Multisensory perception and decision-making with a new sensory skill

James Negen[1*], Laura-Ashleigh Bird[2], Heather Slater[3], Lore Thaler[3] & Marko Nardini[3]

[1]Liverpool John Moores University, School of Psychology

[2]University College London, Institute of Cognitive Neuroscience

[3]Durham University, Psychology Department

*Correspondence: j.e.negen@ljmu.ac.uk

Author Note. Data, code (including analysis code), and methods are available at https://osf.io/wzan2/ .

In press, Journal of Experimental Psychology: Human Perception and Performance – accepted January 2023

ABSTRACT

It is clear that people can learn a new sensory skill – a new way of mapping sensory inputs onto world states. It remains unclear how flexibly a new sensory skill can become embedded in multisensory perception and decision-making. To address this, we trained typically-sighted participants (N=12) to use a new echo-like auditory cue to distance in a virtual world, together with a noisy visual cue. Using model-based analyses, we tested for key markers of efficient multisensory perception and decision-making with the new skill. We found that twelve of fourteen participants learned to judge distance using the novel auditory cue. Their use of this new sensory skill showed three key features: (1) it enhanced the speed of timed decisions; (2) it largely resisted interference from a simultaneous digit span task; and (3) it integrated with vision in a Bayes-like manner to improve precision. We also show some limits following this relatively short training: precision benefits were lower than the Bayes-optimal prediction, and there was no forced fusion of signals. We conclude that people already embed new sensory skills in flexible multisensory perception and decision-making after a short training period. A key application of these insights is to the development of sensory augmentation systems that can enhance human perceptual abilities in novel ways. The limitations we reveal (sub-optimality, lack of fusion) provide a foundation for further investigations of the limits of these abilities and their brain basis.

Keywords: sensory augmentation; multisensory perception; human echolocation; Bayesian models; perception.

Public Significance Statement: Human perception and decision-making has a variety of ways of adapting to sensory substitution and augmentation systems. This article shows that people can use them in a coordinated way with existing perception, increasing both the speed and precision of decisions. There is scope to explore using these systems further for applications such as augmented sports or increasing workplace safety.

Multisensory perception and decision-making with a new sensory skill

New technological and scientific advances offer opportunities to substitute or augment human perception and decision-making (Maidenbaum & Abboud, 2014). Devices such as those translating distance to sound (Maidenbaum et al., 2014) illustrate the scope for tuning human perception and decision-making to alternative or novel sources of information about the surrounding environment. The use of sensory augmentation not only raises, but also enables us to investigate, the fundamental question of how flexibly human perception and decision-making are organized. There is especially a gap in our understanding of how new skills will function when they are embedded in a multisensory context – for example, using a novel auditory cue to distance together with existing visual distance perception. Understanding how people adapt to incorporating new information sources into their perception and action can both guide development of new technologies to enhance human capabilities and provide us with key insights into the organization of perception and decision-making.

We frame this work as the study of people's abilities to work with a *new sensory skill*. We define the term *new sensory skill* to mean a new ability to use information via an available sense – a learned mapping between sensory events and states of the external world. Every time someone learns to use a sensory substitution or augmentation system, they have to acquire a new sensory skill. The new sensory skill is a learned connection between sensory inputs and inferred states of the world. Echolocation provides a useful example for illustrating what we already know about new sensory skills – their potential acuity and their adaptive neural implementation. Echolocation is a technique of using reflected sound (e.g. from mouth clicks) to infer the spatial layout and material properties of objects in the surrounding environment (Kolarik et al., 2014; Thaler & Goodale, 2016). Expert echolocators can discriminate between a disc that is 50cm away versus 53cm away on 75% of trials

(Thaler et al., 2019), discern changes in the angular/azimuthal position of objects that are as small as 1 degree of acoustic angle (Teng & Whitney, 2011), and discern distance changes of only 2cm at 200cm distance (Thaler et al., 2019). These behavioural abilities are associated with brain plasticity in regions that underlie perception. For example, the ‘visual’ (occipital) cortex of expert echolocators responds to spatialized echo sounds (Thaler & Goodale, 2016) in a way that mimics retinotopic organization (Norman & Thaler, 2019). This background of high-level skills leads to an exciting scope for further research involving new sensory skills in more diverse settings. In the present study, we used a novel (echolocation inspired) sound cue for judging distances to investigate the manner in which typically-sighted people may gain and use new sensory skills.

Our interest is not only in whether a new skill is acquired (e.g. usable above chance), but also in the extent to which it may become embedded in typical flexible perception and decision-making. Therefore, we focus on multisensory tasks, multisensory interactions, and dual-task paradigms – situations in which typical human perception and decision-making show key hallmarks of efficiency and flexibility. Multisensory interactions play a central role in the organization of everyday perception and decision making. For example, vision and sound are combined to localise objects (the ventriloquist effect; Alais & Burr, 2004), and brief sounds and flashes are combined to count events (the sound-induced flash illusion; Shams et al., 2002). When signals from different modalities are combined to drive decision-making, this combination of estimates provides some key advantages such as the reduction of sensory uncertainty via Bayesian principles (Ernst & Banks, 2002; Maloney & Mamassian, 2009). To understand how new sensory skills are integrated and coordinated with existing sensory skills, we test for multisensory interactions of this kind. This allows us to gain a better understanding of how flexibly perception and decision-making are organized, and how well newly-learned skills can be embedded within typical perception and decision-making.

The present study builds on previous results with new sensory skills in multisensory tasks in several ways to learn more about how flexibly perception and decision-making adapt to new sensory skills. Our previous results indicate that participants can learn to use a new sensory skill to enhance precision with only a few hours of training (Negen et al., 2018). Participants learned to use either an echo-like auditory cue or a vibrotactile cue to judge distance to a target that was sometimes also indicated visually. A crucial result was that when the new signal was available together with vision, precision was enhanced (uncertainty was reduced) in line with Bayesian principles, although not to the full extent predicted for an ideal perceiver. Here we use the same new sensory skill to examine several crucial outstanding questions about the underlying mechanisms.

The overall goal was to answer a series of specific critical questions to illuminate mechanisms that could possibly drive improved precision with new sensory skills in multisensory tasks (e.g. Negen et al., 2018). How is this precision increase accomplished? We examine the possibility that participants use verbal resources to reason explicitly as their method of using the new sensory skill and combining it with vision; in layman terms, we ask if they ‘talk themselves through’ the task. Are there other multisensory benefits as well? We examine the possibility that the new sensory skill can also augment the speed of easy decisions in a timed task. Does this alter the early processing of the visual cues? We examine the possibility that the new sensory skill is forced to fuse with a visual cue, much like multiple visual cues to depth become forced to fuse during development (Nardini et al., 2010). Does the precision increase depend on the unusual visual cue that was used in the previous study? Previous results on the use and combination of new signals in navigation and sense-of-direction tasks are mixed (Goeke et al., 2016; König et al., 2016; Nagel et al., 2005; Weisberg et al., 2018), so it is important to make sure that this previous result can be robustly replicated in variations of the basic task. As a whole, the study is designed to examine if perception and decision-making adapt to new sensory skills in key multisensory functions and mechanisms.

As a testbed, this study uses a new sensory skill inspired by echolocation (Kolarik et al., 2014; Thaler & Goodale, 2016). Participants hear two identical clicks in series. A longer delay between two auditory clicks indicates that the target is further away. After guessing, participants have visual feedback of the target's true distance. This is less complex than real echolocation, because it does not involve any change in amplitude or power spectrum between the emission and the echo, the potential for different materials or shapes to reflect sounds differently, nor any variation in the emission itself (Zhang et al., 2017). This of course means the information is relatively restricted. However, it makes for a tightly controlled model system, in which participants must learn to map a single cue (auditory delay) to a single parameter (distance). It also parallels the case of a device for sensory substitution/augmentation that translates a single physical property to a sensory cue – for example, the EyeCane (Maidenbaum et al., 2014) or even the simple audio cue that many cars provide to assist with parking while reversing. In our study, participants are trained to judge distances with a new auditory delay cue. In other words, they are trained to use a quantitative mapping from specific features in the time domain to specific features in the space domain (rather than a qualitative, likely existing ability to generically map two magnitudes). This training backdrop is used to test six specific hypotheses that will shed light on the nature and utility of multisensory interactions with new sensory skills.

Hypotheses

In the following sections, for each of the six hypotheses, we include four paragraphs that state (a) what the hypothesis is and a sketch of how it is tested; (b) why this is important; (c) why one might expect the hypothesis to be true; and (d) why one might expect the hypothesis to be false.

Cue Learning Hypothesis (Preliminary)

This hypothesis states that participants will learn the new sensory skill – that they will acquire a useful mapping between audio stimuli and target distances. This is tested by looking for a simple correlation between target distance and response distance in trials where only the novel audio cue is present.

This is an important prerequisite for testing further hypotheses. It would not, for example, be sensible to test if the new sensory skill enhances the speed of decision-making unless we know that they can use the new sensory skill in the first place.

This hypothesis is highly plausible because previous research has demonstrated that untrained adults can quickly learn this specific mapping (Negen et al., 2018).

We have no specific reason to think that many participants will fail to learn the mapping, but there are reports of individual differences in this area (Thaler et al., 2014), so it is possible that some will fail.

Resistance to Dual Task Interference Hypothesis

This hypothesis states that processing of the new sensory skill, either on its own or alongside an existing (visual) sensory skill, is largely non-verbal and thus resists interference from a simultaneous verbal task. Verbal processing would mean that participants covertly used language (‘in their heads’) to reason about the distance of an object based on the auditory signal. If the processing of the new sensory skill is verbal, then it should show interference from additional tasks that interfere with verbal processing. The simplest way to interfere with verbal processing is to layer on an additional verbal task (Wickens, 2002), which is especially effective if both rely on audio stimuli rather than visual (ibid). Verbal working memory tasks are known to interfere with a wide variety of tasks that participants may want to ‘talk through’ (reason about verbally), such as a difficult Tetris board (Epling et al., 2017). This hypothesis is tested with a dual-task paradigm; we asked participants to judge echoic distances in the presence or absence of a simultaneous digit span task with audio stimuli. If participants use the new sensory skill non-verbally, the simultaneous task should not interfere with the ability to judge distance.

This is important because it reflects the usefulness of new sensory skills in a wider variety of settings: non-verbal processing leaves verbal resources free for other uses. New sensory skills would be much less useful if they require explicit verbal resources. This would mean, for example, that a person might not be able to navigate with the new sensory skill and also hold a conversation while out for a walk. It would also imply that the use of the new sensory skill is relatively slow, which impacts the usefulness of the skill in many circumstances.

The best reason to think this might be true is because it would match the subjective descriptions of human echolocation given by expert echolocators. Note the lack of a slow verbal processing aspect in this description from Daniel Kish (Hurst, 2017):

> “It’s flashes. You do get a continuous sort of vision, the way you might if you used flashes to light up a darkened scene. It comes into clarity and focus with every flash, a kind of three-dimensional fuzzy geometry. It is in 3D, it has a 3D perspective, and it is a sense of space and spatial relationships. You have a depth of structure, and you have position and dimension.”

This hypothesis might be false because the new sensory skill, lacking years of practice, may still require verbal support to retain its precision – especially in a multisensory context. This would fit with a number of other differences between expert and novice echolocators (Thaler & Goodale, 2016).

Redundant Signals Hypothesis

This hypothesis states that a new sensory skill enhances the speed of multisensory decision-making. In other words, a new sensory skill can be used to re-create the classic redundant signals effect (Hershenson, 1962). This is tested by giving participants a very easy decision to make, giving them two potential targets that are far apart and asking them to indicate which of them corresponds to a stimulus. We then examine whether they are faster with both the new sensory skill and an existing visual stimulus together (i.e. audio-visually) than with either single stimulus alone (i.e. only audio or only visual). Please note that our interest here is specifically in the redundant signals effect and specifically not in the Miller Race Inequality (Miller, 1982); we are first interested in whether a new sensory skill can create a speed increase and leave aside whether any possible speed increase reflects co-activation.

This is important because many realistic uses of perception and decision-making rely on rapid decision-making for their effectiveness. In everyday interactions with objects and environments – from handling objects, to crossing roads, to playing sports – speed is crucial. Efficient use of the new skill in a speeded context is therefore another important test.

This hypothesis is plausible because new sensory skills can be used to do things like navigate through simple mazes (Maidenbaum et al., 2014), which could imply a fairly low reaction time. If that is the case, it is plausible that the new sensory skill will at least ‘race’ the existing visual skill and create an average decrease in reaction time.

This hypothesis might be false because it remains unknown if the use of a new sensory skill happens fast enough to possibly provide a benefit. It is therefore possible that the visual system will always finish its processing before the new sensory skill and control the response. It is also possible that the new sensory skill will not be able to take control away from the visual skill even if it finishes first. Either case would reduce any speed gains to zero.

Forced Fusion Hypothesis

This hypothesis states that the new sensory skill and an existing (visual) sensory skill will become subject to forced fusion. This means that participants will lose (some) access to the perception of distances indicated by each individual cue and instead only have access to their combined perception of the distance. We test this using a standard oddball task, where participants are given three stimuli and asked to indicate which is different from the other two. This method compares congruent versus incongruent stimuli with the specific prediction that forced fusion should lead to worse performance in the incongruent case (Hillis et al., 2002). This also links at a theoretical level with causal inference theory, which provides an account of when and why perceptual estimates are averaged together (Shams & Beierholm, 2010).

This is important because, if true, it would suggest that the new sensory skill – a learned cue to depth – was already being fused with other sensory information at an early stage in the process of sensation, perception, and decision-making. This would suggest a more specific explanation for the increase in precision, paving the way for a more mechanistic understanding of the basic phenomenon.

This hypothesis is plausible because other depth cues do eventually become partially-fused during late childhood (Nardini et al., 2010) and the mechanism for creating this is still largely unclear. Previous research suggests such an effect can possibly be induced by simple training with a new correlation between two signals (Ernst et al., 2007), so we may observe something similar here.

This hypothesis might be false because 10 training sessions might not be enough to reshape perception in low-level sensory areas, and multisensory interactions might instead reflect processing at later decision-making stages (Rohe & Noppeney, 2018). It is also notable that a concurrent project has largely failed to find fusion indicators (Witzel et al., 2021).

Precision Hypothesis

This hypothesis states that participants will use the new (auditory) sensory skill together with an existing (visual) sensory skill to gain precision. In other words, responses will be more precise when participants are asked to estimate a distance with both cues (audio-visually) than with either cue alone (only audio or only visual). This is predicted by maximum likelihood models of decision-making (Ernst & Banks, 2002) as well as full Bayesian frameworks of decision-making with explicit priors and gain optimization (Maloney & Mamassian, 2009). This is tested by asking participants to make distance estimates with the (auditory) new sensory skill, an existing visual skill, and both together. The precision of the best single cue is then compared to the precision with both cues.

