Relating Sound and Sight in Simulated Environments

Kevin Y. Tsang* and Damien J. Mannion
School of Psychology, UNSW Sydney, Sydney, 2052, Australia
*Corresponding author; e-mail: kevin.tsang@unsw.edu.au
ORCID iDs: Tsang: 0000-0001-9427-0148; Mannion: 0000-0001-5520-632X

Received 10 March 2022; accepted 17 August 2022; published online 8 September 2022

Abstract
The auditory signals at the ear can be affected by components arriving both directly from a sound source and indirectly via environmental reverberation. Previous studies have suggested that the perceptual separation of these contributions can be aided by expectations of likely reverberant qualities. Here, we investigated whether vision can provide information about the auditory properties of physical locations that could also be used to develop such expectations. We presented participants with audiovisual stimuli derived from 10 simulated real-world locations via a head-mounted display (HMD; n = 44) or a web-based (n = 60) delivery method. On each trial, participants viewed a first-person perspective rendering of a location before hearing a spoken utterance that was convolved with an impulse response that was from a location that was either the same as (congruent) or different to (incongruent) the visually-depicted location. We find that audiovisual congruence was associated with an increase in the probability of participants reporting an audiovisual match of about 0.22 (95% credible interval: [0.17, 0.27]), and that participants were more likely to confuse audiovisual pairs as matching if their locations had similar reverberation times. Overall, this study suggests that human perceivers have a capacity to form expectations of reverberation from visual information. Such expectations may be useful for the perceptual challenge of separating sound sources and reverberation from within the signal available at the ear.

Keywords
reverberation, visual environment, auditory inference, perception, head-mounted display

1. Introduction

Sounds reaching the ear have contributions from both the acoustic energy arriving directly from a source and indirectly via reflections from objects and surfaces in the environment. These reflections, known as reverberation,
arrive at the ear as delayed and distorted copies of the direct source signal and can degrade the intelligibility of the sound (e.g., Harris and Reitz, 1985). The ambiguity in separating the relative contributions of the source and environment in the signal available at the ear (e.g., Kopčo and Shinn-Cunningham, 2011) is likely to contribute to this perceptual difficulty.

Despite this ambiguity, the human auditory system is often able to accurately infer the relative contributions of the source and environment to a sound. Traer and McDermott (2016) proposed that listeners may be aided by knowledge of the characteristics of the reverberation that are typical of real-world listening environments. They reported that participants were more accurate in discriminating between sound sources when they were delivered with naturalistic reverberation, relative to reverberation that deviated from the identified real-world regularities. This suggests that human listeners can use expectations about the likely structure of reverberation when identifying the contributions of the source. Consistent with an important role of reverberation expectation, Brandewie and Zahorik (2010) showed that the intelligibility of speech under moderate to high levels of reverberation increased when listeners had been previously exposed to auditory samples produced with the same reverberant qualities.

Because reverberation is related to the material and geometric structure of the environment, there is the potential for information derived from the visual sense, which is often able to provide reliable estimates of such environmental conditions, to also convey expectations about the likely properties of reverberation. Supporting the potential for vision to generate useful expectations about reverberation, Sandvad (1999) reported that participants were able to accurately identify the image of the environment in which a sound was recorded and Defays et al. (2014) found that participants were better able to sort a set of sounds by their reverberation if the sounds were accompanied by images of reverberation-consistent (‘congruent’) locations. Furthermore, Calcagno et al. (2012) showed that visual information can have practical consequences for the interpretation of auditory signals; they reported that previous visual exposure to an environment improved the accuracy of subsequent blindfolded auditory distance judgements in that environment.

However, other studies have indicated that visual exposure does not necessarily permit accurate or useful expectations about reverberation. McCreery and Calamia (2006) asked participants to adjust the level of reverberation based on a photograph of an environment and reported that such matches were inconsistent with the true reverberant properties. Additionally, Schutte et al. (2019) reported that the simultaneous presence of a visual depiction of an environment did not affect judgements of reverberation. Hence, the potential role of vision in the perceptual capacity to separate the source and environmental contributions to the sound signal at the ear is unclear.
Here, our goal was to measure the ability of human participants to relate the auditory and visual characteristics of physical locations. We used both head-mounted display (HMD) and web-based presentation methods to simulate 10 real-world locations. By presenting each pairwise combination of visual and auditory location and asking participants to judge their congruence, we were able to assess the ability to form accurate cross-modal expectations — which could provide the foundation for vision to assist with the problem of separating the source and environmental contributions to auditory signals.

2. Method

2.1. Participants

We recruited 56 participants for the HMD experiment and 130 participants for the web experiment. The majority of participants were university students who were studying a first-year psychology course at UNSW Sydney (43/56 HMD participants, 130/130 web participants), who received course credit for their participation. The remaining 13 participants were recruited for the HMD experiment from the paid research participant database in the School of Psychology at UNSW Sydney, who were paid A$15 for their participation. All participants gave their informed consent in accordance with the experiment protocols approved by the Human Research Ethics Advisory Panel in the School of Psychology, UNSW Sydney (HMD experiment: #3251; web experiment: #3436). The entire participation session lasted approximately 45 min.

Participants were required to self-assess as having no known hearing abnormalities and normal or corrected-to-normal vision to be eligible to participate. For the HMD experiment, participants were excluded from participation if they self-assessed as being tired, sleep-deprived, under the influence of alcohol/drugs, hung-over, having digestive problems, under emotional stress, or suffering from a cold, flu, or migraine (in accordance with the Health and Safety Warnings for the Oculus virtual reality device). For the web experiment, participants were required to be using a desktop or laptop computer with headphones and a Firefox or Chrome browser and be in a distraction-free environment with a stable internet connection for the duration of the participation session.

