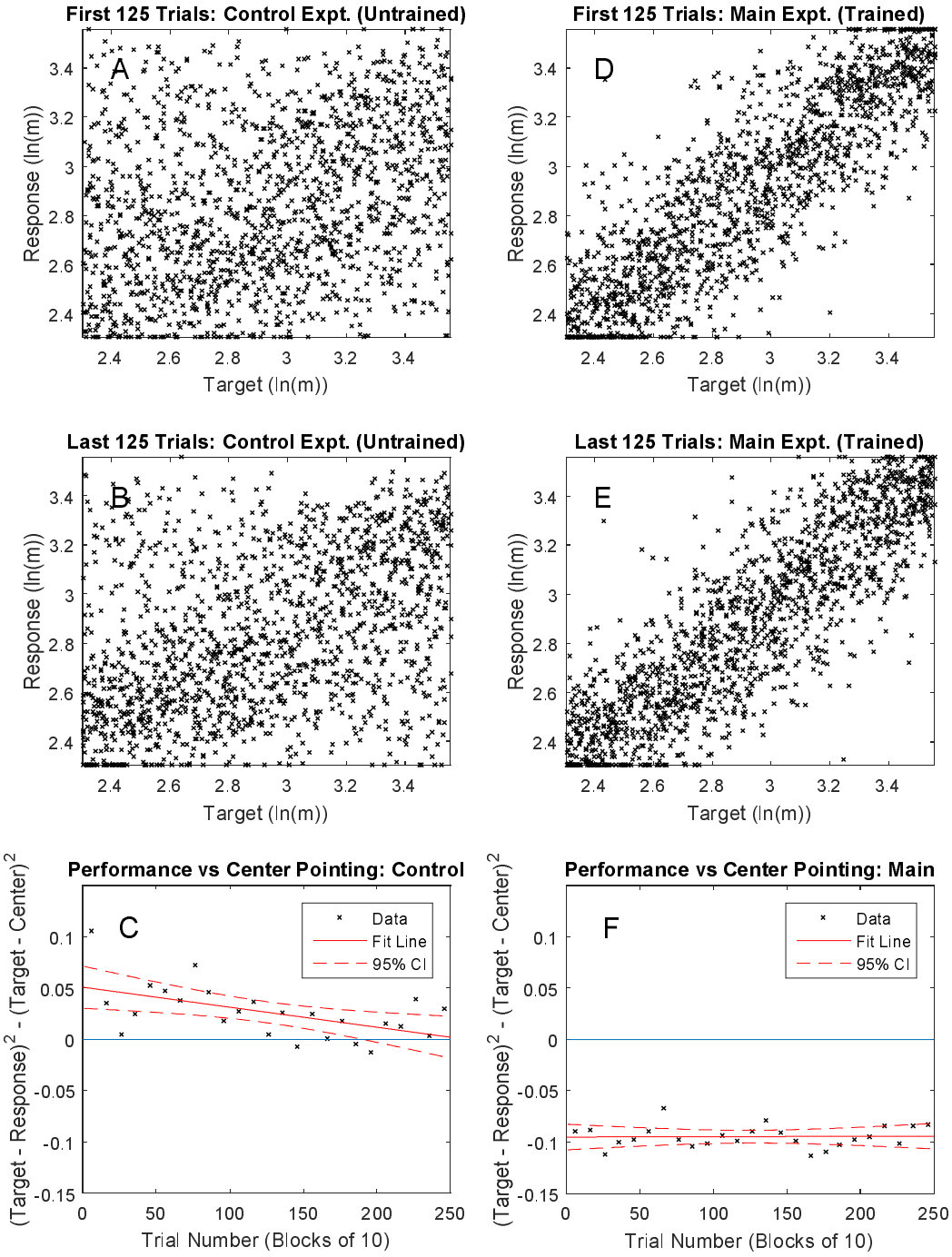

Figure S2. Comparing trained versus untrained performance with the audio cue. The top four graphs show targets versus responses, broken down by the first 125 trials (A, D) versus the last 125 trials (B, E), and also untrained (A, B) versus trained (D, E). In the bottom row (C, F), the red slanted line is the fit to the data, the dashed lines are 95% confidence interval, the blue line is a reference at 0, and the black crosses are means for blocks of 10 trials. Data above the blue line indicate worse performance than a degenerate strategy where they always point to the center of the response line. Left (C) is again untrained and right (F) is again trained.

Supplementary Tables and Figures

Table S2. Exact statistics and Effect Sizes for Sign-Rank Tests.