This hypothesis is important because precision gains via cue combination are a hallmark of perception and decision-making with the native senses (Alais & Burr, 2004; Burr & Gori, 2011; Körding & Wolpert, 2004; *but see also* Rahnev & Denison, 2018). Perception and decision-making take place in noise (under uncertainty) – Bayes-like cue combination provides a way to reduce this noise and so make decisions more efficient. For a new sensory skill to participate efficiently in perception and action, it should join this multisensory process. An important consideration is the kind of noise (uncertainty) in the task. The present study modifies a previous study that used *external noise* in the existing visual skill (Negen et al., 2018) to use *internal noise* instead. With internal noise (only), the signal itself is in theory perfectly reliable, but there is noise (imprecision) in the process of perception (Macmillan & Creelman, 2004). This internal noise could take many forms including an improper understanding of which visual aspects are relevant and how they are calibrated; the point is that a perfect observer would achieve perfect precision. In contrast, external noise is when the signal itself, even if processed perfectly, indicates a world state that varies around the correct value. Most noise in everyday environments is internal. It is therefore important to check whether our previous results (Negen et al., 2018), finding a precision increase when combining a new sensory skill with an existing visual skill with *external noise*, are also found in when the visual skill has *internal noise*.

This hypothesis is plausible because this section of the experiment is much like a previous experiment where a precision increase was found (Negen et al., 2018). The change from external to internal noise has no particular bearing on the underlying mathematical foundations of cue integration – either can be used to create the exact same likelihood function – so it is possible that participant behaviour will also be similar.

The biggest reason to think this hypothesis may be false is that results from previous studies examining precision gains via cue combination have been mixed (Goeke et al., 2016; Negen et al., 2018). However, there are specific methodological choices that could have a major impact. The study that showed a precision increase (Negen et al., 2018) used a cue with external noise and trained participants on its use. The study that did not show a precision increase (Goeke et al., 2016) used a cue with internal noise and did not train participants on its use. It is not clear which of these two (or another aspect) controls the result. The present study uses internal noise and trains the participants. It will therefore clarify if internal noise somehow resists precision increases.

Optimal Weight Hypothesis

This hypothesis states that participants will not only gain some precision given both cues (the new sensory skill and an existing visual skill) versus any single cue, but the precision gain will approach the optimal prediction. In theory, if two cues are subject to noise that is Gaussian and the noise is not correlated across the cues, then it is possible to take a weighted average of them in a manner where the resulting precision (1/variance) is the sum of the two individual precisions. This will be tested by asking participants to make distance estimates with the new sensory skill (audio only), an existing visual skill (visual only), and both together (audio-visual). The precision in the audio-visual trials will be compared against the sum of the precisions in the single-cue trials. Further, the weight given to each cue will be estimated and compared with the weight that results in the optimal noise reduction.

This is important because it would imply that a new sensory skill is not only used to increase precision, but that perception and decision-making are so good at adapting to the use of new signals that they achieve the highest possible efficiency. This would indicate a very high degree of flexibility in perception. It would also point towards potential scope for new sensory skills to be maximally effective (in a certain sense) for augmenting multisensory perception and decision-making.

The main reason to think this may be true is simply because perception and decision-making achieve such near-optimal levels of noise reduction in many other tasks (Alais & Burr, 2004; Clark, 2013; Knill & Pouget, 2004; Körding & Wolpert, 2004, 2006). This ability has to be learned during development for existing sensory skills (Burr & Gori, 2011; Nardini et al., 2010) and it is possible that adults can re-create that learning episode with a new sensory skill. In addition, near-optimal visual integration has been demonstrated by adults who have surgically regained sight after prolonged deprivation (Senna et al., 2021) – this suggests that adults retain their ability to learn near-optimal integration.

The main reason to think this may not be true is that normal optimal combination takes many years to develop. For example, children mis-weight visual vs haptic cues to object size compared with an ideal observer (Gori et al., 2008). Our earlier study (Negen et al., 2018) also showed less-than-optimal precision improvements with a new sensory skill.

Method

This study was not pre-registered.

Summary

Because the design is extensive and detailed, we first summarize it here. This summary will give enough information to understand the results, while later sections give all details required to replicate the study.

Throughout the study, a virtual cartoon whale named Patchy was presented in an immersive 3D environment and gave instructions, encouragement, and feedback to participants (see Figure 1, right panel). Judging the distance to Patchy using one or more sensory cues was the main task on each trial. He could hide under the virtual sea along a line that stretched out in front of the participant, who was sitting on a chair at the very front of a boat. As a baseline, it is helpful to look at an AV trial (audio-visual; Figure 1). The participant is looking at the line in front of the boat. The goal is to estimate the distance to the non-visible target (Patchy). At the same time, they hear a pair of clicks (auditory cue) and see a cluster of blocks distributed along the line (visual cue). The auditory cue is informative about Patchy's position because the delay between the first click and a second click is proportionate to distance (like an echo). The visual cue is informative about Patchy's position because the blocks move (contract) towards a convergence point which is Patchy's location (see Figure 1). During 250ms (overlapping with the audio delay), the blocks move 20% of the way towards the target, then disappear. The participant uses an Xbox controller to control a marker which they use to indicate their estimate of Patchy's distance along the line. When they have moved the marker to the correct location, they select the location by pressing a button. Patchy then surfaces at the correct location, serving as visual feedback of the correct location. He also gives feedback in the form of a percentage expressing their degree of under / overestimation. This completes an AV trial. In summary, we first trained participants to estimate distance using only the novel auditory stimuli. We then introduced the visual stimuli to the trials. Finally, we carried out several variations on these basic trial types to test specific hypotheses.

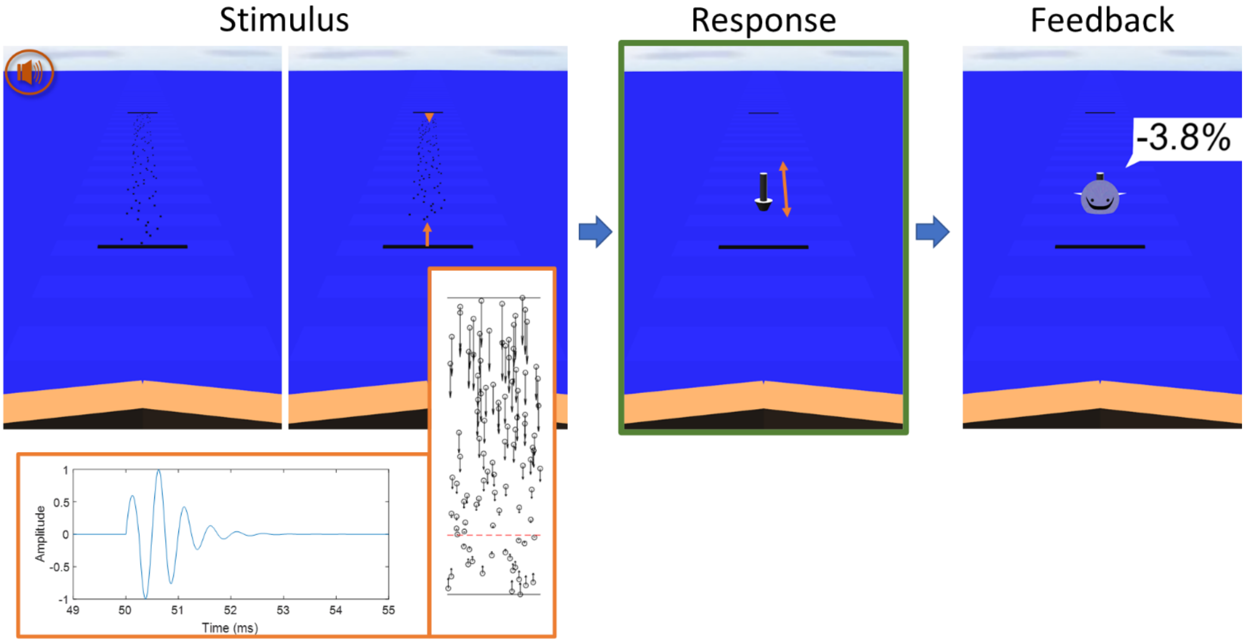

Figure 1: Example AV (audio-visual) trial. Everything in orange here (audio icon, orange arrows, orange-bordered inserts) was not seen by the participant and is edited on top for the reader's clarity. The participant is seated on the top deck of the boat. Looking out over the front, the participant sees two black bars that mark the range of potential locations for the target. They hear an audio stimulus, two clicks in series where the delay signals distance. They also see a distribution of black blocks. The blocks all move 20% of the way towards the target over 250ms. In the "Stimulus" views (left), the orange arrows show the general movement direction in the 3D space, compressing inwards. The left orange-bordered insert shows the waveform of a single click. The right orange-bordered insert illustrates a flow field for this target, i.e. the movement of each block for this target. During the "Response" (next panel), the participant then controls a grey 3D marker which they can move between the black bars. They can move the marker out or in, but not left or right (along the direction marked in orange). When they press A and enter a response ("Feedback" panel), Patchy appears at the correct location and provides feedback in terms of percent error. To be as clear as possible: the black bars, the black blocks, the gradient in the sea, the grey 3D marker, the cartoon whale, and the cartoon whale's speech bubble are all seen by the participant in stereoscopic 3D. See supplemental movie for a few examples of the visual stimulus and its

longer variant (though without disparity or headset tracking, at a lower framerate, and at a lower resolution).

The experiment involved a total of ten sessions, each lasting around 1 hour and carried out over a 2- to 10-week period. The first two sessions gave participants practice with the audio cue and the visual cue separately. Training followed a scaffolding approach, in line with our previous learning paradigm (Negen et al., 2018). Specifically, training with the audio cue began with initial trials made as simple as possible: 2-alternative forced-choice judgments for a 10m vs 35m target distance. The number of options was then increased via a 3- and then 5-alternative forced-choice design. This progressed to continuous judgement trials in which participants can use the joystick to respond anywhere along the line. Continuous responses in the second session are analysed to test the Cue Learning Hypothesis (Preliminary).

In session 3, AV trials were introduced. These were accompanied by Audio trials and Visual trials (like AV but with only one cue). This allows for a test of the Precision Hypothesis. Session 4 repeated session 3. From session 5 onwards, there were variations to test more hypotheses. Session 5 asked participants to do Dual Task trials, in which they had to do an Audio/Visual/AV trial while remembering a string of six verbal digits. This was for the Resistance to Dual Task Interference Hypothesis. Sessions 6 and 7 had perturbation trials, in which the audio and visual cue were offset by 10%. This allowed us to measure the relative weight given to each cue (through multiple regression) and test the Optimal Weight Hypothesis. Session 8 was the oddball task (Figure 2, top row), in which participants were shown the stimuli for 3 AV trials and asked to indicate which one was different from the other two. This was for the Forced Fusion Hypothesis, which specifically predicts that performance will be better when the oddball stimulus is congruent than incongruent (i.e. easier when the visual cue and audio cue both deviate in the same direction). Session 9 was a speeded task (Figure 2, bottom row). Instead of making a slow and careful judgement, participants were given two possible distances and asked to rapidly indicate which of the two was correct. This tested the Redundant Signals Hypothesis. Finally, in session 10, there were dual task control trials. In these, participants completed the verbal working memory task without having to judge any distances. This was done to further test the Resistance to Dual Task Interference Hypothesis.

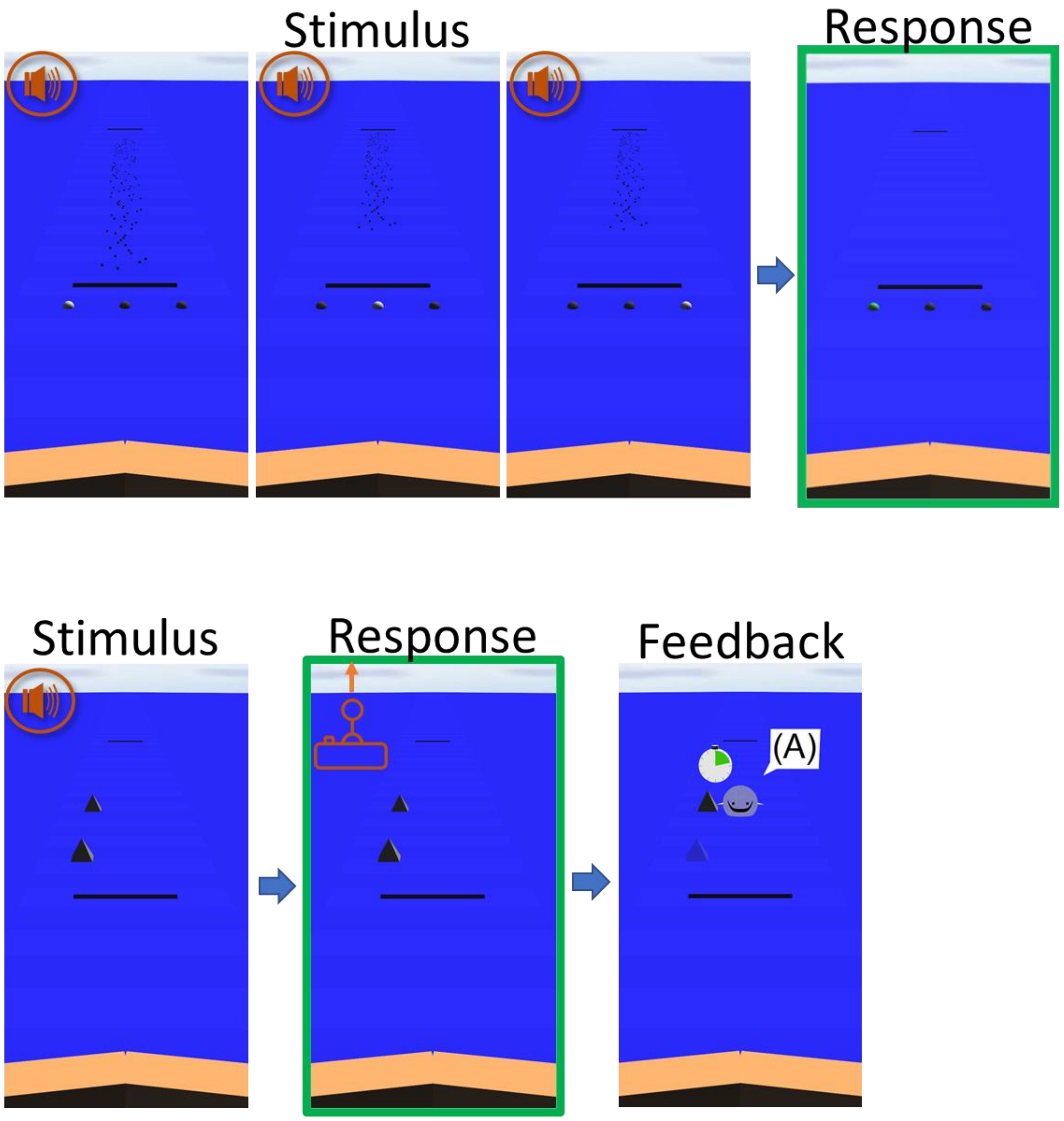

Figure 2: Oddball (top) and Speed Task (bottom) example trials. Everything in orange is drawn on top of the screenshots for the benefit of the reader (audio icon, joystick icon). For oddball, the participant is presented with three sets of AV stimuli, each matched with a sphere changing from black to white. Two stimuli are the same and one is different. The task is to indicate which AV stimulus was different (here, the first stimulus; blocks are shown at the final frame) by highlighting the sphere corresponding to the stimulus with the controller (here, the left sphere). There is no feedback. For the speeded task, the participant is given two possible choices. As fast as possible after the stimulus (trial shown here is audio-only, but there were also visual and AV), the participant pushes the joystick up or down to indicate the nearer or further option. If the choice is correct, Patchy appears and shows them their reaction time on a 3-second clock.