We included 104 participants in the analysis (44 from the HMD experiment and 60 from the web experiment), following application of our exclusion criteria (described in subsection 2.6. Analysis below). The age and gender distributions for these participants is shown in Fig. 1.
Figure 1. Summaries of the demographics and device properties of participants included in the analysis, separated by the experimental delivery conditions (HMD and web). Participants self-reported their age and gender (top row). The experiment automatically collected the relevant hardware and software properties associated with the devices that the participants in the web delivery condition used to access the experiment (middle row). Participants in the web delivery condition also self-reported the properties of the headphones that they used during the experiment (bottom row).

2.2. Apparatus

The HMD experiment was conducted in one of two similar testing cubicles. Visual stimuli were presented through a virtual-reality headset (Oculus, California, USA; Oculus Rift CV1 model) with a spatial resolution of 1080 × 1200 per eye and a temporal resolution of 90 Hz. A pupilometer (SNOWINSPRING, Shenzhen, China; Optical Digital Pupilometer) was used to measure the distance between the participants’ pupils, allowing for an accurate alignment of the lenses inside the headset to reduce discomfort. Based...
on their handedness, the participant used either the right or left hand controller to interact with the experiment. Auditory stimuli were presented via an AudioFile device (Cambridge Research Systems, Rochester, UK) and were heard through circumaural over-ear headphones (Beyerdynamic, Heilbronn, Germany; model DT990 Pro). The experiment was controlled using Python 3.6.8, with the PsychXR package (Cutone and Wilcox, 2018) used to interface with the HMD.

The web experiment was conducted using the hardware that was available to the participant at the time that they completed the session. A summary of the properties of such hardware is shown in Fig. 1. The web experiment was hosted using a local instance of JATOS (version 3.3.6; Lange et al., 2015). The experiment was controlled using custom JavaScript, with audiovisual functionality provided by the three.js library (revision 122dev; https://threejs.org). Additional functionality was provided by the jsPsych library (revision v1.53-1979-g66c8452d; de Leeuw, 2014).

2.3. Stimuli

Auditory and visual stimuli originated from different real-world locations from the EchoThief impulse response library (http://www.echothief.com), where each location had a representation of its visual (panorama) and reverberant (impulse response recording) properties. From the 62 locations in the library that had a high-resolution panorama, we chose 10 locations for use in this study.

Each location had a panoramic image that depicted its visual environment, as shown in Fig. 2. In the HMD experiment, these images were rendered in the headset such that the observer was surrounded by the environment, with the visible region of the location depending on the observer’s momentary spatial orientation. In the web experiment, these images were rendered in fullscreen on the participant’s monitor in first-person perspective, with the visible region of the location depending on the participant’s movement of their mouse.

Each location also had a recording of an auditory impulse response that captured its reverberant properties. Spectrograms of the impulse response waveforms and estimates of their reverberation times are shown in Fig. 3. The $\text{RT}_{60}$ estimates were obtained at octave-spaced frequencies between 63 Hz and 8000 Hz using the ‘optimal’ method of the roomeqwizard application (http://www.roomeqwizard.com; version 5.19). The estimates for any combination of frequency band of channel where the correlation obtained from the best fit was poorer than $-0.99$ were discarded, and the estimates from the left and right channels were then averaged. The set of frequency-specific $\text{RT}_{60}$ estimates were summarised into a ‘broadband $\text{RT}_{60}$’ measure by taking the median (after Traer and McDermott, 2016).
Figure 2. Depictions of each of the 10 locations used in the experiment. The locations are shown here under equirectangular projection to convey the entire light field and are ordered by reverberation time (RT60; see Fig. 3). During the experiment, the locations were presented either in HMD and surrounded the viewer or on a computer monitor in first-person perspective.
Figure 3. Summaries of the auditory impulse response of each of the 10 locations used in the experiments. Each panel shows a spectrogram (averaged over left and right channels) of the impulse response, with log-scaled time (horizontal) and frequency (vertical) axes. The points show the valid RT$_{60}$ estimates at octave-spaced frequencies, with neighbouring frequencies joined by solid lines. The solid vertical line shows the ‘broadband RT$_{60}$’, which summarises the set of frequency-specific RT$_{60}$ values at a given location. The locations are ordered by the estimates of broadband RT$_{60}$. 
To select the 10 locations that were used in this study, we first characterised each location on the ‘global scene features’ (dominant depth, openness, and perspective) that were described by Oliva and Torralba (2006). The first-author of this article (KT) provided ratings between 1 and 6 on each of these dimensions (following Ross and Oliva, 2010), for each of 10 regions on a cube-mapped representation of each location’s panorama (comprised of eight rotations and an upward and a downward elevation), in addition to a qualitative labelling of the most prevalent surface material (e.g., wood, carpet). Each location was then summarised by the mean and standard deviation of the ratings across the eight rotations and the rating for the downward elevation, for each of the three global scene feature dimensions — giving nine dimensions in total. We then performed a principal components analysis to reduce the dimensionality from 9 to 2 and identified 15 locations that were approximately equally distributed across the positive and negative values of both principal components. Finally, we removed five locations to obtain an approximately equal distribution across the range of impulse response reverberation times (RT60), and an equal number of locations with reflective (e.g., cement, metal, etc.) and absorbent (e.g., wood, sand, etc.) surface materials.