Analysis	N trial pairs	*z*	*p*	Cohen's D
Omnibus				
Best Single vs Bimodal	2736	5.020	< .001	.1029
Bimodal vs Optimal	2736	-5.839	< .001	.1519
All Trained (1D)				
Best Single vs Bimodal	996	2.994	.003	.1149
Bimodal vs Optimal	996	-2.602	.009	.1690
New Frequency (1E)				
Best Single vs Bimodal	744	3.053	.002	.1317
Bimodal vs Optimal	744	-3.325	.001	.1440
New Visual Reliability (1F)				
Best Single vs Bimodal	996	2.659	.008	.0676
Bimodal vs Optimal	996	-4.269	< .001	.1446
Changed Vis. Rel. Higher (2A)	498	-5.698	< .001	-.4133
Changed Vis. Rel. Lower (2A)	498	7.197	<. 001	.4604
Lower Visual Reliability (2B)				
Best Single vs Bimodal	1285	2.150	.032	.0804
Bimodal vs Optimal	1285	-4.171	< .001	.1531
Higher Visual Reliability (2C)				
Best Single vs Bimodal	1451	4.914	< .001	.1218
Bimodal vs Optimal	1451	-3.908	< .001	.1546

Note. For the top item, 1D-F, and 2B-C, data were analysed in a two-tailed sign-rank test on the variable error, which is the remaining error after correcting for constant error (difference between mean response and target for each participant session) squared. For the remaining items in 2A, data were again analysed in a two-tailed sign-rank test but instead entering the squared distance between the center of the visual cue and the response. This was all done after converting targets and responses to a log scale.

Table S3. Descriptive Statistics for Variable Error.

Source	Mean	Median	Interquartile Range
Session 3			
Worse Single	0.0505	0.0192	0.0553
Best Single	0.0318	0.0130	0.0371
AV	0.0265	0.0100	0.0313
Session 3			
Worse Single	0.0510	0.0181	0.0617
Best Single	0.0335	0.0130	0.0373
AV	0.0269	0.0086	0.0313
Session 3			
Worse Single	0.0487	0.0176	0.0518
Best Single	0.0275	0.0109	0.0316
AV	0.0244	0.0082	0.0241

Note. Because of very high skewness in a squared error term, interquartile range is given rather than variance or standard deviation.

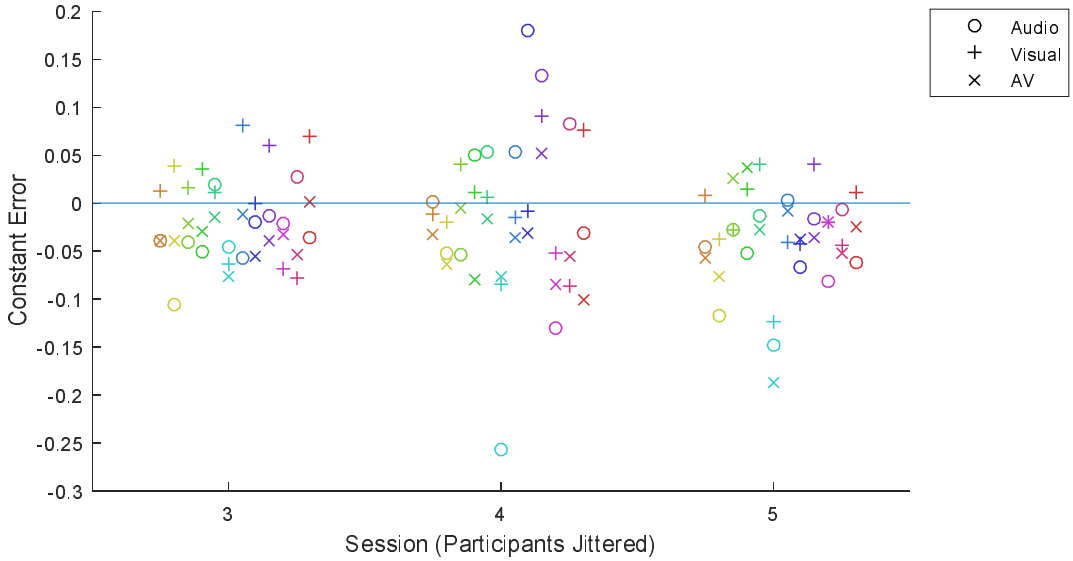

Figure S3. Constant error (mean difference between target and response, on a log scale) was generally low in this dataset. Specifically, it was less than 0.15 log units in 105/108 cases, which is low enough for combination to still be useful with a variable error of 0.03 and 0.05 (Figure 1A in main text). The largest issues with constant error were in the audio cue in Session 4, where the base frequency was altered and feedback was removed.