Participants

Twelve participants (4 male) took part in this study with an average age of 23.8 years (range 20 to 34). An additional two participants were excluded from analysis due to the fact that they failed to learn the audio cue by the second session (female, 20; male, 19). One participant was only able to provide partial data (6 of 10 sessions) due to social distancing measures introduced by the government in March 2020 to counteract the spread of Covid-19. Participants were recruited by word of mouth around Durham, UK. They were paid £10 per session, totalling £100 of compensation. This study was approved by Durham Psychology's Ethics Board (Reference: PSYCH-2018-12-04). Informed consent was given by participants in writing.

Power calculations suggest that power is over 95% for the anticipated effects and this exact design. The present study is interested in effects that tend to be unusually large by Psychology standards: cue learning for these stimuli, Spearman's rho > 0.80 (Negen et al., 2018), correlation of log-response and log-target distance; dual task interference with verbal skills, d = 1.88 (Epling et al., 2017), paired t-test of single-task versus dual-task performance; redundant signals effect, d = 2.75 (Miller, 1982), paired t-test of best single cue versus redundant cues reaction time; forced fusion in adulthood, d = 1.65 (Nardini et al., 2010), paired t-test of congruent versus incongruent discrimination thresholds; cue combination with a new sensory skill, d = 1.1 (Negen et al., 2018), paired t-test of best single cue versus both cues variable error; and weight changes alongside a change in reliability, d = 1.2 (Negen et al., 2018), paired t-test of lower versus higher reliability visual weights. GPower shows the power to detect r = 0.80 with 210 trials is over 99.99%. For the remaining hypotheses, the smallest d value of 1.1 with 11 participants in a paired t-test gives 95.9% power.

Apparatus

Virtual environment

A custom seascape was created in WorldViz Vizard 5 (Santa Barbara, CA, USA) and presented using an Oculus Rift headset (Consumer Version) (Menlo 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). Patchy was 2m long and 1m wide. Participants were seated on the ship and 4.25m above the sea surface so that different depths of the sea plane had more vertical differences in the projection from their viewpoint. 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 colour 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 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 (example, Figure 1, far right panel). The sea surface remained still, and the ship did not move. The major advantages of using virtual as compared with real objects was that trials could move much faster, and that we could have control over reliability of visual cues. Full VR via a head-mounted display (rather than a smaller screen) allowed us to immerse people in the virtual environment, avoiding any conflicting spatial information about the real surrounding lab space.

Different trial types also had different kinds of response mechanisms in the virtual world as appropriate. For continuous judgement, there was a grey 3D marker pointed downwards towards the sea (Figure 1). Participants could adjust this along the response line and then enter a response by pressing the A button. For 2-alternative forced-choice (2AFC), 3AFC, and 5AFC trials, the marker would ‘snap’ to the discrete possibilities. For the digit span sections, participants needed a way to enter a recalled digit in a series. Grey 3D numbers could appear 1m above the sea and 15m out (in a position where it does not occlude the response range). They were 1m tall. To fill in the missing digit, participants could use the controller shoulder buttons to make their selection. During the Oddball trials, there were three grey spheres that were 9m away from the ship (Figure 2, top). While the first stimulus played, the left sphere was white; the middle during the second; the right during the third. To respond, participants could use the controller shoulder buttons to select one of the three spheres and then press the A button. For the speed trials, participants would see two small pyramids along the response line (Figure 2, bottom). To indicate that the target was located next to the further pyramid, participants moved the joystick upwards; for the nearer, downwards. As soon as a response was entered by moving the controller joystick, the selected pyramid would grow slightly larger and the non-selected one would turn 50% transparent.

The speed trials also had a special virtual feedback object (Figure 2, bottom). It was a set of shapes arranged to look like a small stopwatch that floated over Patchy’s head. The maximum time on the clock was 3s. There was a patch on the front that could cover an appropriately sized slice of the clock and change colour. The colour shade varied from red to green based on the time (more red meaning longer).

Headset

The Oculus Rift headset has a refresh rate of 90 Hz, a resolution of 1080 × 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.

Audio Equipment

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. Volume was set to a comfortable but clear level and remained constant across all testing sessions and participants (approx. 60dB SPL).

Controller

Participants used an Xbox One controller (Redmond, WA, USA).

Stimuli

Audio

This was a pair of short 'clicks' where the time between clicks signals the distance to the target. See Figure 1 for an image of the waveform of a single click. The audio stimuli were created by first generating a 5ms sine wave 2000 Hz in frequency with an amplitude of 1 (in effect, 10 periods or 20 half-periods). The first half-period of the wave was scaled down by a factor of 0.6. The next full period had an amplitude of 1.0. An exponential decay mask was created starting after 1.5 periods and ending at 5ms. To be specific, the amplitude during the last 17 half-periods was multiplied by the function e^x. The variable x started at 0 and decreased linearly to -10 (and thus the amplitude started at 1 and decreased exponentially to .0004). These exact choices are somewhat arbitrary but they serve to create a strong 'click', an audio cue with a sharp attack that is relatively easy to temporally localize. 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 10 m, 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 of the information the participants were using. Our stimuli also allowed us to use a range of distances at which real echoes are typically very faint, minimizing the scope for participants to have prior experience with them. For the purposes of reaction time, timing began as soon as the 50ms initial delay ended and an actual sound began playing.

We view this stimulus as *new* or *learned* in the sense that naïve participants are unable to use it to perform better than simple degenerate strategies like always pointing at the center of the response range (Negen et al., 2018). This is likely because the use of this stimulus requires a new mapping from (parts of) the time domain to (parts of) the spatial domain.

Visual

This stimulus is essentially a type of coherent motion or motion integration stimulus. There were 100 black blocks with a side length of 0.05m. From the participant's perspective, the x axis is left/right and the z axis is near/out. At the beginning of the stimulus, the boxes were spread evenly from 10m to 35m along the z axis. They were placed uniformly randomly from -0.5m to 0.5m along the x axis. Over the course of 250ms, beginning as soon as they appeared, each box moved 20% of the way towards the target at a constant speed. As soon as the 250ms ended, the blocks disappeared. This means that the full set of dots compresses inwards. Any inaccuracy represents internal noise, since perfectly estimating the block trajectories' convergence point (or even perfectly extrapolating the remaining 80% of any one block's trajectory) would perfectly localise the target. There is also a longer version of this stimulus, intended to make it more reliable, for certain trial types. This is the same, except that blocks move 40% of the way towards the target over the course of 500ms. For the purposes of reaction time, timing begins as soon as the blocks appear. Supplemental Movie 1 shows a few examples of the visual stimulus and its longer variant (though without disparity or headset tracking, at a lower framerate, and at a lower resolution).

This specific visual stimulus was chosen for three reasons. Piloting showed that responses around the correct target are a good approximation of a normal distribution. Participants are able to understand it very quickly, which means that additional training did not have to be added to the already lengthy design. The movement parameters (time and distance percentage) make a convenient and simple way to increase or decrease reliability, which helps with the analysis regarding cue weights. Pilot studies were also used to find a range of parameters for the visual cue that resulted in comparable precision to the auditory cue.

We view this visual stimulus as *native* or *existing* in the sense that it is already perceived as spatial by naïve participants. The visual stimulus is very easy to use without any particular training (see Supplemental Movie 1). This contrasts with the audio cue, which requires a newly learned mapping from (parts of) the time domain onto (parts of) the space domain.

Digit Span

Six digits were selected randomly. A text-to-speech voice read out the six digits in order at a rate of one per second.

Procedure

There were 10 sessions, each with up to 300 trials. Each session was done on a different day and allowed to span up to 10 weeks. Breaks were given whenever requested. The goal and types of trials were slightly different for almost every session. Figure 3 gives an overview of which trial types each session used and in what order. The following text explains each session in detail. Whenever a session uses a new trial type that has not yet been used, that trial type will be explained under a subheading.

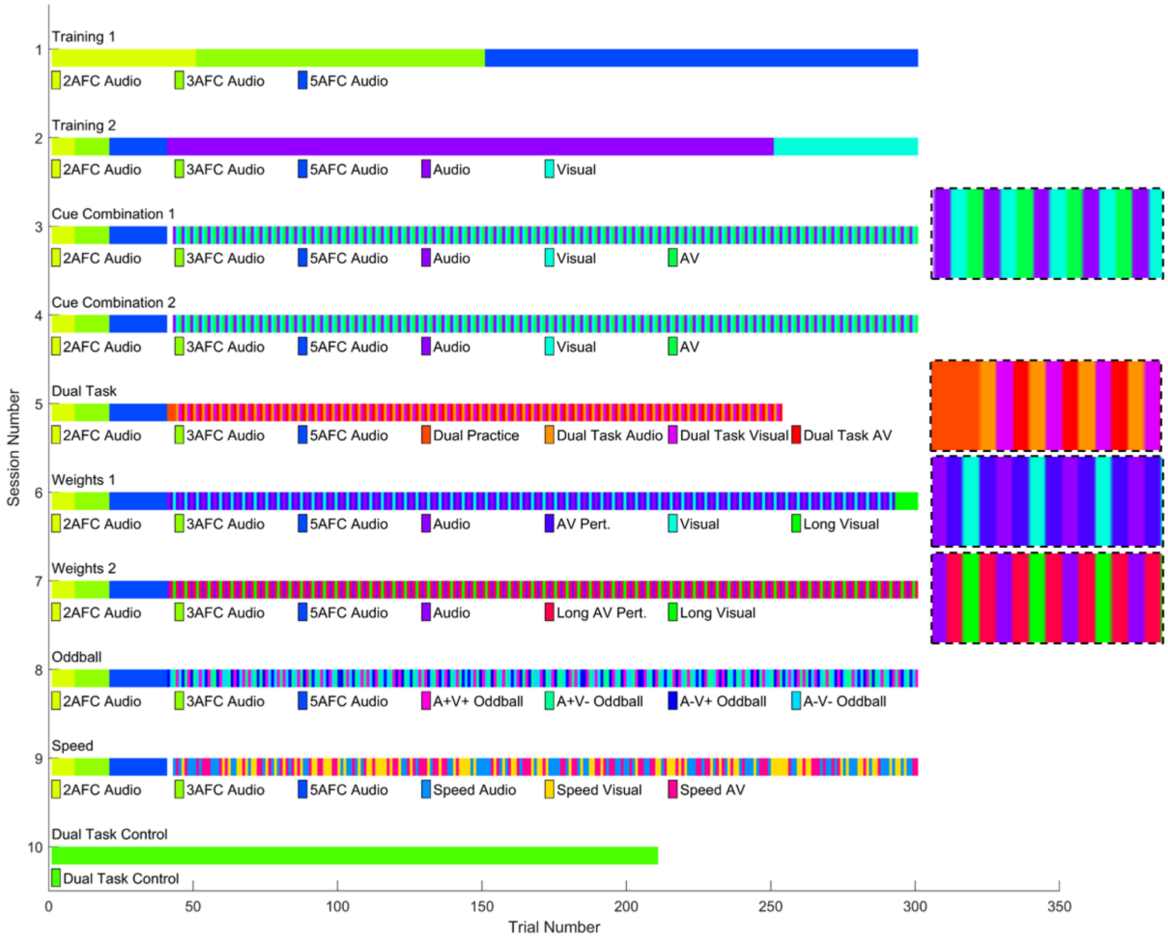

Figure 3: Reference guide for the different sessions and trial types. See *Procedure* for details of each trial type and an explanation of each session's purpose. Callouts in dashed lines contain zoomed sections from the nearest line to show the repeating pattern. (There is no callout for session 4 because it is the same as session 3. There is no callout for 8 or 9 because the trial order was random.)

Training 1 (Session 1)

The aim of the first session was to train the participants to use the audio cue. No data from this session were analysed. The session consisted of 50 trials of 2AFC Audio, followed by 100 trials of 3AFC Audio, and then 150 trials of 5AFC Audio. Targets were used as evenly as possible and in a random order.

2AFC Audio / 3AFC Audio / 5AFC Audio. 2AFC stands for two-alternative forced choice. With an audio stimulus, participants judge if the target was 10m or 35m away. 3AFC also included 22.5m. 5AFC also included 16.25m and 28.75m. Before the first 2AFC Audio trial, Patchy would demonstrate the difference by alternatively appearing 10m and 35m away while the matching audio stimulus played. This demonstration cycled six times. 3AFC demonstrated the three targets in four cycles. 5AFC demonstrated the five targets in three cycles. During testing trials, feedback was given: if the correct choice was made, Patchy would appear and make a ‘nodding’ motion. If the incorrect choice was made, Patchy would appear at the correct location and move his head left and right in a ‘no’ gesture.

Training 2 (Session 2)

Session 2 was further training with the audio cue, a test of the Cue Learning (Preliminary) Hypothesis, and an introduction to the visual stimulus. First the participant was presented with 8 trials of 2AFC Audio, followed by 12 trials of 3AFC Audio, and then 20 trials of 5AFC Audio which served as a reminder of the previous training. Targets were used evenly in a random order. For the following text, that set of 40 trials will be referred to as ‘the warmup block’ for brevity. The warmup block is not analysed here or in any session. After the warmup block, there were 210 Audio trials. Targets were spread evenly on a log scale and presented in random order. These were used to test the Cue Learning (Preliminary) Hypothesis, which predicts a significant correlation between target and response, and to give further training with the audio cue. Next there were 50 Visual trials, spread evenly on a log scale and presented in a random order. This was to ensure that the participants were adequately familiarised with the visual cue and its use before the tests of cue combination (i.e. we would not want people to fail to combine the audio and visual cue due to unfamiliarity with the visual cue). These Visual trials are not analysed.

Audio / Visual. Participants were only given one cue (the new sensory skill or the visual cue) to use. They were required to judge the distance to a target along a continuous line stretching from 10m to 35m. They used a joystick on the controller to move a marker to their estimated target location and pressed the A button. Patchy would appear and give them feedback in terms of percentage. For example, a target of 20m and a response of 18m would get the feedback “-10.0%”.

Cue Combination 1 (Session 3)

Sessions 3 to 5 were designed to test the Precision Hypothesis. This hypothesis states that variable error will be lower in the AV trials than with the best single cue (i.e. whichever of Audio or Visual has the lower variable error). These sessions also provided a basis for estimation of the optimal possible precision for the Optimal Weight Hypothesis. After the warmup block (not analysed), there were 86 Audio, 86 Visual, and 86 AV trials. They were presented in the order of one Audio trial, then one Visual, then one AV, then one Audio, and so on. The targets were spread evenly on a log scale and random in order.

AV. During AV trials, both the audio and visual stimuli are presented. For this trial type, the two always agree perfectly (i.e. there is no offset in the distances they signal) and begin at the same time.

Cue Combination 2 (Session 4)

We repeated the procedure from session 3 in session 4 in order to gather more data and provide greater statistical power for the Precision Hypothesis and the Optimal Weight hypothesis.