The auditory stimuli used in the experiments were created by convolving short utterances with the impulse response associated with each of the locations. These utterances were dry recordings of two male and three female speakers uttering the words ‘one’ or ‘five’ (Eaton et al., 2016). The auditory stimuli in the HMD experiment produced an output level between 70 dB SPL (sound pressure level) and 85 dB SPL (LAFmax), as determined by an artificial ear, microphone, and analyser (Brüel and Kjær, Nærum, Denmark; models 4152, 4144, and 2250, respectively). In the web experiment, each participant was asked to set an output level such that the stimulus with the highest root-mean-square signal was “loud but not uncomfortable”.

The auditory stimuli used in ‘catch’ trials were created by convolving the utterance with a temporally-reversed impulse response. The purpose of catch trials was to have stimuli for which a particular response was always expected if participants understood the task requirements and were engaged in the completion of a task on a given trial. We chose to use a temporal reversal of the impulse response in catch trials as Traer and McDermott (2016) demonstrated that such reversals cause the resulting sounds to be perceived as deviant from those arising from natural impulse responses (i.e., impulse responses originating from the 10 selected locations).

2.4. Design

Each pairwise combination of auditory and visual environment was presented once, for a total of 100 trials (10 auditory environments and 10 visual environments) per participant. We also included 20 catch trials, which used the
temporally-reversed impulse responses as auditory stimuli. The responses on catch trials were used to exclude participants from the analysis and to incorporate the propensity for an off-task response into the analysis of the experimental trials (see subsection 2.6. Analysis). Each participant thus completed a total of 120 trials, composed of 100 experimental trials and 20 catch trials.

2.5. Procedure

For participants in the HMD experiment, their interpupillary distance was first measured and used to adjust the distance of the lenses inside the headset. The participant then put on the headset and headphones with assistance from the experimenter, ensuring a comfortable fit by fastening the fitting straps. The participants confirmed their handedness, and were given the corresponding controller to interact with the experiment.

For participants in the web experiment, they first completed a set of questions relating to their demographics and computing and audio hardware. To assess compliance with the requirement to participate using headphones or earphones, participants then completed a headphone screening process consisting of ‘binaural beat’ and ‘Huggins pitch’ tests (see Milne et al., 2021, for details).

Each participation session consisted of an exploration phase followed by an experimental phase. Both phases of the experiment involved participants viewing various locations. For participants in the HMD experiment, they observed the entirety of the visual environment by turning their heads and turning the viewpoint rotation using the thumbsticks on the controller. The thumbstick could be pushed left or right to rotate the viewpoint of the virtual environment by 45° left or right respectively, for a total of eight viewpoints. The participants were also able to move the angle of viewpoint within the virtual environment by rotating the headset. However, if the participant rotated their head too far left or right, they would be notified that they should instead use the thumbstick to rotate the viewpoint. This rotation restriction was put in place to prevent awkward head movements and risk of neck discomfort. For participants in the web experiment, the environments were presented with their browser in fullscreen mode and with a pointer lock applied to their mouse. Participants observed the entirety of the visual environment by moving their mouse to rotate the viewpoint of the camera. If participants happened to manually exit fullscreen mode at any point during an exploration or experimental trial, the environment was replaced with a grey screen that required them to re-enter fullscreen mode.

The exploration phase allowed participants to familiarise themselves with the visual aspects of each location that were to be used in the experimental trial phase. During the exploration phase, participants were consecutively presented with the 10 visual environments of the locations in random order. They
were instructed to explore visual environments to “understand the environments as best as they could” while considering how imaginary speech might sound if it were produced in the environment shown. Auditory stimuli were not presented during the exploration phase. To ensure that exploration was diligent, participants could only advance to the next environment after having viewed the entirety of the current environment for an ample period of time. Each time the participant entered a visual environment, they would start from a randomly selected viewpoint. From this viewpoint, they were required to explore the environment for at least 10 s and to have rotated around the complete environment. Exploration had no maximum time restrictions, so participants could continue to explore the environment if they did not feel comfortable with their knowledge of the location after the required time period. After they were finished exploring, they could advance to the next trial by simply pressing a button (on the controller for the HMD experiment or on the keyboard for the web experiment). When advancing trials in the HMD experiment, the screen faded to black briefly (1.5 s) before fading in the next visual environment.

After the completion of the exploration phase, participants engaged in the experimental phase. During this phase, a visual target (a faceted red sphere) was rendered at a particular location within the environment (see Fig. 4). The positioning of the target within each location was identical for all trials. When the headset (HMD) or mouse (web) was aligned with the visual target, the target turned blue and the utterance for that trial was delivered. After the auditory stimulus ceased, participants were required to respond whether or not the sound was produced from within the depicted visual location (“Was the speech sample recorded in the visual environment that you saw?”).

The 120 trials of the experiment were presented in random order, consisting of 100 experimental trials and 20 catch trials, separated into eight runs of 15 trials each. Between each run, there was a mandatory resting period of 30 s, in which the display in the headset faded to black (HMD) or the fullscreen mode was exited (web).

After completing all the experimental trials, participants completed the Igroup Presence Questionnaire (IPQ; Schubert et al., 2001) to assess their subjective experience of the environment. The questionnaire consisted of 14 statements that were consecutively presented through text within the headset (HMD) or browser (web). The first statement assessed the validity of the scale in the construction of the questionnaire. The following 13 statements each assessed one of three facets (spatial presence, involvement and experienced realism) regarding the sense of presence experienced in the virtual environments used in the experiment. Participants answered each question by selecting one response that depicted their judgement towards a statement on a
Figure 4. Example renderings of the visual environment and target. The embedded target was red in colour (A) before changing to blue when aligned with the central crosshair (B) following appropriate movement of the participant’s head (HMD experiment) or the participant’s mouse (web experiment).

seven-point Likert scale. The endpoints of the scale were labelled with negative (e.g., “fully disagree”, “not really at all”) and positive (e.g., “fully agree”, “completely real”) assessments, and participants used the thumbstick (HMD) or mouse (web) to indicate their response within this scale.