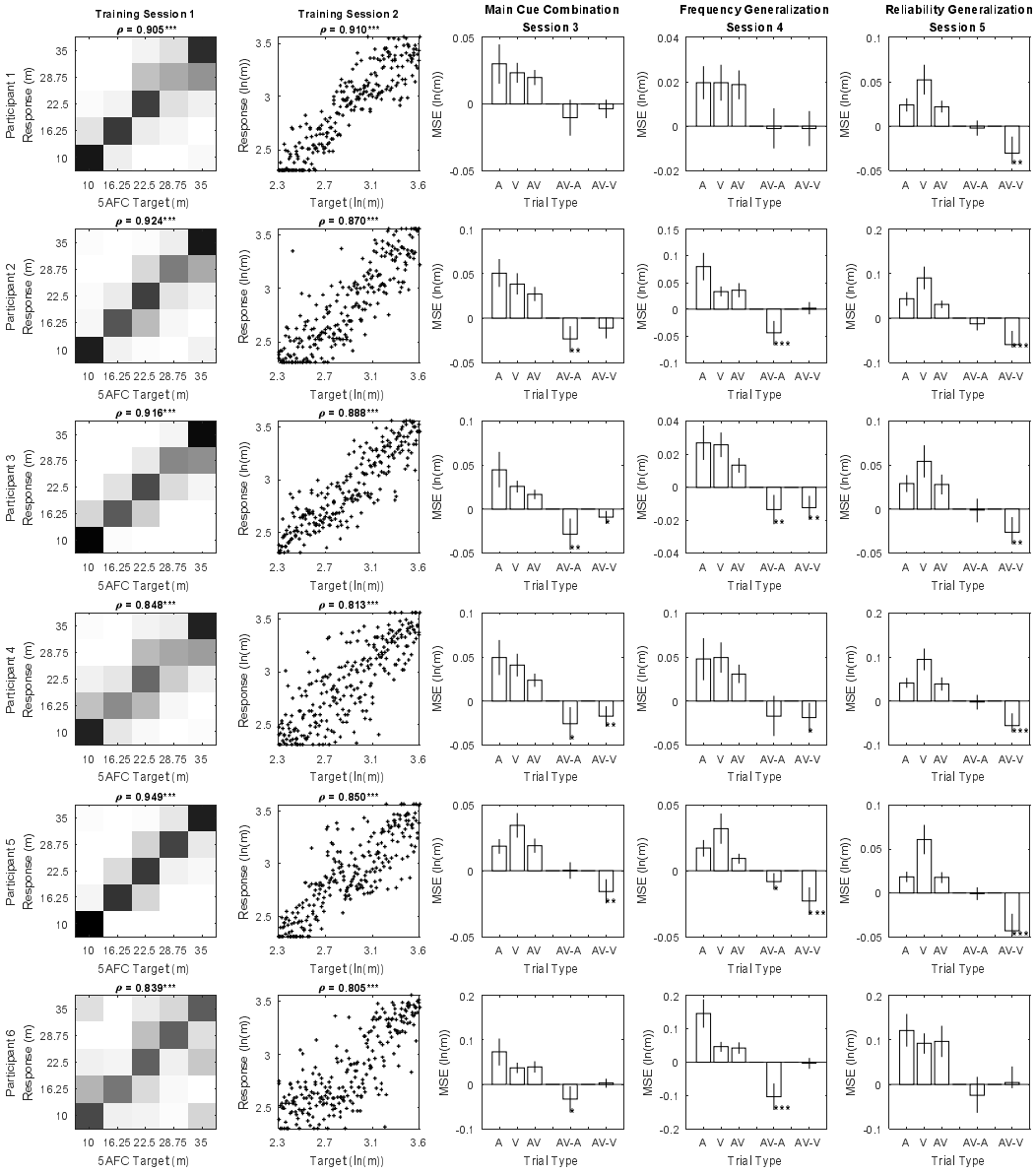

Figure S4. Detailed results by participant and session (participants 1-6). The first column is a heatmap of targets vs responses during session 1. The second column is target versus response on a log scale during session 2. The remaining chart variable error during sessions 3-5 for the audio only trials (A), the visual only trials (V), the bimodal trials (AV), the bimodal variable error minus each matching unimodal one (AV-A and AV-V), with a 95% confidence interval.