Dual Task (Session 5)

Session 5 served several purposes at once. The primary purpose was to test the Resistance to Dual Task Interference Hypothesis. This states in part that performance on the Audio, Visual, and AV trials is independent of verbal working memory and thus will not be impaired by a simultaneous verbal working memory task. This set of trials also provided further data for the Precision Hypothesis and the Optimal Weight hypothesis. After the warmup block (not analysed), there were three Dual Task Practice trials (not analysed). Following this, there were 70 Dual Task Audio, 70 Dual Task Visual, and 70 Dual Task AV trials. These were presented in the order of one Dual Task Audio, one Dual Task Visual, one Dual Task AV, one Dual Task Audio, and so on. The targets were spread evenly on a log scale and random in order.

Dual Task Practice. Participants were presented with six random digits (audio presentation, 1s per digit). There was then a delay of 2.0 seconds before they completed a memory probe. For the probe, they were presented with a visual display containing five of the six digits. The missing digit is replaced with a question mark. To identify the missing digit from memory, they used the controller to cycle through digits 0-9 and clicked to indicate their response. No feedback was given.

Dual Task Audio / Dual Task Visual / Dual Task AV. The participants were presented with the random digits before being given the stimulus or stimuli (i.e. digits then clicks/blocks/both). They were then required to judge the distance to a target (continuously 10m to 35m), and then complete a memory probe. No feedback was given on the memory probe. Feedback on the distance judgement was given after completing the memory probe.

Weights 1 (Session 6)

Sessions 6 and 7 were designed to test the Optimal Weight Hypothesis. This required us to estimate the actual weights placed by the participant on each cue. It also required that we estimate the optimal weight each cue should be so that optimal weights can be compared with the actual weights. After the warmup block (not analysed), there were 63 Audio trials, 63 Visual trials, and 126 AV Perturbation trials. The AV Perturbation trials allowed us to estimate the weight given to each of the two cues (and also any central tendency bias or prior on the center of the response line) via multiple regression. The Audio and Visual trials were performed in order to estimate the precision (1/variance) of responses with each single cue, which then determined the optimal weight to give each one (higher precision means more weight). These were presented in the order of one Audio trial, one AV Perturbation trial, one Visual trial, one AV perturbation trial, one Audio trial, and so on. There were 63 targets spread evenly on a log scale. Each target was used once for the Audio trials, once for the Visual trials, and twice for the visual stimulus of the AV Perturbation trials. For the AV Perturbation trials, the audio distance was +/- 10% of the visual distance, with the sign chosen randomly unless one choice would place the audio stimulus outside of the response range of 10m to 35m. Targets were presented in random order. At the end of the session, 8 Long Visual trials (not analysed) were presented to introduce participants to the new visual reliability for the next session.

AV Perturbation. The participants were given an audio cue and a visual cue that differed from each other by 10%. To be very specific, once a location was selected, the visual cue indicated that location while the audio cue signalled a distance that was plus or minus ln(1.1)=0.0953 on a natural log scale. The participant made a judgement of the distance to the target along a continuous line. No feedback was given.

Weights 2 (Session 7)

Session 7 was a further test of the Optimal Weight Hypothesis. For this session, the longer version of the visual stimulus was used. We expected this to lead to higher precision in visual judgments, and therefore to change the optimal-predicted weighting towards vision. After the warmup block, there were 65 Audio trials, 65 Long Visual trials, and 130 Long AV

Perturbation trials. The scheme for trial order and target placement mirrored Session 6. This again allowed us to estimate the actual weight and optimal weight given to each.

Long Visual / Long AV Perturbation. Participants were presented with a visual stimulus lasting twice as long as the usual Visual trials (500 vs 250ms). These were otherwise like Visual / AV |Perturbation.

Oddball (Session 8)

Session 8 was designed to test the Forced Fusion Hypothesis. The standard method for testing this is an oddball task with congruent versus incongruent trials, with forced fusion inferred if the incongruent trials are more difficult (Hillis et al., 2002). After the warmup block (not analysed), there were 65 A+V+ Oddball trials (congruent), 65 A-V- Oddball trials (congruent), 65 A-V+ Oddball trials (incongruent), and 65 A+V- Oddball trials (incongruent). These were presented in a random order. The standards were spread evenly from 15.7m to 22.3m. Targets were selected through a staircase method, described below. The was an independent staircase for each of the four Oddball trial types (i.e. 4 staircases: A+V+, A-V-, A+V-, A-V+). These staircases were designed to converge at approximately 2/3 correct (against a chance rate of 1/3).

A+V+ Oddball / A-V- Oddball / A+V- Oddball / A-V+ Oddball. A standard distance was chosen along the line from 15.7m to 22.3m. An oddball distance was generated that differed from the standard by a certain amount. The participant was presented with three sets of AV stimuli. The standard was played twice and the oddball once. The second set began 1.0s after the first and the third began 1.0s after the second to be sure they did not overlap. The task was to select the oddball, which was randomly chosen to be one of stimuli 1-3. In the instructions, this was specifically phrased as: “Two are the same. One is different. Pick the odd one out.” In an A+V+ Oddball trial, the audio and visual components of the oddball presentation both signalled a further distance than the standard. A-V- Oddball trials had both cues nearer than the standard. A+V- and A-V+ had one further and one nearer. The staircases operated on a log scale. Each staircase began at 0.2. With every correct response this was multiplied by 0.9. With every incorrect response this was multiplied by 1/(0.9^3). This was capped at 0.4. For example, suppose a 20m standard and a difference of 0.2 on a log scale. The oddball would signal a distance of e^(ln(20)+0.2) = 24.4m. No feedback was given.

Speed (Session 9)

Session 9 was designed to test the Redundant Signals Hypothesis. This hypothesis suggests that the new audio cue will facilitate faster responses alongside the visual cue than could be achieved with either single cue alone. After the warmup block (not analysed), there were 83 Speed Audio, 83 Speed Visual, and 83 Speed AV trials. These were presented in random order. Targets were spread evenly on a log scale. Incorrect alternatives differed by 40%. Since the decisions were relatively easy, the instructions instead stressed the importance of speed for these trials: “This one is about SPEED!” and “Ready? Remember, FAST!”.

Speed Audio / Speed Visual / Speed AV. The participant was shown two possible distances to choose which differed by 40%. They were marked by small pyramids appearing. After the pyramids appeared, but before the stimulus began, there was a delay with a randomized length: 0.5 seconds plus an amount drawn from an exponential distribution with a rate of 3 (mean of 1/3s). This is capped at 2.5 seconds. The stimulus (Audio, Visual, or AV) indicated one of the two possible distances. The participant was required to move the joystick upwards to indicate the further distance or downwards to indicate the nearer distance. Feedback was then given in the form of a small clock displaying their time if they were correct, otherwise, Patchy appeared and slowly shook his head in a ‘no’ gesture. Several randomly selected trials were also designated as ‘catch’ trials where the delay was increased by 2.5 seconds but the trial was otherwise the same.

Dual Task Control (Session 10)

Session 10 (along with session 5) was designed to test the Resistance to Dual Task Interference Hypothesis. These trials were designed to measure performance on the verbal working memory task without the need to make a distance judgment. Participants still used the joystick to place the marker but could now see the location of the target (Patchy) directly rather than having to infer it via uncertain audio or visual cues. There was no warmup block. There were 210 Dual Task Control trials.

Dual Task Control. Participants heard six random single digit numbers. Following a delay of 2.0 seconds, they saw Patchy appear. They then used the joystick to place the marker onto Patchy and press A. They then completed a memory probe (i.e. they were presented with the previously heard six digit sequence visually, with a question mark in place of the probe digit). They then use the controller to select the correct number to place into the sequence. No feedback is given.

Data Processing and Analysis Plan

The procedure detailed above resulted in a total of 2,857 trials for each participant. In this section, we describe how we extracted measures from this dataset and analysed them. In addition to the formal analyses described here, we also briefly examined the factor of time to complete. In short, it did not have a significant impact on audio variable error, mean reaction time, or cue weights. This is potentially due to the design details; every session began with an unanalysed warmup block that allowed participants to remember how the audio cue works. We therefore neglect time to complete as a factor in the analyses below.

Cue Learning (Preliminary) Hypothesis

For the audio trials from session 2, for each participant the target and response were transformed onto a logarithmic scale. This was done to account for Weber’s law (Getty, 1975). The data were then analysed for a correlation between the targets and responses. A positive correlation in an individual participant was interpreted as evidence for that participant learning how the audio cue works. For comparison to expert echolocators, we also estimated a weber fraction for each participant. This was be done by creating synthetic two-alternative forced-choice (2AFC) trials from the audio trials. We assumed that if an estimate when given one target is further than an estimate given another target, then the participant would have chosen the first target as further in a 2AFC task. All possible pairings were used within each participant. This method is not as ideal as actual 2AFC data but should at least give some indication of how far the overall performance is from expert performance.

Resistance to Dual Task Interference Hypothesis

For this and further analyses, we needed to be able to calculate the variable error (i.e. to estimate the amount of perceptual noise in the judgements, separated from systematic distortions). This was done with the data from sessions 3-5, using the trial types that involve a continuous judgement of distance (Audio, Visual, AV, and their Dual Task variants). To begin, we trimmed outliers. This was done by calculating the standard deviation of responses minus targets, on a logarithmic scale, for all the data being used here. This was 0.23 natural-log-meters (lnm). Any trial with an error greater than three times this (0.69 lnm) was excluded as an outlier. This corresponded roughly to any response that was more than twice the target distance or less than half the target distance. This excluded 145 trials (1.66%). Next, we trimmed the outer 10% of targets to remove any heavy distortion due to the bounds on the response range. Unfortunately, even after this, there was still a central tendency bias; the slope of responses regressed onto targets was less than 1.0 m/m on average, t(107) = -4.07, p < 0.001, d = -0.39. One way of thinking of this is to conceptualize the participants as using a high-variance prior with a peak at the center of the response range. We therefore had to employ a novel method of calculating the variable errors to avoid biasing the results. This method is described, analysed, and justified in detail in a separate publication (Aston et al., 2021) and described in outline next.

The next steps were done separately for each participant, session, and trial type. We regressed the responses onto the targets, with both expressed on a logarithmic scale. This resulted in a slope of the regression line and a set of residuals (errors from the regression line). Next, the variable error was calculated by finding the standard deviation of the residuals, then dividing that by the slope, with the slope capped at 1.0 m/m. This was done because a central tendency bias could mask the amount of perceptual noise. For example, suppose a participant takes their actual perceptual estimates and moves them 50% of the way towards the centre of the response range. This would reduce the standard deviation of the residuals by 50% without changing the perceptual noise at all. It would also result in a slope of 0.5 m/m. Dividing the standard deviation of the residuals by the slope recovers the original perceptual noise. This measure is an attempt to directly capture the kind of perceptual noise that distraction should increase and cue combination should reduce, separate from systematic distortions like central tendency bias. This results in a 12 (participants) x 3 (session 3, 4, or 5) x 3 (Auditory, Visual, or AV) matrix of variable errors. These variable errors are taken as the estimate of perceptual noise.

The rest of the process has two sections. First, we wanted to see if the digit span task affected distance judgements. The variable error in sessions 3 and 4, without the digit span task, was compared against the variable error for session 5, with the digit span task. This was done as a one-tailed t-test since the hypothesis is specifically that the digit span would increase variable error. To facilitate this, the variable errors for sessions 3 and 4 were averaged together over sessions, within participants and within trial types. Variable error was then averaged across trial types, within participants. This replicates the mechanics of an ANOVA main effect (essentially averaging everything without the dual task and averaging everything with the dual task) but allows for one-tailed analysis. Second, we wanted to see if the distance judgement affected accuracy on the digit span task. To do this, for each participant we computed the average of the percent of correct memory probes in Dual Task Audio, Dual Task Visual, and Dual Task AV. This was a 12-participant vector of percent correct scores. This was compared with a paired t-test against the percent of correct memory probes in session 10 (i.e. Dual Task Control).

Redundant Signals Hypothesis

This section involves mean reaction times and accuracy rates. Since reaction times often have many outliers, we used the 90% trimmed mean for all responses. For each participant, we calculated the 90% trimmed mean of reaction times for Speed Audio trials, for Speed Visual trials, and for Speed AV trials. This resulted in an 11 (participant) x 3 (Speed Audio, Speed Visual, or Speed AV) matrix of average reaction times. (The twelfth participant only provided data for sessions 1-5.) All trials (including incorrect responses) were included. We also calculated the percent correct for Speed Audio trials, Speed Visual trials, and Speed AV trials.

First, we compared the mean percent correct in Speed Audio trials against the chance rate (50%) with a one-sample t-test. Second, we wanted to compare the average reaction time with each cue alone (i.e. Speed Audio and Speed Visual) versus the average reaction time in Speed AV trials. We compared Speed Audio versus Speed AV with a paired t-test and then compared Speed Visual versus Speed AV with another paired t-test. Third, we wanted to check that any speed gains could not be explained purely by adoption of a different speed-accuracy trade-off, which would be evident in a loss of accuracy. Two paired t-tests were again used to compare accuracy in Speed AV against Speed Audio and then Speed Visual. We also ran additional analyses to confirm that results were robust to changes in analysis details (i.e. trimming, inclusion of incorrect trials).

Forced Fusion Hypothesis

There is unfortunately no standard model for fitting and analysing oddball data. It is not a 2AFC task and there is no consensus theoretical reason to believe that performance will follow a specific curve. We therefore analysed these data in two ways. The first remains very close to the raw data and does no fitting. For this, we used the average congruent deviation and the average incongruent deviation. On each oddball trial, there are two standard choices and the oddball choice, which deviates from the two standard choices. Since these trials used a staircase design, the average deviation is an index of performance. The deviations were expressed on a log scale (i.e. on the same scale as the staircase). The average congruent deviation was the average deviation in A+V+ trials and A-V- trials. The average incongruent deviation was the average deviation in A+V- trials and A-V+ trials. A paired t-test was used to compare congruent and incongruent average deviations. Forced fusion predicts greater deviations for incongruent stimuli. In addition, if this measure is capturing the ability to discriminate the oddball from the standard in a reasonable fashion, then we expect it to capture individual differences. To check that this is true, we examined the correlation between the congruent versus incongruent outcome measures.

For the second method, we adapted a typical 2AFC model based on the Gaussian cumulative density function (CDF). We assumed that the probability of a correct choice is:

$$(\mathrm{N}(x = |D|, \mu = 0, \sigma^2 = s) - 0.5) * 2 * .65 + {}^{1}\!/_{3}$$

Where N() is the normal CDF, *D* is the deviation of the oddball from the standard on a log scale, and *s* is the fitted parameter. This equation has a minimum of 1/3 where D=0, has a limit of 0.98333… as *D* moves away from 0, and is monotonically increasing as *D* moves away from zero. While there is little in the way of deep theory behind this equation, it at least satisfies some basic properties that we would like: the probability of a correct response is at least equal to chance guessing (1/3), at most near 100% but with a small overhead for lapses, and increases as the oddball becomes increasingly different from the standard. With smaller values of *s*, the equation moves more steeply towards its upper limit as a function of *D*. We fitted the value of *s* to each participant by maximizing the likelihood of the observed data. We then solved for the value of *D* that results in a probability of 2/3 for a correct response and took that as the fitted threshold. Appendix 2 reports a small simulation study that examined the fit of this model, finding it acceptable for present purposes.