The entire participation session lasted approximately 45 min. A demonstration of the experimental procedure is given by the recording of an example experimental session for a participant in the web delivery method (see Supplementary Video S1).

2.6. Analysis

The code used to perform the analysis is available alongside the raw data and a computational platform (Clyburne-Sherin et al., 2019) at doi: 10.24433/CO.3600071.v3.
2.6.1. Exclusions
We excluded participants who exceeded the threshold of incorrect responses on the catch trials that we considered to be indicative of a lack of diligence or understanding in completing the task (four or more of the 20 catch trials incorrect). From the 56 participants we tested in the HMD experiment, we excluded 12 participants (21%) from the data analysis. From the 130 participants we tested in the web experiment, 47 participants (36%) did not meet this exclusion criterion. For the web experiment, we also evaluated three addition exclusion criteria. First, we identified participants who gave a negative response to the prompt “Did you complete the task with appropriate diligence and attention?” (14/130; 11%). Second, we identified participants who took longer than 45 min to complete the primary task (9/130; 7%). Finally, we identified participants who did not respond correctly to more than four of the six trials on one of the two headphone check tasks (27/130; 21%). Based on the data of Milne et al. (2021), we would expect this criterion to successfully identify approximately 67% of participants using speakers and to mistakenly identify approximately 7% of participants using headphones. Participants were excluded from further analysis if they met any of these four exclusion criteria, resulting in 70 participants from the web experiment (54%) being excluded.

2.6.2. Observations
Each experimental session produced a set of binary values that each indicated whether the judgement on the associated trial was that the sound was (1) or was not (0) produced in the depicted visual environment. There were 100 of such values per participant, composed of trials from each pairwise combination of the auditory and visual components of each of the 10 locations. With 104 participants satisfying the exclusion criterion, the statistical analysis was based on data from 10 400 trials. We represent this data as a three-dimensional array \((y)\), indexed by the coordinates \(i\) [participants; \(i \in \{1, \ldots, 104\}\)], \(j\) (auditory location; \(j \in \{1, \ldots, 10\}\)), and \(k\) (visual location; \(k \in \{1, \ldots, 10\}\)).

Participants also completed 20 catch trials per session, which we represent as the number of incorrect trials per participant \((y^w)\). Each experimental session also produced a set of 14 integer responses (coded as values between 1 and 7, negatively-scored where appropriate) from the IPQ. We summed the responses of the first 10 items, after subtracting 1 from each item’s score to convert its range to between 0 and 6, to give a single ‘presence’ score per participant out of 60 \((y^q)\). The remaining items, which corresponded to the ‘experienced realism’ factor, were excluded from the analysis because our experimental manipulation of congruence was confounded with the experienced realism in the simulated environments; the incongruent pairs were designed to have an audiovisual mismatch, and therefore were inherently unrealistic.
2.6.3. Statistical Model

Our primary analysis goal was to estimate the effect sizes of a set of potential influences on participant judgements of the congruence of auditory and visual environments. To achieve this goal, we constructed a statistical model that encompassed each of the three observational variables (congruence judgements, catch trial responses, and presence scores) and contained predictors that quantified each of the potential influences of interest. Here, we provide an overview of the key details of our statistical modelling approach — we encourage the reader to consult the thorough description that is provided in the Appendix for specific details.

The foundation of the model structure for each of the three observational variables was provided by a generalised linear mixed model (see Moscatelli et al., 2012, for an introduction in a psychophysics context). The use of the generalised linear mixed model allowed for predictions to be made on the probability of a match response for a given trial, while accounting for fixed and random effects. The model for the congruence judgements included fixed effects for the intercept, the within-subjects factor of audiovisual congruence, the between-subjects factor of delivery method, and the interaction between audiovisual congruence and delivery method. It also included random participant effects on the intercept and on the audiovisual congruence factor, in addition to random ‘item’ effects (Judd et al., 2017; Rouder and Lu, 2005) for auditory locations and visual locations on the intercept and their interactions with participants. The probability of a match response based on the outcome of this model was also adjusted to account for the propensity of participants to ‘lapse’ on a given trial (i.e., to respond randomly and independently of the congruence task requirements). These lapse rates were informed by the model of the catch trial responses, which included a fixed intercept and a participant random effect. Finally, the model for the presence scores included fixed effects for the intercept and the between-subjects factor of delivery method and a random participant effect on the intercept. The random participant effects, present across the models for the observational variables, were allowed to covary.

We used a Bayesian framework (Lee, 2018; van de Schoot et al., 2021; Wagenmakers et al., 2018) to determine the posterior probability distribution of the statistical model parameters given our observed data. The Bayesian framework requires the specification of prior distributions for the model parameters. We adopted the strategy of providing weakly informative priors, where the informative aspect is on the approximate scale of the parameter values (see the Appendix for the definition of each prior). The model was implemented in PyMC3 (version 3.11.4; Salvatier et al., 2016), and Markov chain Monte-Carlo (MCMC) sampling was performed using its implementation of a No-U-turn sampler (Hoffman and Gelman, 2014). A total of 4000 draws were used for each of four independent chains in the sampling process,
after discarding the initial draws (1000) used in initializing the sampler. Sampling quality was assessed by evaluation of posterior predictive distributions (Betancourt, 2020), sampling traces, autocorrelations, and sampling metrics.