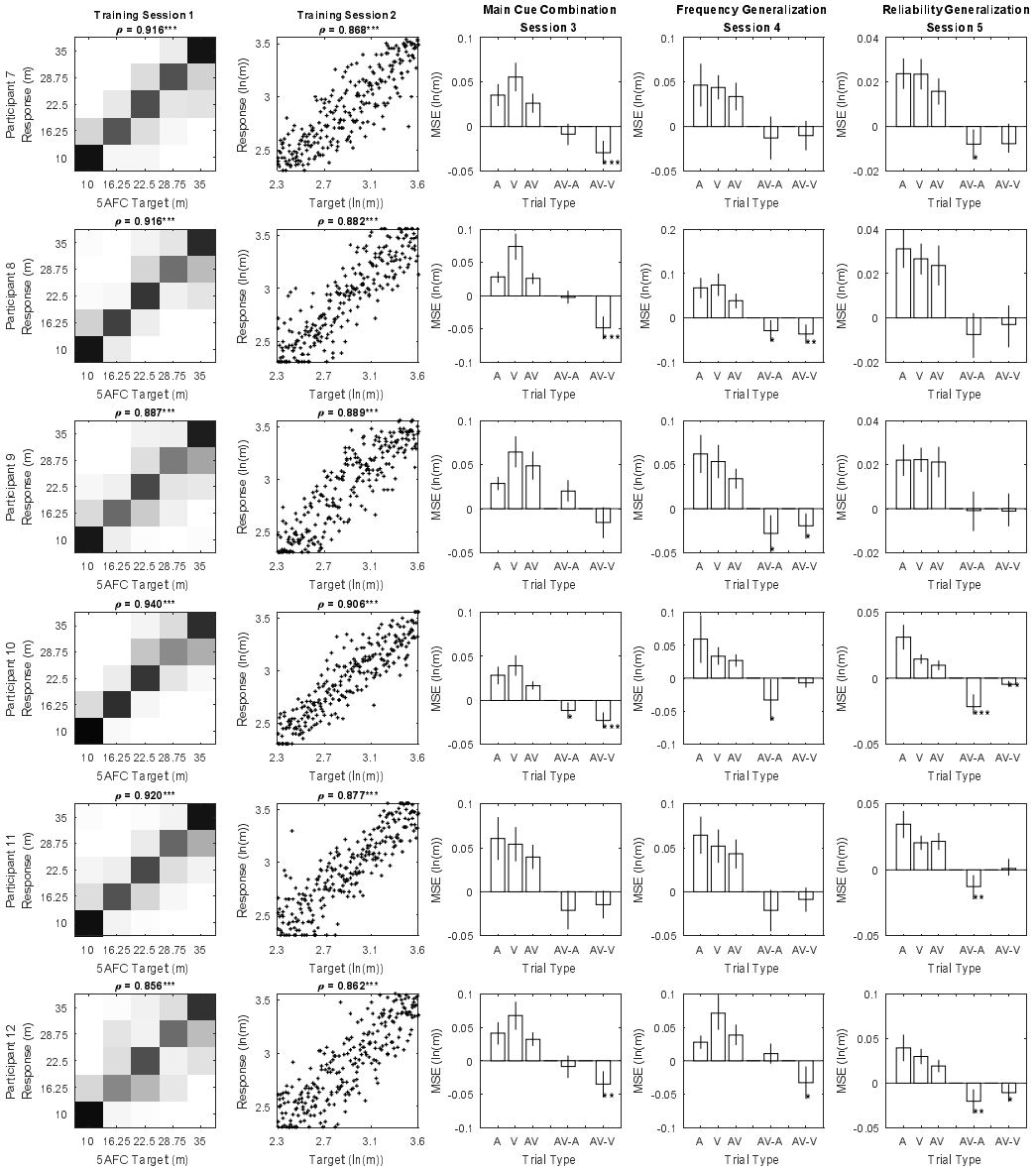

Figure S4 continued. (participants 7-12)