Precision Hypothesis

Here, we wanted to compare the variable error for the best single cue against the variable error for AV trials. The variable error for the best single cue was taken as the variable error for either auditory or visual, whichever was lower. This could be different for each combination of participant and session. This resulted in a 12 participants x 3 sessions x 2 (best single, AV) matrix. The data were then analysed with a repeated-measures ANOVA (best single versus AV). The section above regarding the Resistance to Dual Task Interference Hypothesis describes the calculation of variable errors.

Optimal Weight Hypothesis

For this we needed the optimal variable error (i.e. perceptual noise as a standard deviation) for each participant during sessions 3-5, the weights given to each cue during sessions 6 and 7, and the optimal weights to give each cue during sessions 6 and 7. The optimal variable error for sessions 3-5 was found with the formula:

$$(\sigma_{Audio}^{-2} + \sigma_{Visual}^{-2})^{-1/2}$$

The variable errors are standard deviations. This formula transforms them to precisions (1/variance), adds them, and then transforms them back to standard deviations. This results in a 12 (participant) x 3 (session 3, 4, or 5) matrix of optimal variable errors. This was compared to the AV variable error for each participant in sessions 3-5 with a repeated-measures ANOVA.

The optimal weight to give each cue was calculated separately for each participant and session (6 or 7). The optimal weight to give the visual cue (Rohde et al., 2016) is

$$\frac{\sigma_{Visual}^{-2}}{\sigma_{Visual}^{-2} + \sigma_{Audio}^{-2}}$$

The optimal weight to give the audio cue is one minus the optimal visual weight. Standard deviations were found for the audio trials and the visual trials in the same manner as described above (Resistance to Dual Task Interference hypothesis). This resulted in an 11 (participant) x 2 (session 6 or 7) x 2 (Auditory or Visual) matrix of optimal weights. These are compared against the observed weights, described below.

The observed weights were found by multiple regression. This was done separately for each participant and session (6 or 7). The natural logarithm of the responses was the outcome variable. The distance signalled by the audio and visual cue, both on a logarithmic scale, were the predictors. The weights were taken to be the estimated beta value for each predictor. In four cases where this was below zero or above one, they were adjusted to be zero or one. Note that this procedure does not always result in observed weights that sum to exactly one, but did tend to remain close to this (average of 0.95) – though as it happens, forcing them to sum to 1 does not alter the significance pattern of the results here. This results in an 11 (participant) x 2 (session 6 or 7) x 2 (auditory or visual) matrix of observed weights.

Observed visual weights were compared to optimal visual weights with a repeated-measures ANOVA. The same was done for auditory weights in a separate ANOVA. The correlation between observed visual weights and optimal visual weights was also analysed; again, the same for auditory. Finally, a paired t-test was used to see if the visual weight was higher in session 7, which had a longer (i.e. more reliable) visual stimulus, than in session 6, which had the normal (i.e. less reliable) visual stimulus.

Data, code (including analysis code), and methods are available at https://osf.io/wzan2/.

Results

The raw data are attached as supplemental material.

Cue Learning (Preliminary) Hypothesis

This hypothesis was an essential first check that participants individually learned to use the new sensory skill. Results were consistent with this hypothesis. In Audio trials, participants had to judge the distance to a target with the new sensory skill. Of the fourteen participants, twelve showed significant correlations between target distances and response distances in the second session for Audio trials. For those twelve, correlations ranged from 0.55 to 0.91, all p-values below $1x10^{-26}$; the remaining two were excluded.

A post-hoc test for a practice effect was also run by entering the variable error on Audio trials for sessions 3 to 7 in a repeated measures ANOVA. This was not significant, F(4, 40) = 1.28, p = 0.294, $\eta^2_{partial}$ = .11. suggesting that performance with the new sensory skill had largely stabilized by the end of the second session. Weber fractions ranged from 0.16 to 0.52 (mean of 0.28, median of 0.27, SD of 0.10). In other words, while performance was well above chance, it was also well short of the performance levels that experts achieve with their own clicks and non-virtual stimuli (Thaler et al., 2019).

Resistance to Dual Task Interference Hypothesis

This hypothesis suggests that the new sensory skill can be used with minimal interference from a verbal working memory task. Results were consistent with this hypothesis. First, we found no evidence that the verbal working memory task caused interference when added onto the basic judgement task. On judgement-only trials, participants judged a distance with the new sensory skill (Audio), a visual cue (Visual), or both (AV). On Dual Task trials, they also had to remember six digits. Judgement-only variable error was not significantly better (lower) than Dual Task variable error, t(11) = 2.13, p = 0.972, d = -0.61 (Figure 4, left). If anything, performance was trending towards better performance (lower variable error) with the verbal working memory task than without. The corresponding Bayes factor is $BF_{01} = 8.76$, generally considered moderate evidence for the null hypothesis. Second, we found no evidence that the distance judgement task caused interference when added onto the verbal working memory task. In the digit span control task, participants only had to remember six digits for a delay that matched the Dual Task trials. The mean proportion of digits correctly recalled was not significantly different for the Dual Task trials (i.e. averaging Audio Dual Task, Visual Dual Task, and AV Dual Task) versus the digit span control task, t(10) = -2.15, p = 0.057, d = 0.65 (Figure 4, right). All post-hoc comparisons between control versus individual Dual Task trial types were non-significant after correction. In short, neither the judgement task nor the working memory task significantly decreased performance when added to the other. If distance judgements were a fundamentally verbal task, then the digit span task should have interfered with them in the same way that many other pairs of verbal tasks interfere with each other (Epling et al., 2017; Wickens, 2002). This suggests that processing these distance cues was largely parallel to verbal working memory.

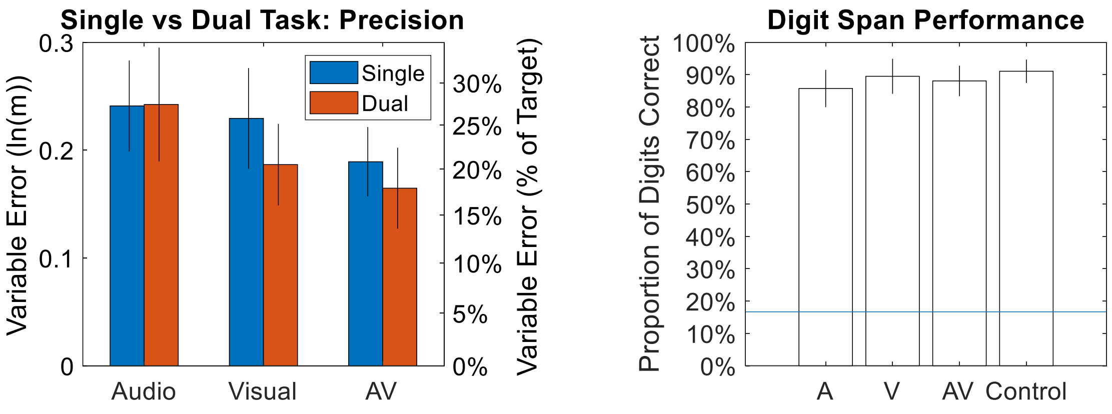

Figure 4: Performance for the dual task versus single-task variants. The left side charts variable error for different trial types. The right side charts the proportion of correctly

answered memory probes. The blue line represents chance performance. Error bars are 95% confidence intervals.

Further post-hoc testing also showed that variable error in the Dual Task AV trials was lower than variable error in the best single-cue dual-task trials, t(11) = -3.10, p = .010, d = -0.90. In other words, we found a cue combination effect during trials where participants also had to complete a verbal working memory task. This further clarifies that the cue combination effect described later (Precision Hypothesis) is not dependent on having verbal resources available.

Redundant Signals Hypothesis

This hypothesis tested whether participants could improve speed with both cues versus the best single cue. Results were consistent with this hypothesis. Speed trials asked participants to decide which of two targets were indicated by an audio cue (Speed Audio), a visual cue (Speed Visual), or both (Speed AV). For all 11 participants, the best (fastest) cue was the visual cue. Mean reaction times were faster in Speed AV trials than Speed Visual trials, t(10) = 5.11, p <.001, d = 1.54 (Figure 5, bottom left). Choices were not significantly less accurate in Speed AV trials than Speed Visual trials, t(10) = 0.71, p = 0.496, d = 0.21 (Figure 5, bottom right). In other words, we found that adding the audio cue onto the visual cue caused an improvement in speed that was not due to a trade-off in accuracy. Further research will be needed to clarify if this is due to racing or co-activation (Miller, 1982).

For completeness, we also report descriptions and comparisons of the Speed Audio trials. For Speed Audio trials, the mean reaction time was 740ms (95% CI: 628 to 851ms). Accuracy was significantly higher than the 50% accuracy benchmark for the Speed Audio trials, t(10) = 9.83, p < .001, d = 2.96, mean of 72% (95% CI: 68% to 78%). Compared to Speed AV trials, Speed Audio trials were both slower, t(10) = 7.92, p < .001, d = 2.39, and less accurate, t(10) = -4.45, p = 0.001, d = -1.34.

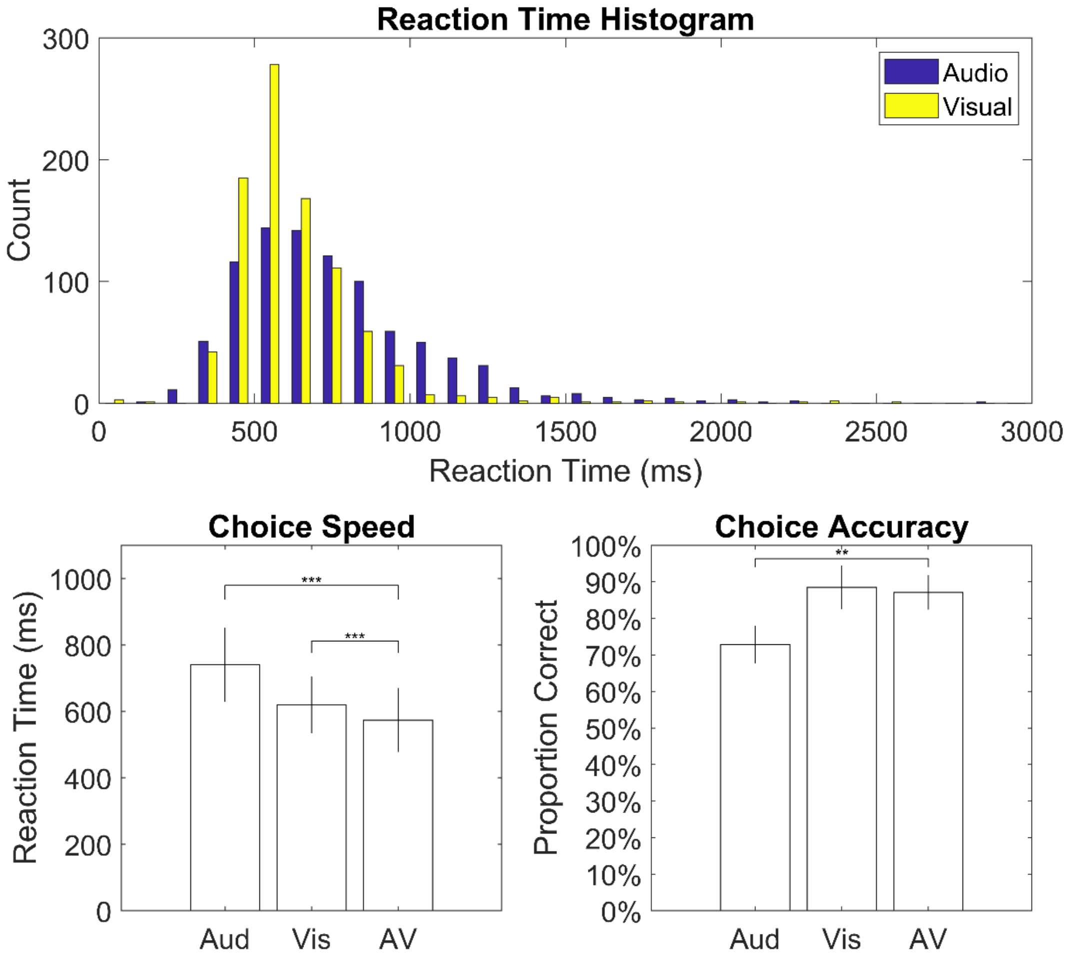

Figure 5: Performance on the speeded task by trial type. On the top, a histogram of reaction times to a choice of target versus target +/- 40%. On the bottom left, 90% trimmed means for reaction times. On the bottom right, accuracy of choices (50% chance). Error bars are 95% confidence intervals. Participants were faster to make a choice with both stimuli (i.e. AV) than either single cue alone. Participants were also more accurate with both stimuli than with the audio cue alone.

We also examined how robust these results are to specific choices in terms of the analysis details. The main planned analysis used all trials and trimmed the outer 10% of reaction times to remove outliers. We also ran the analysis post-hoc with either all trials or only the correct trials, as well as trimming 0% to 99% in steps of 1% (with 99% trim only leaving the median). Of these 200 tests, the maximum p-value was 0.014. This suggests that the significant speed increase is not a particular artefact of the analysis details or a distortion from including all trials.

Forced Fusion Hypothesis

This hypothesis suggests that participants will not be able to avoid averaging the audio and visual cues. We did not find any evidence in favour of this hypothesis. Oddball trials asked participants to inspect three AV displays and "pick the odd one out". These trials are classified as congruent (A+V+ and A-V-) or incongruent (A-V+ and A+V-) depending on how the audio and visual aspects of the oddball display differ from the other two displays. If forced fusion is occurring, we expect performance in incongruent trials to be worse. Performance thresholds were not significantly different for congruent oddball trials versus incongruent oddball trials, t(10) = 0.30, p = 0.768, d = 0.09 for the mean deviation measure, t(10) = 1.32, p = 0.217, d = 0.40 for the fitted threshold measure (Figure 6, left). Corresponding Bayes factors are BF_{01} = 4.12 and 6.67, which is generally considered moderate evidence in favour of the null hypothesis. This fails to support the presence of forced fusion. However, congruent and incongruent scores were highly correlated, r(9) = 0.92, p < .001 for the mean deviation measure, r(9) = 0.98, p < .001 for the fitted threshold measure (Figure 6, right). This suggests that the measure was reasonably sensitive to individual variation in precision at judging distance.

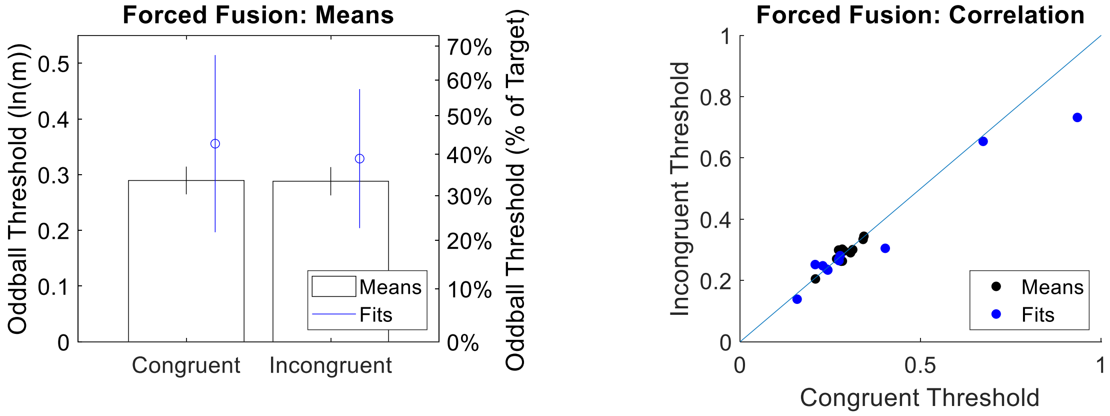

Figure 6: Results for the tests of forced fusion. On the left, thresholds were similar for congruent and incongruent trial types. Under forced fusion, we would expect incongruent performance to be worse. On the right, correlation between the two measures shows that the task was sensitive to individual variation in performance. Black dots and bars reflect estimating thresholds by the mean deviation. Blue dots and bars reflect estimating thresholds through a fitting method based on the normal distribution. Error bars are 95% confidence intervals.