As our primary goal is to estimate the values of the parameters, rather than to perform formal hypothesis testing or model comparison (Calin-Jageman and Cumming, 2019), reporting summaries of the posterior distributions is our primary method of communicating the outcomes of the study (Kruschke and Liddell, 2018). We use the median as the measure of centrality and credible intervals (equal-tailed intervals, calculated using quantiles; Makowski et al., 2019) as the measure of uncertainty when summarising the posterior distributions. To increase the interpretability of the reported summaries, we typically transform the posterior samples from the model units (log-odds) to more familiar units. The use of log-odds as units in the model arises because of our use of a logit link function in the generalised linear mixed models. We use this link function because it allows the model to be specified on an unbounded scale; each of the observational variables was assumed to be generated by processes with parameters that were constrained to the [0, 1] interval (the probability of responding ‘match’, the probability of producing an incorrect response on a catch trial, and the proportional degree of presence), and the logit function maps a value in the unit interval (p) to an unbound interval by taking the logarithm of the odds, where the odds are defined as p/(1 − p).

To investigate the aspects of the congruence task responses that were not accounted for by our statistical model, we examined the residuals (Gelman et al., 2000). We computed the residuals as the observed congruence task data (y) minus the posterior probability of a match response for each trial according to the model. The distribution of residuals that would be expected based solely on the model was simulated by using posterior predictive draws of congruence task data in place of the observed data in the above calculation.

3. Results

In this study, we presented participants with a series of first-person perspectives of scenes either through a HMD or through a computer monitor. In each scene, participants viewed the panorama of a location and heard a sound that was rendered with the acoustic properties of a (potentially different) location. Participants judged the congruence of the auditory and visual signals; that is, whether they thought the sound had been produced in the depicted visual environment (a ‘match’). Such judgements were collected for each pairwise combination of the auditory and visual components of 10 locations, giving 10 congruent trials and 90 incongruent trials per participant.
We formed a statistical model of the congruence judgements and used Bayesian procedures to estimate the model parameters. To motivate the construction of the model and provide an intuition for the model parameters, we begin by describing a decomposition of the average of the observed responses across all participants. As shown in Fig. 5A, these data can be represented as an image in which the horizontal and vertical axes represent the auditory and visual locations, respectively, and the brightness of each cell is related to the average proportion of trials in which participants responded ‘match’.

We are particularly interested in whether the proportion of match responses differs for trials in which the auditory and visual locations are congruent, relative to trials in which the auditory and visual locations are incongruent. As shown in Fig. 5B, this can be found by comparing the average proportion of match responses where the auditory and visual locations are the same (the leading diagonal) to the average proportion of match responses where the auditory and visual locations differ (the remaining cells). In subsection 3.1. Congruence below, we report the posterior distribution for a model parameter which captures this difference — we also quantify how this difference varies with the two delivery methods used in this study and with the participant self-reported experience of ‘presence’ in subsections 3.2. Experiment Delivery Method and 3.4. Presence below, respectively.

We also seek to examine how the propensity to respond ‘match’ varies across the 10 different locations used in this study. To do so, we can measure the average proportion of match responses for the auditory component of each location (across all visual components; Fig. 5C) and for the visual component of each location (across all auditory components; Fig. 5D). In subsection 3.3. Location-Specific Influences below, we summarise the posterior distributions for parameters that describe how the propensity to respond ‘match’ varies across locations.

We can use the influences on the propensity to respond ‘match’ that we have described thus far (congruence, auditory location, and visual location) to form a ‘prediction’ of the average proportion of match responses that would be expected if these were the only influences that affected the propensity to respond ‘match’. Subtracting these predicted responses (Fig. 5E) from the observed responses (Fig. 5A) gives us the residuals (Fig. 5F), which reveals the additional influence of each specific auditory and visual location pairing on the propensity to respond ‘match’. In subsection 3.3. Location-Specific Influences below, we examine the magnitude of the residuals from the statistical model both within the congruent pairings and within the incongruent pairings.

3.1. Congruence

Our primary interest in this study was to estimate the effect of the auditory and visual components of a given presentation being ‘congruent’ on the propensity
Figure 5. Illustration of the analysis approach using summaries of the observed data (averaged over all participants). Each panel has the auditory location number and visual location number on the horizontal and vertical axes, respectively, and has the brightness of each cell related to the proportion of match responses (panels A–E) or the value of the residual (panel F). Panel A shows the raw averages for each auditory and visual pairing. Panel B shows the average proportion of match responses for the congruent pairings (the leading diagonal) and the incongruent pairings (the remaining cells). Panels C and D show the average proportion of match responses for each auditory location and each visual location, respectively. Panel E shows the sum of the effects depicted in panels B, C, and D. Panel F shows the difference between the observed data and such predicted responses (the residuals).
Figure 6. Summary of the fixed-effect contributions to the match responses. In each panel, the dotted lines show the aggregate of the observed match responses for a given marginalisation. The histograms show the distribution of such aggregations performed on posterior predictive samples from the fitted model. Panels A, B, and C show summaries of the congruence effect, the experiment effect on the intercept, and the interaction between the congruence and experiment effects, respectively.

for a participant to respond that the speech sample was produced within the depicted visual environment. As shown in Fig. 6A, the observed proportion of match responses (aggregated over all participants, experiments, and locations) was higher for congruent trials (0.64) than for incongruent trials (0.47). From the statistical model, we estimate that the congruent presentations were associated with an increase in the probability of a match response of between about 0.22 (95% credible interval: [0.17, 0.27]) relative to the incongruent presentations.
3.2. Experiment Delivery Method

Because the study comprised two different experiments, we were also interested to estimate the potential effects of the between-experiment difference — whether the session was delivered using a HMD or web method. As shown in Fig. 6B, the experiment had little apparent effect on the overall propensity to respond ‘match’ (the intercept); the observed proportion of match responses was 0.48 for the HMD experiment and 0.48 for the web experiment, with the statistical model indicating that the web experiment was associated with a change in the probability of a match response of between about a 0.06 decrease and a 0.08 increase (95% credible interval; posterior median: 0.01) relative to the HMD experiment. The proportion of match responses for congruent trials was higher for the web experiment (0.65) than for the HMD experiment (0.61) despite there being similar proportions on match responses on incongruent trials (0.47 for both HMD and web), as shown in Fig. 6C. From the statistical model, we estimate that the difference in the probability of responding ‘match’ on congruent relative to incongruent presentations was between about 0.04 lower and about 0.13 higher (95% credible interval; posterior median: 0.04) for the web experiment relative to the HMD experiment.