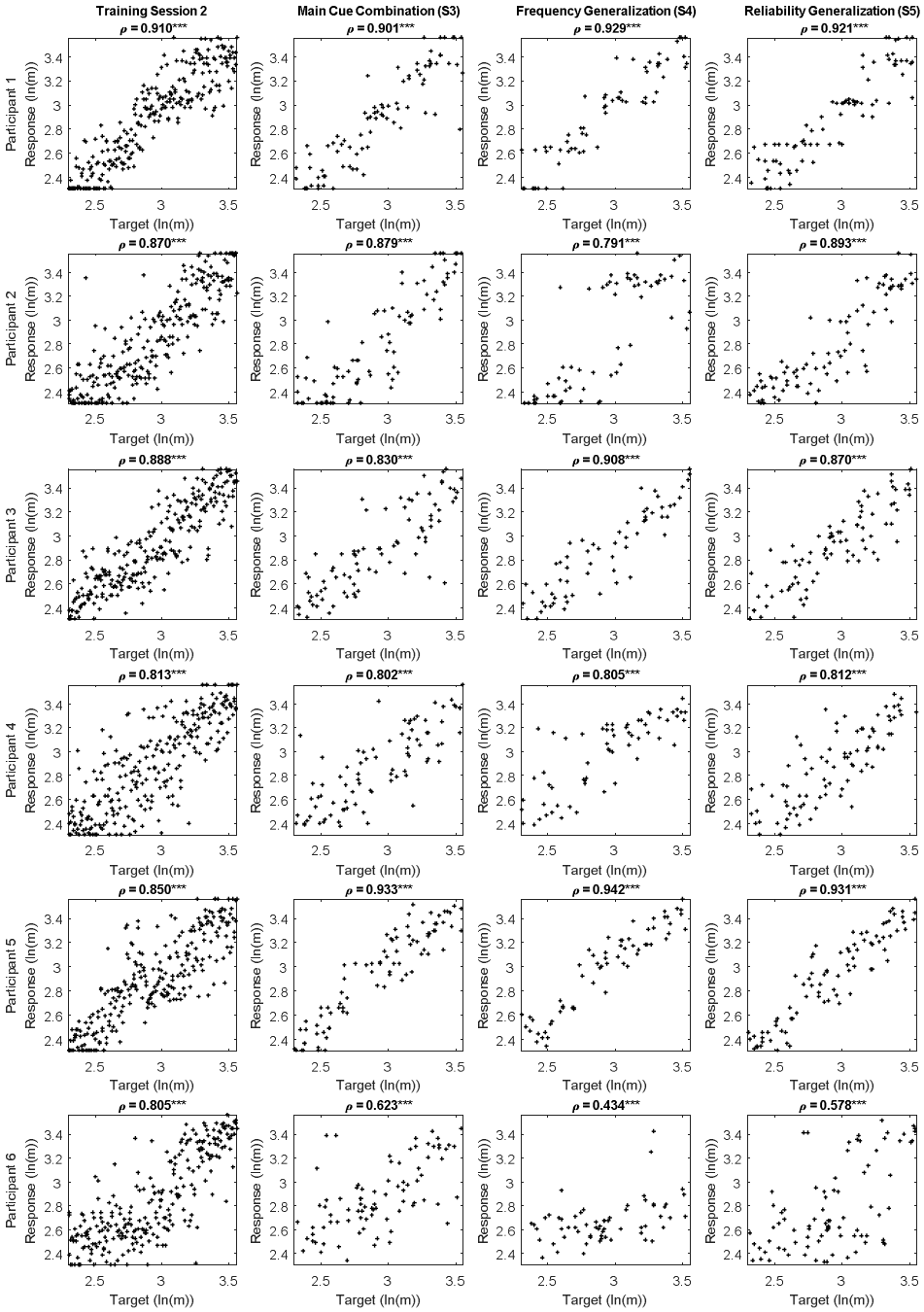

Figure S5. Targets versus responses for the audio-only trials in all sessions (2-5) with a continuous response (participants 1-6). For the frequency generalization sessions, the trials for the new frequency are displayed.

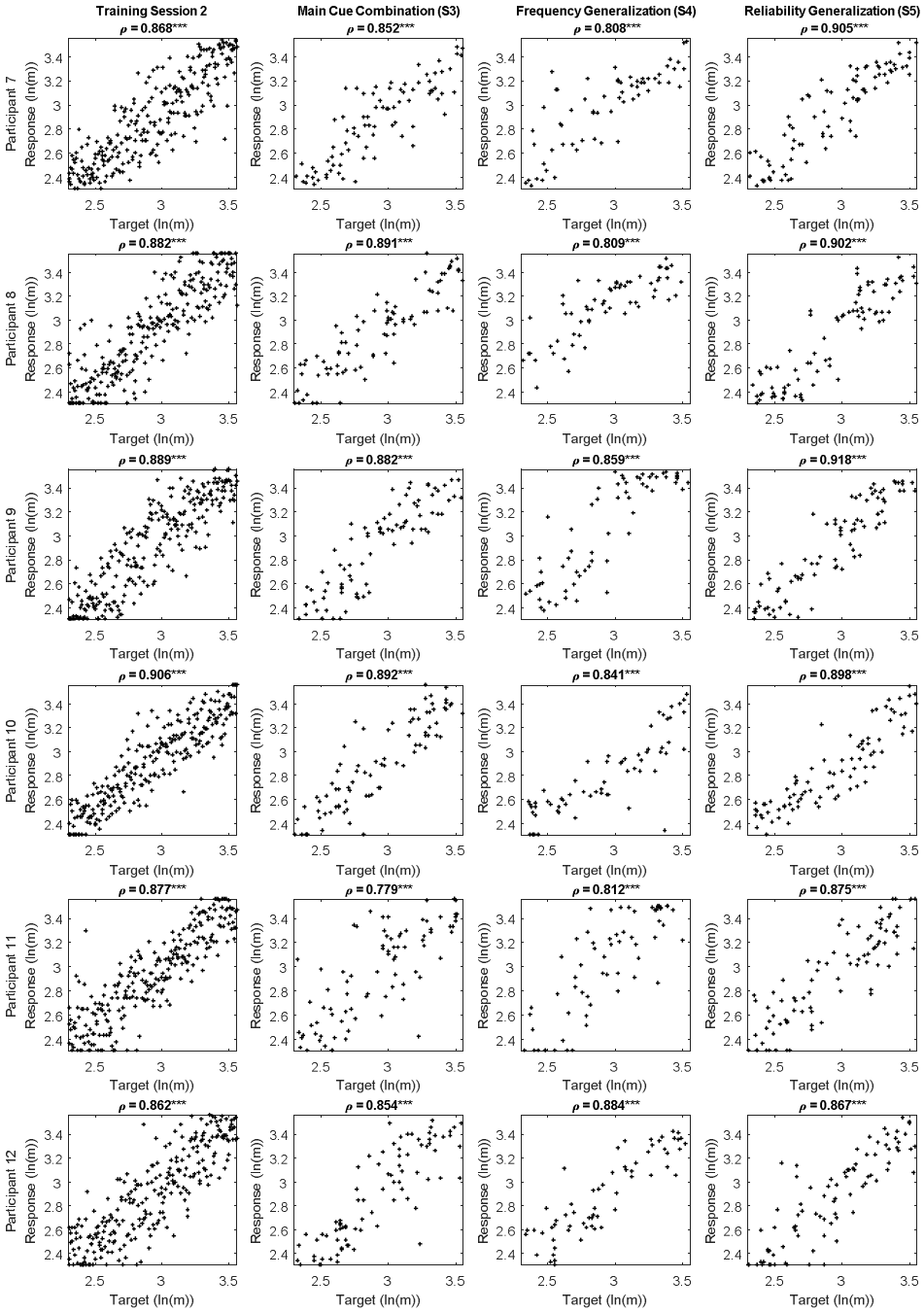

Figure S5 continued. (Participants 7-12)