Precision Hypothesis

This hypothesis tested whether participants gained precision (reduced variable error) by combining the new sensory skill with a visual cue when both were available. Results are consistent with this hypothesis. AV Trials asked participants to judge the distance to a target with both an audio and visual cue. The best single cue refers to trials that also ask participants to judge a distance to the target – whichever single cue produced the lowest variable error for that participant and session. The variable error for AV trials was lower (better) than the variable error for the best single cue, F(1, 11) = 5.59, p = .038, $\eta^2_{partial}$ = 0.34 (Figure 7), d = 0.68. This was also true for a variation of the analysis where ranks were entered instead of raw scores, F(1, 11) = 5.92, p = .033, $\eta^2_{partial}$ = 0.35. This suggests that participants did combine the visual cue and the new audio cue, extending previous results with external noise (Negen et al., 2018) to the present study using internal noise.

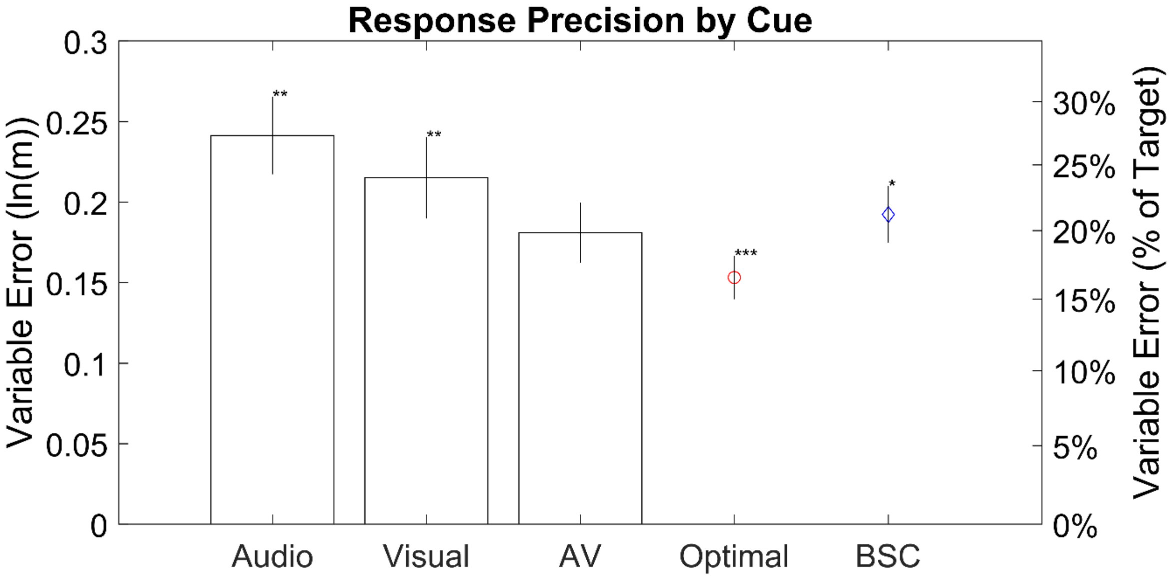

Figure 7: Average variable error by trial type in sessions 3 through 5. Variable errors are used to quantify the amount of non-systematic perceptual noise in judgements (see *Resistance to Dual Task Interference* under *Data Analysis* under *Method* for calculations). Audio and Visual are the trials where only the audio or only the visual stimulus, respectively, were presented. AV stands for audio-visual (i.e. trials where both the audio and visual stimulus were presented). Optimal and Best Single Cue (BSC) are not different trial types; they are values derived from the Audio, Visual, and AV trials. Optimal refers to the best possible variable error that should theoretically be possible with optimal computations. BSC stands for best single cue; for each participant and session, it is the lesser (better) of the auditory variable error and the visual variable error. Error bars are 95% confidence intervals. Asterisks compare each type against AV in a paired t-test: *p < .05, **p < .01, and ***p < .001. Error bars are 95% confidence intervals.

There was also a difference in slopes (see Data Processing and Analysis Plan). Slope corrections were on average higher in AV trials than single-cue trials (Audio: 0.89 m/m; Visual: 0.90 m/m; AV: 0.96 m/m). The difference is significant as a main effect, F(2, 22) = 3.70, p = .041, η^2 = 0.08. However, this is not surprising. Slopes are expected to increase towards 1.0 m/m as perceptual precision increases. See Aston et al. (2021) for a full discussion of this phenomenon and detailed justification of this analysis method.

Optimal Weight Hypothesis

This hypothesis tested whether participants set the optimal weights for each cue and gained optimal precision from the use of both cues. Results are partially consistent with this hypothesis. AV Perturbation trials and Long AV Perturbation trials asked participants to judge a target location when given an audio cue and a visual cue that were offset by 10%. This 'long' variation presented the visual cue for a longer time. First, we did find that cue weights were related to the optimal cue weight. For AV Perturbation and Long AV Perturbation trials, there was a significant correlation between optimal visual weights and observed visual weights, r(20) = 0.78, p < .001 (Figure 8). There was also a significant correlation between optimal auditory weights and observed auditory weights, r(20) = 0.80, p < .001. When the visual reliability was higher in Session 7 than Session 6 (i.e. longer versus shorter), this did lead to higher weight being placed on the visual cue, t(10) = -3.15, p = 0.010, d = -0.95 and lower weight being placed on the audio cue, t(10) = 3.93, p = 0.003, d = 1.18. Mean observed visual weight shifted from 0.588 for less reliable (i.e. shorter) visual cues to 0.776 for more reliable (i.e. longer) visual cues. All of this is consistent with the hypothesis so far.

However, the weights still differed from optimal. The observed visual weights were higher on average than the optimal visual weights, F(1, 10) = 5.53, p = 0.041, $\eta^2_{partial}$ = .356, d = 0.71. The mean observed visual weight was 0.682 versus an optimal 0.608. In addition, the observed auditory weights were lower on average than the optimal auditory weights, F(1, 10) = 16.32, p = 0.002, $\eta^2_{partial}$ = 0.620, d = -1.22. The mean observed auditory weight was 0.259 versus an optimal 0.391. This suggests that participants did vary in how they set integration weights, and that this variation is explained partially by the optimal weight, but that participants also tended to systematically over-rely on the visual cue. In accordance with this, the variable error in AV trials was higher than the optimal variable error, F(1, 11) =

23.00, p < .001, $\eta^2_{partial}$ = 0.677, d = -1.50. This was also true in a variation where the ranks were entered instead of the raw variable errors, F(1, 11) = 32.69, p < .001, $\eta^2_{partial}$ = 0.748.

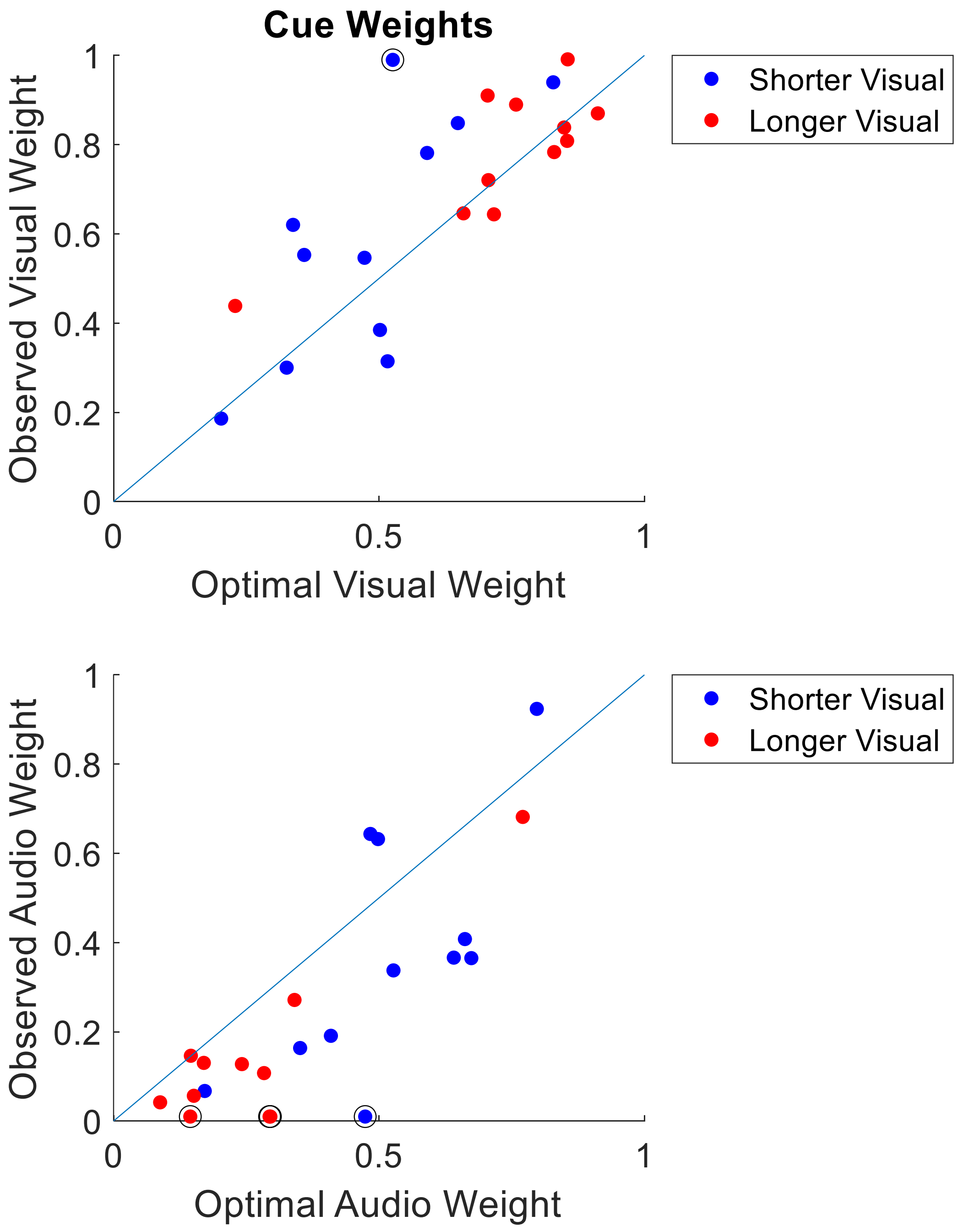

Figure 8: Weights given to each cue. Each blue dot is a separate participant in session 6, when the visual cue was shorter (less reliable). Each red dot is a separate participant in session 7, when the visual cue was longer (more reliable). Dots with a black circle around

them were adjusted to be in the range of zero to one (i.e. into the range of prior plausibility). The blue line is the identity line.

Appendix 1 examines this interpretation via cross-validation. These data are used to compare three models: one where people combine cues optimally, one where people combine cues with too little weight on the audio cue, and one where participants only rely on one cue at a time. In summary, the model where people under-rely on the new sensory skill is favoured for eight of twelve participants, with the optimal model favoured for another three. This largely accords with the interpretation given here (a precision increase through the combination of multiple cues, but not generally to the optimal level).

Discussion

In this study, we wanted to understand the principles by which new sensory skills operate in multisensory environments – a key aspect of how flexibly perception and decision-making operate. We found that use of this new skill met three key criteria: enhancing the speed of perceptual decisions, processing through a non-verbal route, and integration with vision in an efficient, Bayes-like manner. We also show limits: integration was less-than-optimal, and there was no mandatory fusion of signals. It is noteworthy that these skills were attained after only very short initial training and experience. Our results provide further evidence in keeping with proposals that plasticity in perception and decision-making allows new sensory skills to take on new and useful functions (Amedi et al., 2017; Maidenbaum & Abboud, 2014; Striem-Amit et al., 2012). This suggests that perception and decision-making not only have flexibility on the unisensory level, but also at the multisensory level – it is not limited to learning how to use a new sensory skill on its own, but can also adapt to integrating and coordinating it with other perceptual systems.

On balance, this can be interpreted as a new way of further understanding the potential advantages of new sensory skills. Since new sensory skills can improve speed and can be processed non-verbally, they may be quite useful in naturalistic settings where speed can be important and the availability of verbal resources is useful. This is potentially a large shift from thinking of new sensory skills as something that is used only slowly and with full attention. In addition, when we evaluate a new sensory skill, we may want to think less about how well people can use it in isolation and more about the way it will contribute to and coordinate with the overall multisensory processing stream. Failing to consider the benefits of such multisensory processes could mean that new sensory skills are (severely) undervalued.

One of the questions in the scope was whether participants would show efficient cue combination with a cue that is subject to internal noise. We observed a number of specific findings consistent with this: responses were less noisy with both cues than either single cue alone; the weight given to each cue was positively correlated with the optimal weight, and the weight flexibly changed when cue reliabilities changed. In short, participants followed Bayes-like principles to gain measurable benefits from combining multiple cues, much as is seen in native multisensory perception (Alais & Burr, 2004; Ernst & Banks, 2002; Knill & Pouget, 2004; Pouget et al., 2013). This adds important knowledge to the new field of multisensory processing with new sensory skills – the only previous evidence for Bayes-like combination with a new sensory skill came from our previous study in which the other (visual) cue had external noise (Negen et al., 2018). Because internal noise is typically the major issue for real-world perceptual-motor problems, the current results link lab findings more closely to everyday perception and to potential future applications of sensory training.

However, while performance qualitatively followed the hypothesized pattern of noise reduction, performance did not quantitatively meet the predictions of optimal noise reduction. The weight given to the visual cue was too high; the weight given to the auditory cue was too low; noise was not reduced to the optimal level theoretically possible. This suggests that new sensory skills can interact with native perception to reduce noise, but that optimal noise reduction is either out of reach or may require different circumstances (for example, longer training). It should be noted however that full Bayesian optimality is a high standard to compare with, since even native perceptual skills do not always show optimal Bayesian performance (Rahnev & Denison, 2018), and full evaluations of optimality also require consideration of costs and priors, not included in this study.

Flexible cue combination is an especially important result because it expands the scope of sensory training from the traditional model of substitution or replacement to one of augmentation. In our study, the new signal did not only substitute for vision, but improved the precision of existing visual capabilities. This has important potential applications to patients with sensory loss (e.g. partial vision loss) who still have some useful visual function. It also has applications towards developing devices to further enhance healthy perception, such as the use of additional sensors to guide surgery, to navigate, to play enhanced sports, to more efficiently work with heavy objects in a warehouse, or to locate potential hazards.