3.3. Location-Specific Influences

Our statistical model included ‘random’ effects for the auditory and visual components of each location, which captured heterogeneity in the propensity for a match response based on the modality-specific features of a given location. The posterior distributions for each of the locations for the auditory and visual modalities are shown in Fig. 7. For the auditory modality (Fig. 7A), we note an apparent association in which the locations with longer reverberation times (those with higher location numbers) have a decreased probability of responding ‘match’. For the visual modality (Fig. 7B), we find that locations 4 and 5 are notable for their association with a decrease in the probability of a match response.

To explore the potential influence of individual pairings of specific auditory and visual locations, which are not captured in our statistical model, we examined the residuals. We calculated the difference between the observed data and the posterior response probabilities, both averaged over participants. We first examined the congruent combinations (Fig. 8), which showed locations 8, 3, and 9 to be particularly notable — with location 8 showing a higher proportion of match responses and locations 3 and 9 showing a lower proportion of match responses than would be expected based on the statistical model. We
then explored the magnitude of the residuals within the 90 auditory and visual location pairings that comprise the incongruent trials. During such exploration, we noted the apparent relationship that is depicted in Fig. 9. This visualisation and accompanying linear regression model suggests that the probability of incorrectly responding ‘match’ on an incongruent trial is negatively related to the similarity of the reverberation characteristics of the two locations; that is, incongruent trials where the visual modality depicts a location with a reverberation time that is similar to the reverberation time of the location rendered via the auditory modality have a higher than predicted probability of a match response.

**Figure 7.** Posterior distributions for the parameters capturing an effect of auditory (A) and visual (B) locations. In each panel, the vertical axis indicates the change in the probability of a match response (relative to the overall mean) and the horizontal axis shows the location number (which is a ranking based on the location’s reverberation time). The ribbons show credible intervals estimated from the posterior distributions (with the lightest grey representing the 95% credible interval and greys of decreasing intensity representing the 90%, 80%, 60%, 40%, and 20% credible intervals), and the white lines indicate the medians of the posterior distributions.
Figure 8. Comparison of the observed mean proportion of match responses for congruent trials and those derived from posterior response probabilities from the statistical model. The horizontal bars indicate the 95% credible interval of the distribution of the residual summary formed from posterior predictive samples and give an indication of the magnitude of the summarised residuals that would be expected under the model. The ribbons are as described in Fig. 7.

Figure 9. Comparison of the observed mean proportion of match responses for incongruent trials and those derived from posterior response probabilities from the statistical model. Each point shows the median of the difference between the observed data and posterior response probabilities for a particular pair of incongruent auditory and visual locations, with the 95% credible interval shown by the vertical lines. The line and ribbons show the estimated trend from a linear regression model, and are as described in Fig. 7.
3.4. Presence

We also used a questionnaire to measure the degree of ‘presence’ that was subjectively experienced by participants during the session, quantified as a single number between 0 and 60 for each participant. The average presence ratings were higher for participants in the HMD experiment (27.14) compared to those in the web experiment (23.60); from the statistical model, we estimate that the HMD experiment was associated with an increase of about 17.14% (95% credible interval: [0.00%, 37.23%]) in the reported presence relative to the web experiment. We estimate that the correlation of individual differences in the self-reported presence with the overall propensity to respond ‘match’ was between about −0.15 and 0.29 (95% credible interval; posterior median: 0.07) and with the ability to discriminate congruent and incongruent trials was between about −0.52 and 0.72 (95% credible interval; posterior median: 0.20).

4. Discussion

In this study, we investigated the capacity for human participants to accurately relate the auditory and visual characteristics of simulations of real-world locations. The primary finding of this study is that the propensity to perceive a sound as having been produced in a particular visual environment was greater when the sound was indeed produced in the physical location depicted by vision. This suggests that participants were able to extract diagnostic acoustic features from visual information, consistent with the previous reports by Defays et al. (2014) and Sandvad (1999). It is potentially inconsistent with the findings of McCreery and Calamia (2006), who asked participants to reproduce the apparent level of reverberation in a visual reference and reported that “reverberance levels chosen by subjects rarely match those measured from impulse responses recorded in the rooms being presented” (p. 3150) — however, it is difficult to gain further insight into the apparent discrepancy as these findings were only reported in abstract form. More broadly, it suggests that vision may be able to assist in the separation of the source-related and environment-related ingredients to the acoustic signals at the ear by contributing an estimate of the environmental component.

Although the specific visual features that underlie the ability to estimate the acoustical properties of an environment are currently unclear, we can perhaps gain some insight by exploring the observed heterogeneity in congruent trial performance across locations. As shown in Fig. 8, location 8 was notable for its higher than expected proportion of ‘match’ responses on congruent trials. When exploring its panorama (Fig. 2), it is apparent that there is a dominating spatial layout (a long straight tunnel) and surface material (metallic enclosure). The high levels of match responses in this location suggests that these
features are likely to be readily translatable into an accurate auditory expectation. However, the relatively lower levels of match response that are evident for locations 3, 7, and 9 do not seem to have any readably identifiable correlates in their panoramas — clearly more research, with a broader or more controlled set of visual stimuli, is required to understand the visual features that underlie the capacity to infer auditory properties from visual information.