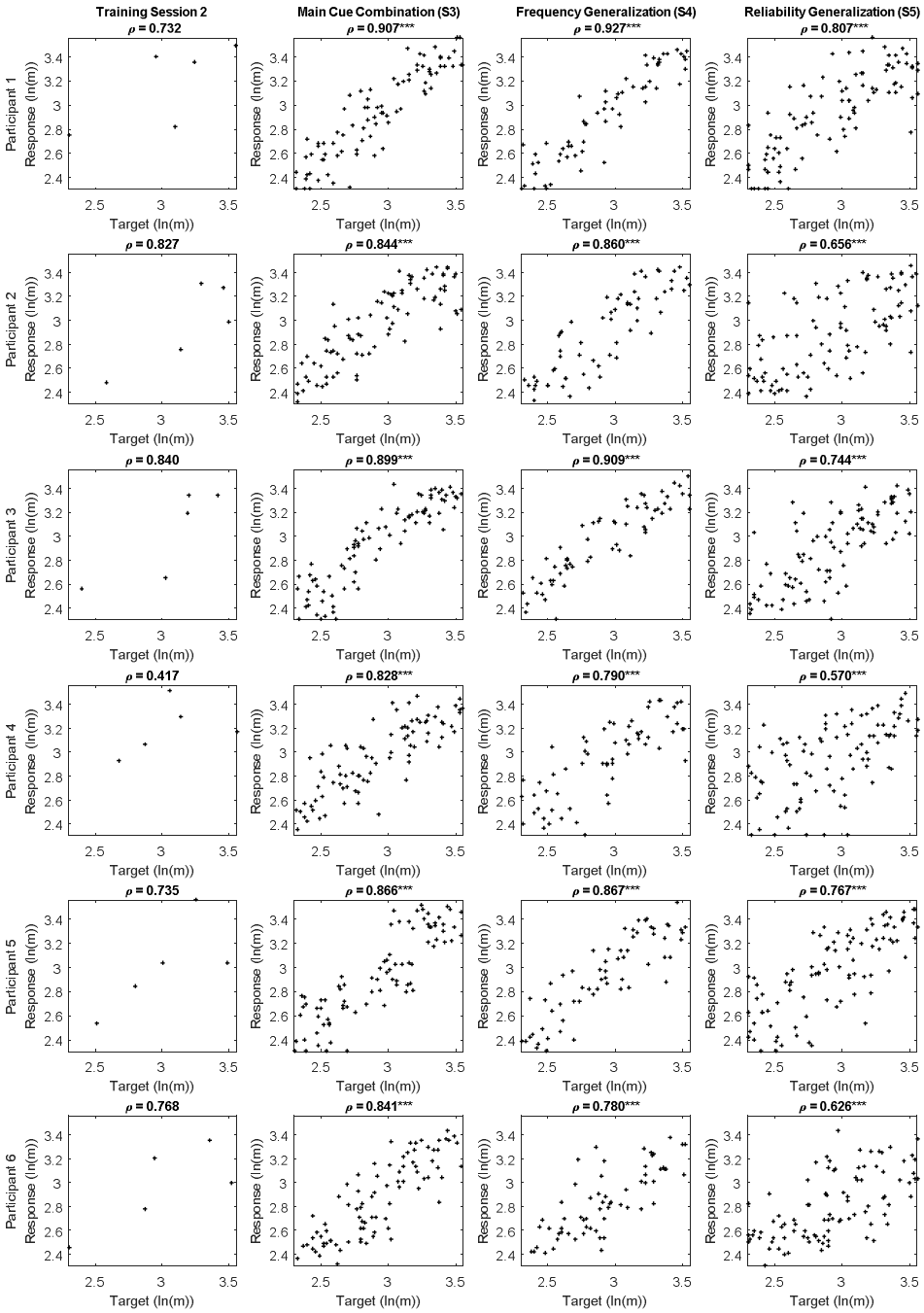

Figure S6. Targets versus responses for the Visual-Only Trials (participants 1-6).