Results also clarify that the precision increase does not appear to be due to forced fusion. In other words, it seems that the combination of the two cues is not so early in the process that people have difficulty working with the two separate estimates and ignoring their average. Further research will be needed to find out more about when and how the combination occurs. Research on the brain basis of multisensory integration shows that integration happens at multiple levels, from forced fusion in early ‘sensory’ areas, to more sophisticated decision-making in line with causal inference in higher ‘decision’ related areas (Rohe & Noppeney, 2015). The extent to which these different levels may be reshaped by different kinds and durations of training with new skills is a question for future research.

There is also room left to further clarify the extent of verbal interference as well. The present study was only designed to detect very large verbal interference effects. We consider this good evidence that performance with the new sensory skill does not, in layman’s terms, depend merely on ‘talking themselves through’ its use. However, it very much remains possible that a more extensive investigation of this issue will reveal smaller interference effects from verbal working memory. There may also be individual differences in terms of ability to process new sensory skills through non-verbal routes, which the present study was also not designed to detect.

The dual task results also bring up a very unexpected possibility for future work. If anything, participants performed better with the dual task. One participant spontaneously described this as such: “with all the numbers, you sort of get out of your head and just point to the right spot.” This is extremely speculative, but it may be that a distracting task can act as a strategy to improve performance with a new sensory skill after some explicit practice. In other words, it is possible that top-down strategies eventually become actively harmful to the process of using a new sensory skill. This would, of course, require more research before any conclusions could be reached.

While we have answered a number of key questions, there are also a number left aside. Perhaps it is most useful to say that many studies of sensory substitution try to intersect psychology and philosophy; our approach here is closer to intersecting psychology and engineering. We have left aside questions like the perceptual feel of the new sensory skill (Witzel et al., 2021), the generalizability of the new sensory skill (Negen et al., 2018), whether the new sensory skill specifically affects perception (Deroy & Auvray, 2012), whether the new sensory skill was learned by simple associative learning versus some more sophisticated method (Nagel et al., 2005), whether the new sensory skill took on a visual character (White et al., 1970), and so on. These are all interesting questions but they lay somewhat aside from our current research aims. We are fully confident that the new sensory skill here is a genuine new sensory skill simply because responses were correlated above chance with targets and because we know this same cue is not used to a meaningful degree without training (Negen et al., 2018). That already qualifies it as a new sensory skill as we intend the term; further questions about the new sensory skill's character in isolation are outside the current scope.

There are a wide variety of different questions left open that will be excellent future directions. We should learn more about generalization – how and when a new sensory skill transfers to untrained distances, environments, and tasks. Since the current study combines many low-level visual cues into one high-level visual cue, it would be interesting to see how individual low-level cues are contributing and if they would each be amenable to cue combination. Viewing this through the lens of a coupling prior (Ernst et al., 2007) could be useful. In addition, it may be useful to learn more about the individual new sensory skill as well. The experience of the new sensory skill, whether it becomes fundamentally spatial for participants, is an open question. It could also be that learning would be faster or different if the sound were self-initiated as it is in everyday practitioners, though this would require the innovation of new simulation techniques. In general, we do not yet have a good characterization of the learning curve involved – for example, an estimate of how long it takes for learning to plateau. Some or all of these could be interesting to explore with model systems like the one here or perhaps even with devices intended for widespread use.

In conclusion, new sensory skills can not only be useful on their own (Thaler et al., 2019) and increase multisensory precision (Negen et al., 2018), but also increase speed and resist dual task interference – even when the new sensory skill has internal noise. There are however some notable limitations after this short training: the lack of quantitatively optimal processing, and the lack of forced fusion. The research opens up important questions for future research, including how potential reorganisation of neural sensory processing (Amedi et al., 2017) may support this kind of multisensory perception and decision-making; how the findings translate to longer training programmes, more complex real-world perceptual-motor tasks; and how they can best be implemented in the design of new devices and approaches for augmenting human perception and decision-making.

Acknowledgments

Thanks to Paula Teo, Rupert Talfourd-Cook, Narjes Kazarooni, Rebecca Leach, and Clara Gallay for help with piloting and data collection. Supported by grant ES/N01846X/1 from the Economic and Social Research Council of the United Kingdom. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 820185).

References

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Amedi, A., Hofstetter, S., Maidenbaum, S., & Heimler, B. (2017). Task Selectivity as a Comprehensive Principle for Brain Organization. *Trends in Cognitive Sciences*, *21*(5), 307–310. https://doi.org/10.1016/j.tics.2017.03.007

Aston, S., Negen, J., Nardini, M., & Beierholm, U. (2021). Central tendency biases must be accounted for to consistently capture Bayesian cue combination in continuous response data. *Behavior Research Methods 2021*, 1–14. https://doi.org/10.3758/S13428-021-01633-2

Burr, D., & Gori, M. (2011). Multisensory integration develops late in humans. In M. M. Murray & M. T. Wallace (Eds.), *The Neural Bases of Multisensory Processes* (pp. 345–362). CRC Press/Taylor & Francis. https://doi.org/10.1201/b11092-23

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Deroy, O., & Auvray, M. (2012). Reading the World through the Skin and Ears: A New Perspective on Sensory Substitution. *Frontiers in Psychology*, *3*(NOV), 457. https://doi.org/10.3389/fpsyg.2012.00457

Epling, S. L., Blakely, M. J., Russell, P. N., & Helton, W. S. (2017). Interference between a fast-paced spatial puzzle task and verbal memory demands. *Experimental Brain Research*, *235*(6), 1899–1907. https://doi.org/10.1007/s00221-017-4938-z

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Ernst, M. O., H., B. H., C., K. D., W., R., Olshausen, B. A., & S., L. M. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 7. https://doi.org/10.1167/7.5.7

Getty, D. J. (1975). Discrimination of short temporal intervals: A comparison of two models. *Perception & Psychophysics*, *18*(1), 1–8. https://doi.org/10.3758/BF03199358

Goeke, C. M., Planera, S., Finger, H., & König, P. (2016). Bayesian Alternation during Tactile Augmentation. *Frontiers in Behavioral Neuroscience*, *10*, 187. https://doi.org/10.3389/fnbeh.2016.00187

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young Children Do Not Integrate Visual and Haptic Form Information. *Current Biology, 18*(9), 694–698. https://doi.org/10.1016/j.cub.2008.04.036

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology*, *63*(3), 289–293. https://doi.org/10.1037/h0039516

Hillis, J. H., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, *298*(5598), 1627–1630. https://doi.org/10.1126/science.1075396

Hurst, N. (2017). How Does Human Echolocation Work? *Smithsonian Magazine*. https://www.smithsonianmag.com/innovation/how-does-human-echolocation-work-180965063/

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kolarik, A. J., Cirstea, S., Pardhan, S., & Moore, B. C. J. (2014). A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research*, *310*, 60–68. https://doi.org/10.1016/j.heares.2014.01.010

König, S. U., Schumann, F., Keyser, J., Goeke, C., Krause, C., Wache, S., Lytochkin, A., Ebert, M., Brunsch, V., Wahn, B., Kaspar, K., Nagel, S. K., Meilinger, T., Bülthoff, H., Wolbers, T., Büchel, C., & König, P. (2016). Learning new sensorimotor contingencies: Effects of long-term use of sensory augmentation on the brain and conscious perception. *PLoS ONE*, *11*(12). https://doi.org/10.1371/journal.pone.0166647

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. https://doi.org/10.1038/nature02169

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326. https://doi.org/10.1016/J.TICS.2006.05.003

Macmillan, N. A., & Creelman, C. D. (2004). Detection Theory: A User's Guide: 2nd edition. In *Detection Theory: A User's Guide: 2nd edition*. Psychology Press. https://doi.org/10.4324/9781410611147

Maidenbaum, S., & Abboud, S. (2014). Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience & Biobehavioral Reviews*, *41*, 3–15. https://doi.org/10.1016/J.NEUBIOREV.2013.11.007

Maidenbaum, S., Hanassy, S., Abboud, S., Buchs, G., Chebat, D.-R., Levy-Tzedek, S., & Amedi, A. (2014). The EyeCane, a new electronic travel aid for the blind: Technology, behavior & swift learning. *Restorative Neurology and Neuroscience*, *32*(6), 813–824. https://doi.org/10.3233/RNN-130351

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(01), 147. https://doi.org/10.1017/S0952523808080905

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247–279. https://doi.org/10.1016/0010-0285(82)90010-X

Nagel, S. K., Carl, C., Kringe, T., Märtin, R., & König, P. (2005). Beyond sensory substitution - Learning the sixth sense. *Journal of Neural Engineering*, *2*(4), R13. https://doi.org/10.1088/1741-2560/2/4/R02

Nardini, M., Bedford, R., & Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(39), 17041–17046. https://doi.org/10.1073/pnas.1001699107

Negen, J., Wen, L., Thaler, L., & Nardini, M. (2018). Bayes-Like Integration of a New Sensory Skill with Vision. *Scientific Reports*, *8*(1), 16880. https://doi.org/10.1038/s41598-018-35046-7

Norman, L. J., & Thaler, L. (2019). Retinotopic-like maps of spatial sound in primary 'visual' cortex of blind human echolocators. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1912), 20191910. https://doi.org/10.1098/rspb.2019.1910

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, e223. https://doi.org/10.1017/S0140525X18000936

Rohde, M., Van Dam, L. C. J., & Ernst, M. O. (2016). Statistically optimal multisensory cue integration: A practical tutorial. *Multisensory Research*, *29*(4–5), 279–317. https://doi.org/10.1163/22134808-00002510

Rohe, T., & Noppeney, U. (2015). Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception. *PLOS Biology*, *13*(2), e1002073. https://doi.org/10.1371/JOURNAL.PBIO.1002073

Rohe, T., & Noppeney, U. (2018). Reliability-Weighted Integration of Audiovisual Signals Can Be Modulated by Top-down Attention. *ENeuro*, *5*(1).

https://doi.org/10.1523/ENEURO.0315-17.2018

Senna, I., Andres, E., McKyton, A., Ben-Zion, I., Zohary, E., & Ernst, M. O. (2021). Development of multisensory integration following prolonged early-onset visual deprivation. *Current Biology*, *31*(21), 4879-4885.e6. https://doi.org/10.1016/j.cub.2021.08.060

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*, 425–432. https://doi.org/10.1016/j.tics.2010.07.001

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147–152. https://doi.org/10.1016/S0926-6410(02)00069-1

Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with Sounds: Sensory Substitution Selectively Activates the Visual Word Form Area in the Blind. *Neuron*, *76*(3), 640–652. https://doi.org/10.1016/j.neuron.2012.08.026

Teng, S., & Whitney, D. (2011). The acuity of echolocation: Spatial resolution in the sighted compared to expert performance. *Journal of Visual Impairment & Blindness*, *105*(1), 20–32. http://www.ncbi.nlm.nih.gov/pubmed/21611133

Thaler, L., De Vos, H. P. J. C., Kish, D., Antoniou, M., Baker, C. J., & Hornikx, M. C. J. (2019). Human Click-Based Echolocation of Distance: Superfine Acuity and Dynamic Clicking Behaviour. *JARO - Journal of the Association for Research in Otolaryngology*, *20*(5), 499–510. https://doi.org/10.1007/s10162-019-00728-0

Thaler, L., & Goodale, M. A. (2016). Echolocation in humans: an overview. In *Wiley interdisciplinary reviews. Cognitive science* (Vol. 7, Issue 6, pp. 382–393). John Wiley & Sons, Inc. https://doi.org/10.1002/wcs.1408

Thaler, L., Wilson, R. C., & Gee, B. K. (2014). Correlation between vividness of visual imagery and echolocation ability in sighted, echo-naïve people. *Experimental Brain Research*, *232*(6), 1915–1925. https://doi.org/10.1007/s00221-014-3883-3

Weisberg, S. M., Badgio, D., & Chatterjee, A. (2018). Feel the way with a vibrotactile compass: Does a navigational aid aid navigation? *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(5), 667–679. https://doi.org/10.1037/xlm0000472

White, B. W., Saunders, F. A., Scadden, L., Bach-Y-Rita, P., & Collins, C. C. (1970). Seeing with the skin. *Perception & Psychophysics*, *7*(1), 23–27. https://doi.org/10.3758/BF03210126

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Witzel, C., Lübbert, A., Schumann, F., Hanneton, S., & O'Regan, J. K. (2021). *Can perception be extended to a "feel of North"? Tests of automaticity with the NaviEar*. https://doi.org/10.31234/OSF.IO/2YK8B

Zhang, X., Reich, G. M., Antoniou, M., Cherniakov, M., Baker, C. J., Thaler, L., Kish, D., & Smith, G. E. (2017). Human echolocation: waveform analysis of tongue clicks. *Electronics Letters*, *53*(9), 580–582. https://doi.org/10.1049/el.2017.0454

Appendix 1 – Cross Validation for Cue Combination

Results indicate that participants combined the visual cue and the new audio cue to judge distance. However, results also point towards sub-optimal combination – specifically a tendency to over-weight the visual cue and under-weight the audio cue. This appendix supplements those analyses with a cross-validation approach to investigate the possibility of convergent evidence for this interpretation. In summary, the results presented here agree with the main text.

Data Used

For these models, we use the trials that involve a distance judgement, starting with session 3 (after training). This includes the dual task trials, the perturbation trials, and the trials with a longer visual cue. The 2AFC trials, 3AFC trials, 5AFC trials, speeded task trials, control task trials, and oddball task trials are not used. In session 6, where there are 8 visual-only trials with the long visual cue and 83 trials with the shorter visual cue, only the shorter ones are used. The data from the final participant with a partial dataset are used where possible. Within each session and participant, the nearest 5% of targets and the furthest 5% of targets are excluded. Each session is modelled separately i.e. the same participant can have completely independent parameters in sessions 3, 4, 5, 6, and 7.

Models

The three models are Single Cue, Optimal, and Audio Confidence. Common to all three, we use the natural logarithm of the cues and the responses. Also common to all three, there is a lapse rate and a mechanic for the application of a prior. The lapse reserves 2% of the probability to be spread evenly along the response line. The prior allows for the participant to have their responses biased systematically towards a particular point on the line. While this is modelled explicitly as a prior, the mathematics can also mimic a more general bias e.g. a central tendency bias. Non-lapse trials are modelled as a normal distribution. Late noise is assumed to be negligible beyond what is already captured by the lapse rate.

Single Cue

Conceptually, single cue means that participants only use a single cue when two cues are available. The analysis in the main text suggests that this model is not favoured, at least for most participants. This is an important alternative because it represents the way that young children, who are still learning to use their native senses for the first time, deal with similar multisensory situations (Burr & Gori, 2011). It is also the way that participants were previously shown to address the simultaneous presentation of a trained new sensory cue and a native vestibular cue (Goeke et al., 2016).

Mechanically, this model has five parameters:

1. σ_{Audio} the standard deviation of perception around audio cues. This (and the next two) are constrained to be positive values only.
2. σ_{Visual} the standard deviation of perception around visual cues.
3. σ_{Prior} the standard deviation of a prior that the participant applies.
4. μ_{Prior} the mean of the prior that the participant applies. This value is not constrained.
5. *V* the probability that the participant uses the visual cue. 1-V is the probability that they use the audio cue. This is constrained to be at least zero and at most one.