However, location 9 does demonstrate an intriguing limitation of vision when inferring acoustical properties due to the presence of transparent surfaces. It is ambiguous from the visual information available in the panorama (Fig. 2) whether or not the frames in the enclosure contain glass. Thus, a set of surfaces that have implications for acoustical propagation are subject to a high degree of perceptual interpretation from the available visual signals. It would be interesting whether the inferred acoustic properties are indeed affected by the interpretation of visually transparent or absent surfaces — and, conversely, whether the interpretation of visually transparent or absent surfaces is affected by auditory signals. This latter point highlights that although we have emphasised the role of vision in being informative of the acoustical properties of a location — both because of the broad motivation of the study (whether vision could assist in the separation of the source and environmental contributions to auditory signals) and because of our implementation of the task (presenting the visual stimulus prior to the auditory stimulus) — the current task could also be performed by generating estimates of visual features from audition. For example, the degree of reverberation could be used to estimate the room size which is then compared against the visual estimate of room size as the basis for the judgement of congruence. Indeed, Pop and Cabrera (2005) interpreted the results of Sandvad (1999) in this direction. While these explanations are seemingly reciprocal in the current study, the distinction may have implications for the involvement of vision in the problem of separating source and environment in auditory signals.

The claim that vision is often capable of supporting accurate estimates of environmental acoustics is supported by the pattern of match responses in the incongruent trials. As shown in Fig. 9, there is an apparent trend in which the propensity for a match response declines as the absolute difference between the reverberation times (broadband RT<sub>60</sub>) increases. That is, visual environments that depict locations with similar (but not identical) reverberation times to the location of the auditory stimulation are more likely to be incorrectly reported as being congruent. This response pattern would be expected for estimates of reverberation time that are subject to uncertainty; although the precision with which we can estimate the reverberation times from the visual and auditory renderings of the current environments is unclear, such estimates would undoubtedly contain a (potentially high) degree of uncertainty.
This apparent pattern of responding in incongruent trials also suggests that signal detection theory may provide a useful computational foundation for expanding on the insights obtained from the current statistical model. In an application of this framework, each trial can be considered to evoke internal estimates of the auditory characteristics of the environment (such as reverberation time) — an estimate from the visual information and an estimate from the auditory information. Each estimate is conceived as being drawn from a normal distribution with a particular mean and standard deviation. To perform the task, the estimates from the two modalities are compared and a match response is elicited if the resulting absolute difference is less than a criterion.

The role of the criterion in such a signal detection theory framework would provide a natural interpretation for the observed variation in the propensity to respond ‘match’ for the different locations. Performing the current task under the application of the signal detection framework mentioned above requires the conversion of the continuous quantity (the absolute difference in estimated reverberation times) into a binary decision according to a criterion value that is under the potential control of the participant. As shown in Fig. 7A, the willingness of participants to respond ‘match’ declined with the reverberation time of the auditory location — which could be interpreted as a change in criterion, with participants requiring stronger evidence to respond ‘match’ with higher reverberation times. Participants might have adopted this strategy in recognition that the higher reverberation times are generally less likely to be encountered in the world. To quantify this likelihood, we estimated the reverberation times in the database of experience-sampled real-world locations that were collected by Traer and McDermott (2016). As shown in Fig. 10, the locations with higher levels of reverberation are particularly unusual with respect to the general distribution of typical reverberation times — supporting the sensibility of using a strategy that errs on the side of responding ‘nonmatch’ for such reverberation times.

Alternatively, a similar association between the propensity to respond ‘match’ across auditory locations with reverberation time could also be present with a constant criterion. With the same criterion, an asymmetry could emerge due to the nonuniform sampling of reverberation times — a location that has a higher number of other locations with similar reverberation times would be expected to attract more match responses than a location with a reverberation time that is more dissimilar to the others. Within the current set of locations, most of the reverberation times are towards the lower end of the distribution and would hence be compatible with the observed pattern of match responses. Indeed, the predictions from the current set of locations would be nonmonotonic with a peak between 0.75 s and 1 s — which resembles the observed pattern shown in Fig. 7A.
Figure 10. Comparison of the reverberation times of the auditory locations used in the current study (crosses) against the cumulative probability of the reverberation times in the environments collected by Traer and McDermott (2016).

Under either interpretation of the variation in the overall tendency to respond ‘match’ across auditory locations, we would also expect a similar relationship with the reverberation time of the visual location — participants should be less willing to respond ‘match’ when the visual environment depicts a location with a high reverberation time. However, the variation in the propensity to respond ‘match’ across visual environments seems unrelated to the reverberation time of its location. It is possible that the likely decrease in the precision with which reverberation time can be estimated from visual relative to auditory signals causes the estimates of reverberation time from vision to be ignored when selecting an appropriate criterion. Instead of reverberation time, the most salient feature that is evident when comparing the panoramas of those locations with differences in their propensity to respond ‘match’ (that is, when comparing the panoramas shown in Fig. 2 against the values shown in Fig. 7B) is whether the location is indoors or outdoors. This distinction between indoor and outdoor environments is relatable to the visual property of ‘openness’, which is defined in scene categorisation research as the degree of spatial enclosure or expanse of a depicted visual environment (Greene and Oliva, 2009; Zhang et al., 2018). The degree of apparent openness may induce a change in criterion in which more open scenes are required to have stronger evidence to elicit a match response. Alternatively, the reverberation time estimates from more open scenes may be less precise than those from more closed scenes — with this difference in precision translating to a difference in the propensity to respond ‘match’ with a constant criterion.
Participants viewed the visual components of this study either using a HMD or using a web-based visualisation via a standard computer monitor and listened to the auditory components of the study using standardised equipment (in the HMD experiment) or diverse and readily-available equipment (in the web experiment). Despite the large differences between these delivery methods, it is notable how little of an effect they seemed to have on participant task-related behaviour in the context of the current study. The magnitude of the key effect, the difference between responses in congruent and incongruent trials, was similar between the two delivery methods. If anything, the congruence effect was higher in the web-based delivery — which is opposite to what would be expected if immersion in the visual environment was critical to the congruence effect. This suggests that both presentation methods are capable of providing the audiovisual features that are used to infer the relationship between the auditory and visual properties of the current set of locations — and demonstrates the viability of web-based delivery methods.