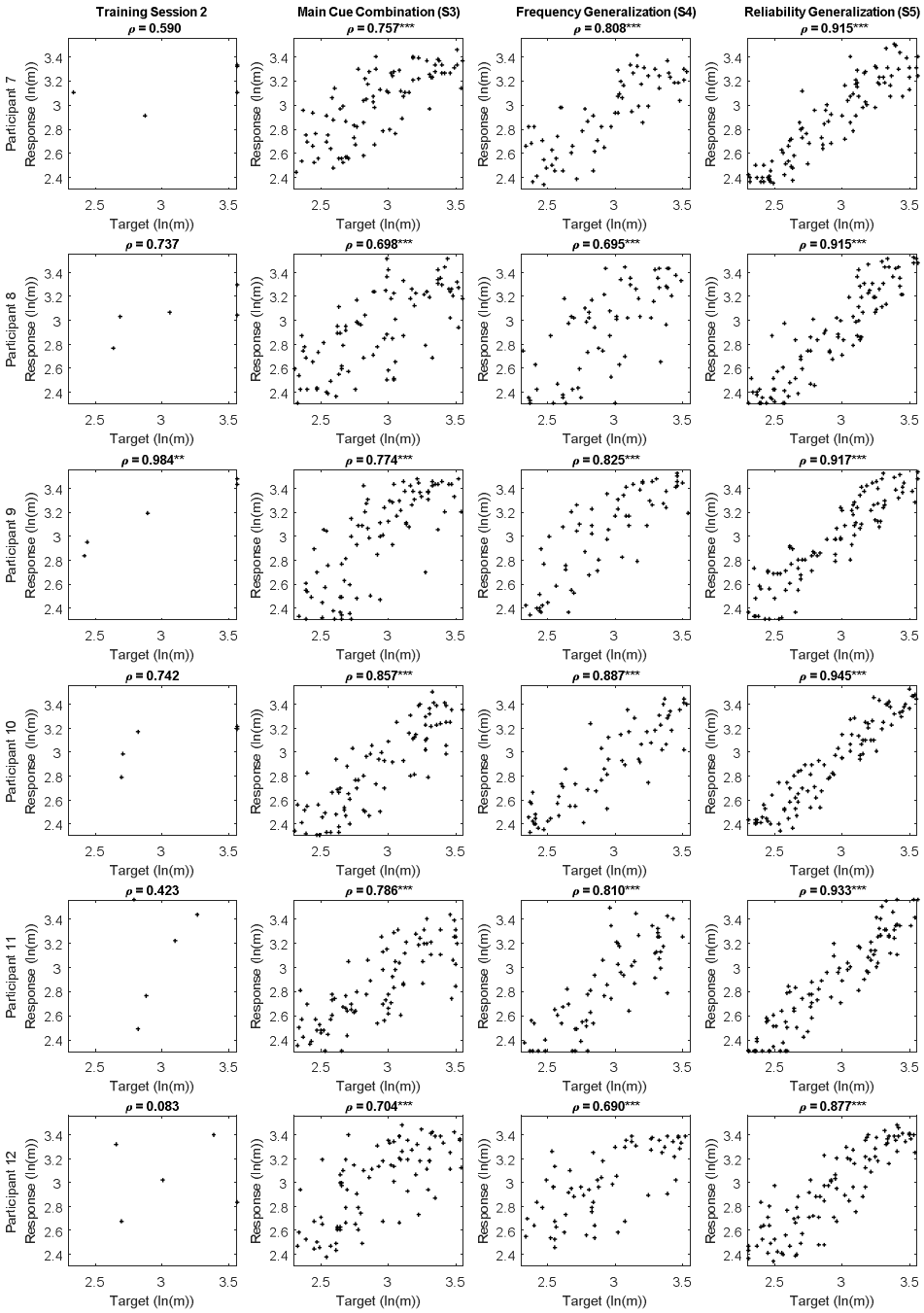

Figure S6 Continued. (Participants 7-12)