The modelled distribution of the single-cue trial types follows:

$$
\begin{aligned}
&(1)\ \mathrm{P}\left(\mathrm{Response}_{\mathrm{Auditory}}\right)\\
&\quad = \varphi\left(\mathrm{Cue}_{\mathrm{Auditory}}\frac{\sigma_{\mathrm{Auditory}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right.\\
&\quad \left.+\,\mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}},\sigma_{\mathrm{Auditory}}\frac{\sigma_{\mathrm{Auditory}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right)*.98\\
&\quad +\frac{.02}{\ln(35)-\ln(10)}
\end{aligned}
$$

$$
\begin{aligned}
&(2)\ \mathrm{P}\left(\mathrm{Response}_{\mathrm{Visual}}\right)\\
&\quad = \varphi\left(\mathrm{Cue}_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right.\\
&\quad \left.+\,\mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}},\sigma_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{\sigma_{Visual}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right)*.98+\frac{.02}{\ln(35)-\ln(10)}
\end{aligned}
$$

$$
\begin{aligned}
&(3)\ \mathrm{P}\left(\mathrm{Response}_{\mathrm{AV}}\right)\\
&\quad = \varphi\left(\mathrm{Cue}_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right.\\
&\quad \left.+\,\mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}},\sigma_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right)*.98*V\\
&\quad +\ \varphi\left(\mathrm{Cue}_{\mathrm{Auditory}}\frac{\sigma_{\mathrm{Auditory}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right.\\
&\quad \left.+\,\mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}},\sigma_{\mathrm{Auditory}}\frac{\sigma_{\mathrm{Auditory}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right)*.98*(1-V)\\
&\quad +\frac{.02}{\ln(35)-\ln\,(10)}
\end{aligned}
$$

where φ is the probability density function of the normal distribution, parameterized with a mean then a standard deviation, and ln() is the natural logarithm. In the first two equations, we can see that the φ() term is multiplied by 0.98 and that a term spreading 2% of the probability across the response range (from 10m to 35m) is added. That is the lapse rate. We can also see that the mean response is a weighted average of the cue and the μ_{Prior} term. The standard deviation is also adjusted for the possibility that the cue has a weight below one. Those are the mechanics of the prior that the participant applies. In equation 3, we see that the first φ() term depends on the visual cue and is multiplied by *V*, while the second φ() term depends on the audio cue and is multiplied by (1-V). When V is set to 1, only the visual cue is used. When V is set to 0, only the audio cue is used. When V is in intermediate, the participant switches back and forth between the two available cues at random, but does not use both on the same trial.

Optimal

The optimal model means that participants use the two cues together optimally (though they can still apply an incorrect prior). The analysis in the main text again suggests that this model is not favoured for most participants. This is an important alternative because it reflects the way participants deal with many multisensory tasks that only involve native cues (Pouget et al., 2013).

Mechanically, this model uses the same parameters as Single Cue, except without the final *V* parameter. It also uses the same equations for the audio-only and visual-only trials (i.e. equations 1 and 2). It differs in the equation for the AV trials:

(4a) σ_{AV}

$$= \sqrt{\left(\sigma_{\text{Auditory}} \frac{\sigma_{\text{Auditory}}^{-2}}{\sigma_{\text{Auditory}}^{-2} + \sigma_{\text{Visual}}^{-2} + \sigma_{\text{Prior}}^{-2}}\right)^2 + \left(\sigma_{\text{Visual}} \frac{\sigma_{\text{Visual}}^{-2}}{\sigma_{\text{Auditory}}^{-2} + \sigma_{\text{Visual}}^{-2} + \sigma_{\text{Prior}}^{-2}}\right)^2}$$

$$(4b)\ \mathrm{P}(\mathrm{Response}_{\mathrm{AV}})$$

$$= \varphi\left(\mathrm{Cue}_{\mathrm{Auditory}}\frac{\sigma_{\mathrm{Auditory}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right.$$

$$+\mathrm{Cue}_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}$$

$$\left.+\mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}},\sigma_{AV}\right)*.98+\frac{.02}{\ln(35)-\ln(10)}$$

Here we see that the mean response is a weighted average of the audio cue, the visual cue, and the prior mean. The standard deviation is a pool of the auditory standard deviation times the auditory weight as well as the visual standard deviation times the visual weight. These weights are optimal.

Audio Confidence

Conceptually, Audio Confidence allows a participant to under-estimate the reliability of the audio cue. Otherwise, it is equivalent to the optimal model. The analysis in the main text favours this model. If this model is not favoured here, at least for most participants, that would cause us to carefully re-examine the interpretation given in the main text.

Mechanically, this model uses the same first four parameters. It also uses an additional free parameter *C* that is a factor to reflect under-estimation of auditory precision. *C* is constrained to be at least zero and at most one. *C* can be inferred by looking at the AV trials, especially the ones with a perturbation, to see if they over-rely on the visual cue. Equation 2 is still used for the visual-only trials. The other two equations are

(6) $\mathrm{P}(\mathrm{Response}_{\mathrm{Auditory}})$

$$= \varphi\left(\mathrm{Cue}_{\mathrm{Auditory}}\frac{C\sigma_{\mathrm{Auditory}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}} + \mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}, \sigma_{\mathrm{Auditory}}\frac{C\sigma_{\mathrm{Auditory}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right) * .98 + \frac{.02}{\ln(35)-\ln(10)}$$

(7a) σ_{AVC}

$$= \sqrt{\left(\sigma_{\mathrm{Auditory}}\frac{C\sigma_{\mathrm{Auditory}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right)^2 + \left(\sigma_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}\right)^2}$$

(7b) $\mathrm{P}(\mathrm{Response}_{\mathrm{AV}})$

$$= \varphi\left(\mathrm{Cue}_{\mathrm{Auditory}}\frac{C\sigma_{\mathrm{Auditory}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}} + \mathrm{Cue}_{\mathrm{Visual}}\frac{\sigma_{\mathrm{Visual}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}} + \mu_{\mathrm{Prior}}\frac{\sigma_{\mathrm{Prior}}^{-2}}{C\sigma_{\mathrm{Auditory}}^{-2}+\sigma_{\mathrm{Visual}}^{-2}+\sigma_{\mathrm{Prior}}^{-2}}, \sigma_{AVC}\right) * .98 + \frac{.02}{\ln(35)-\ln(10)}$$

Basically, the actual precision of the audio cue, $\sigma_{\mathrm{Audio}}^{-2}$, is multiplied by *C* in every instance where we are calculating a weight.

One could equivalently state that the Optimal model is the Audio Confidence model with C constrained to equal one.

Results and Discussion

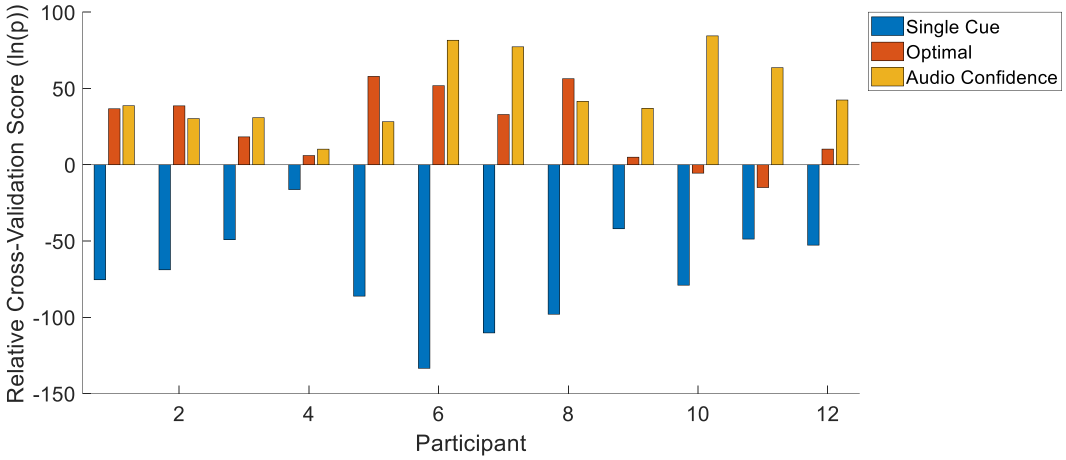

Figure A1. Relative cross-validation scores are the raw log probability (i.e. the final result of the cross-validation procedure) minus the mean score across models and within participants. Higher scores are better. Scores are combined within participants but across sessions.

Results here are broadly consistent with the results in the main paper. For all 12 participants, the Single Cue model was meaningfully worse than the other two. The Optimal model was meaningfully better than the other two models for three participants (2, 5, and 8). One participant (1) had similar cross-validation scores for both the Optimal model and the Audio Confidence model. The remaining eight participants were meaningfully better fit with the Audio Confidence model than the other two. This suggests conceptually that all participants were combining the audio and visual cue. However, a minority were combining optimally and most were under-weighting the audio cue. This fits with the conclusions presented in the main text (i.e. as an average, the audio cue was underweighted, and the cues were not used to maximum efficiency).

In summary, the cross-validation analysis bolstered our confidence in the interpretation that is given in the main text.

Appendix 2 – Goodness of Fit for Oddball Task

Here we report an examination of how well the ad-hoc model of the oddball task fits the data. The main goal is to assess if the actual data majorly violate the pattern of correct and incorrect responses predicted by the fitted model. The alternative (ultimately favoured here) is that the model's fitted predictions and the actual data are within the range expected from simple sampling error.

For each participant, separately for congruent and incongruent trial types, for every oddball deviation (absolute distance between standard and oddball on a log scale) tested, we counted the number of correct responses and the total number of trials. We also used the formula in the main text, copied below, to calculate a predicted rate of correct responses from the fitted values and specified oddball deviations:

$$(\mathrm{N}(x = |D|, \mu = 0, \sigma^2 = s) - 0.5) * 2 * .65 + {}^1/_3$$

This gives us three values for every combination of oddball deviation, participant, and trial type (congruent versus incongruent): P, the probability of a correct response predicted by the model, K, the actual number of correct responses, and N, the number of relevant trials. To quantify deviation from the model predictions, we found the sum of |PN – K|. This was 288.23. To assess if this value indicates a significant violation, we ran a small simulation study. One million times, we simulated a random draw from a binomial distribution for every oddball deviation, participant, and trial type. We will call this K*. We then calculated the sum of |PN-K*| for each of the million repeats. The resulting value was approximately normally distributed with a mean of 291.4 and a standard deviation of 14.6. In other words, the observed value of |PN – K| was slightly below (better) than we would expect on average if the model were perfectly specified; the difference between observed rates of correct response (separated by oddball deviation, participant, and trial type) and the predicted rates of correct response (separated the same way) was within the range of what we would expect via sampling error in this experiment.

We conclude from this that while the fitted model is almost undoubtedly incorrect at some detailed level, it is good enough to approximate the actual response data to a point where the predictions and the available data are not readily statistically discernible. In context of its desirable mathematical features (e.g. probability of a correct response increases with oddball deviation, changing any correct response to incorrect increases the fitted threshold, etc.), we interpret this as being a reasonable model for present purposes.

For further examination by the reader, we also provide charts of the correct response rate against the fitted rates, separated in the same way as the analysis (Figure A2).

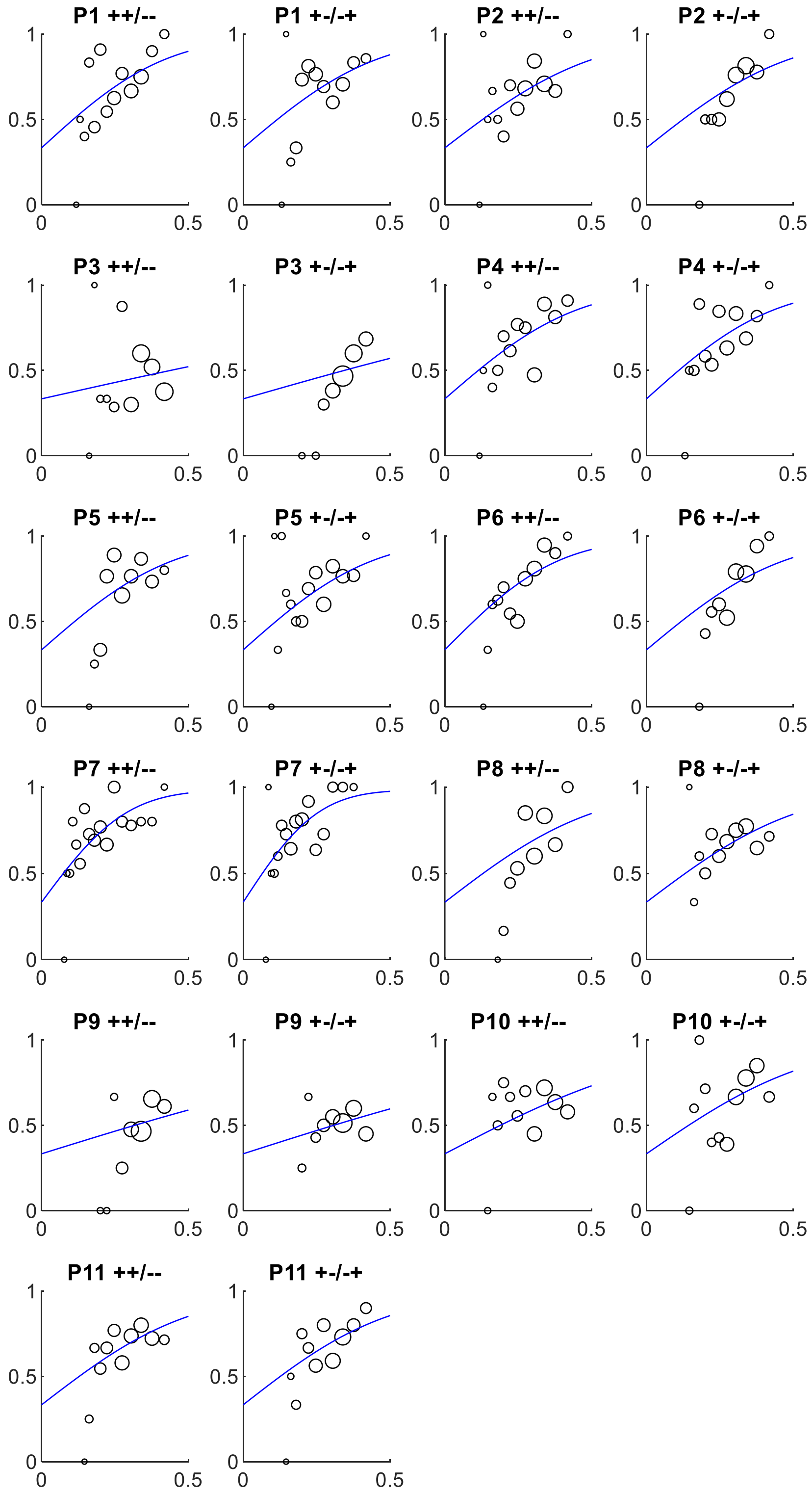

Figure A2. Observed rates of correct response and fitted predictions. The x axis is the absolute oddball deviation on a log scale. The y axis is the rate of correct responses. The size of the circle indicates the number of trials at that oddball deviation. The blue line is the fitted prediction. Each panel represents 130 trials.