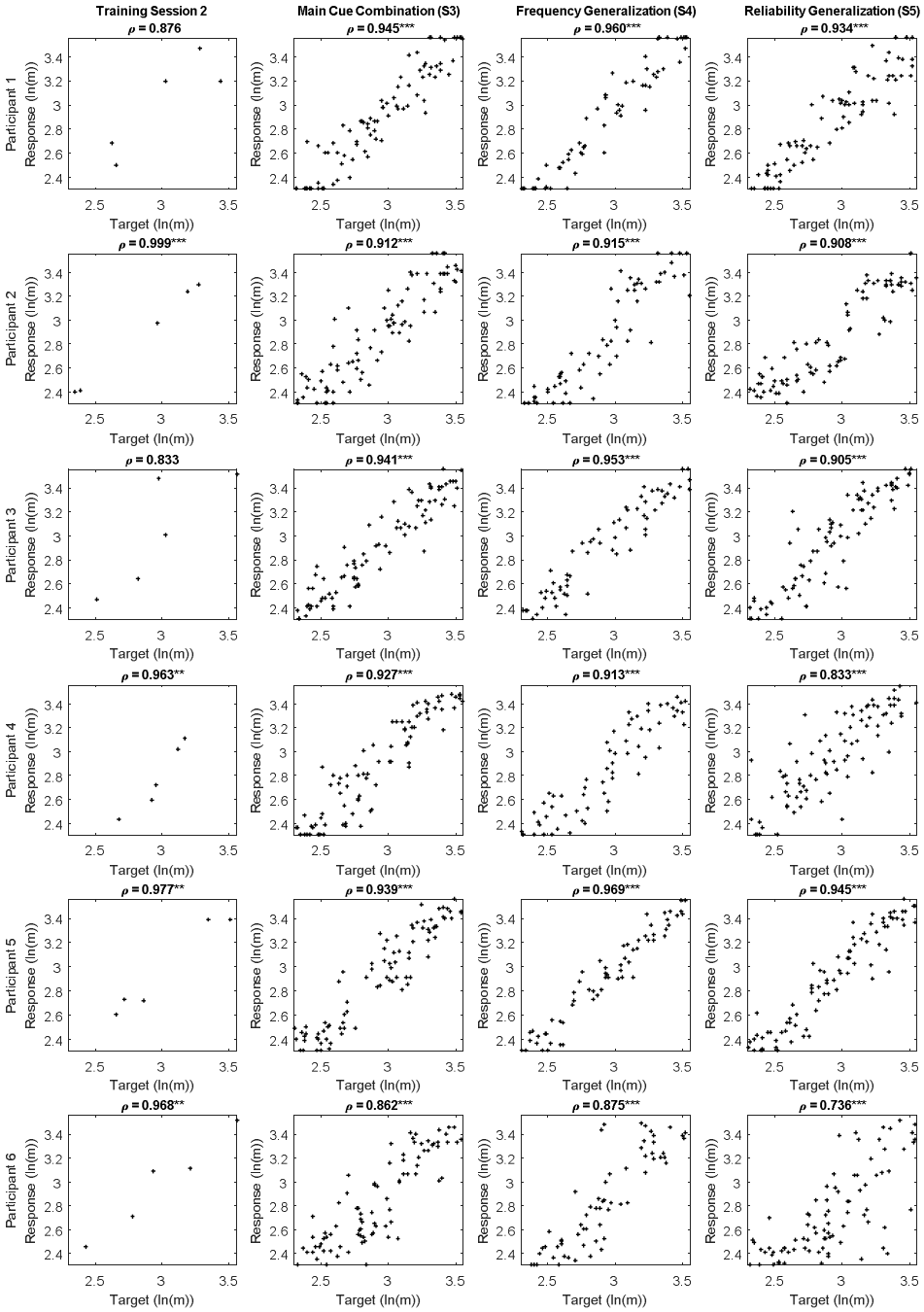

Figure S7. Targets versus responses for the bimodal trials.

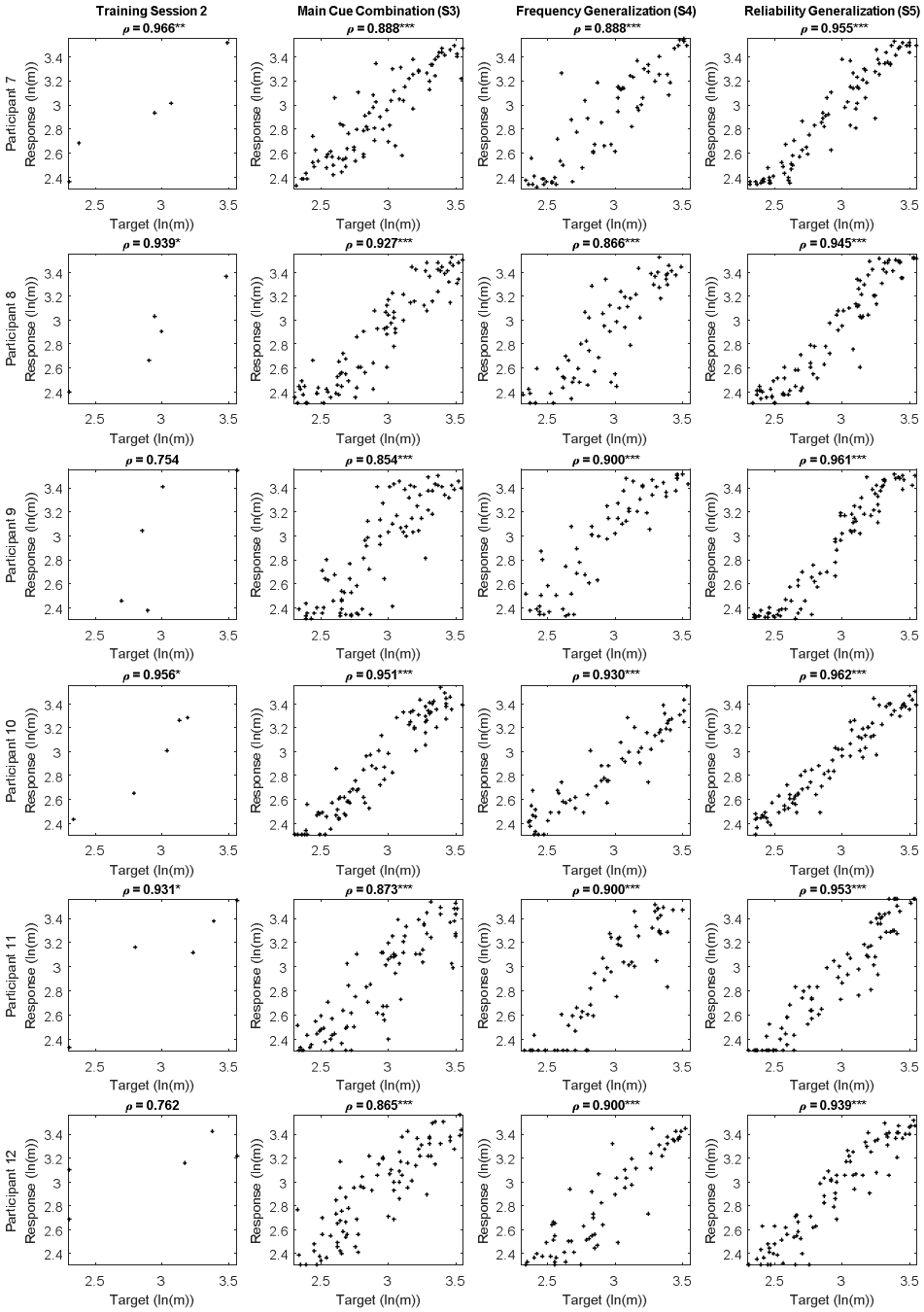

Figure S7 continued. (Participants 7-12)