The delivery method may have had a small effect on the subjective experience of participants during the session, with participants in the HMD experiment tending to report higher degrees of presence than those in the web-based experiment. The variation in presence across participants, relative to the experiment means, was weakly associated with an increase in both the overall propensity to respond ‘match’ and to respond ‘match’ on congruent trials more than on incongruent trials — although the estimates of such associations have a high degree of uncertainty due to the sample sizes used in this study, which are relatively low for examining individual differences. The apparent viability of web-based delivery increases the feasibility of future studies with larger sample sizes, which can increase the precision of such estimates.

We note a number of potential limitations that require consideration when interpreting the results of this study. First, the generality of the conclusions to a different set of locations is uncertain. We were only able to use a relatively small number of locations (10), and those locations are somewhat nonrepresentatively sampled in that their parent database comprised locations that were unique and unusual in their auditory properties (also see Fig. 10). Second, the high proportion of participants that were excluded from analysis (particularly in the web-based experiment) may make the generality to different participants questionable. However, while acknowledging the issues around excluding such a large number of participants, we believe that our exclusion criteria were appropriate and, importantly, are unlikely to bias the results by selectively targeting subpopulations of participants who differ in their outcome-relevant behaviour (for example, those unable to distinguish congruent and incongruent presentations are not selectively excluded on that basis). Third, the visual elements used in the current study do not fully engage with the strengths of virtual reality available to HMD users, which has implications for the generalisation...
of the comparisons between the two delivery methods. By using panoramic images rather than renderings of simulated three-dimensional environments, we were unable to utilise the stereoscopic capabilities of the HMD. Furthermore, panoramic images such as those used in this study typically suffer from image quality issues and stitching errors (Ritter and Chambers, 2021), which may be more readily apparent in the wide field of view of the HMD. Finally, we have focused on reverberation time (quantified via the broadband RT<sub>60</sub>) as a summary of the auditory properties of a location and have not considered other potential metrics. For example, the frequency dependence of reverberation (Traer and McDermott, 2016) may also be an auditory feature that is able to be related to vision. Furthermore, there may be additional proximal metrics (i.e., those present in the auditory waveform) that are indicative of the auditory properties of a location, such as those described by Kolarik et al. (2016) and Kopčo and Shinn-Cunningham (2011) as covarying with the ratio of direct to reverberant energy.

In summary, we report that participants showed a capacity to relate the sound and sight of a set of simulated real-world locations; audiovisual congruence was associated with an increase in the probability of participants reporting an audiovisual match of about 0.22 (95% credible interval: [0.17, 0.27]). This capacity is consistent with a potential role for vision in inferring the relative contributions of source and environment to the auditory signals available at the ear. Future research could be directed towards clarifying the audiovisual features that are used to relate the two modalities and towards assessing the functional consequences of visual information for the perceptual interpretation of auditory signals.

Acknowledgements

This research is supported by an Australian Government Research Training Program (RTP) Scholarship (to KT). We thank Chris Warren for making the Echo Thief database publicly available.

Supplementary Material

Supplementary material is available online at: https://doi.org/10.6084/m9.figshare.20554872

References

Betancourt, M. (2020). Towards a principled Bayesian workflow. Retrieved from: https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html.
Brandewie, E. and Zahorik, P. (2010). Prior listening in rooms improves speech intelligibility, J. Acoust. Soc. Am. 128, 291–299. DOI:10.1121/1.3436565.
Calcagno, E. R., Abregú, E. L., Eguía, M. C. and Vergara, R. (2012). The role of vision in auditory distance perception, *Perception* 41, 175–192. DOI:10.1068/p7153.

Calin-Jageman, R. J. and Cumming, G. (2019). Estimation for better inference in neuroscience, *eNeuro* 6, ENEURO.0205-19.2019. DOI:10.1523/ENEURO.0205-19.2019.

Clyburne-Sherin, A., Fei, X. and Green, S. A. (2019). Computational reproducibility via containers in psychology, *Meta-Psychology* 3, MP.2018.892. DOI:10.15626/MP.2018.892.

Cutone, M. and Wilcox, L. (2018). PsychXR (version 0.2.0). Available from https://github.com/mdcutone/psychxr.

de Leeuw, J. R. (2014). jsPsych: a JavaScript library for creating behavioral experiments in a web browser, *Behav. Res. Methods* 47, 1–12. DOI:10.3758/s13428-014-0458-y.

Defays, A., Safin, S., Billon, A., Decaestecker, C., Warzée, N., Leclercq, P. and Nyssen, A.-S. (2014). Bimodal interaction: the role of visual information in performing acoustic assessment in architecture, *Ergonom. Open J.* 7, 13–20. DOI:10.2174/1875934301407010013.

Eaton, J., Gaubitch, N. D., Moore, A. H. and Naylor, P. A. (2016). Estimation of room acoustic parameters: the ACE challenge, *EEE/ACM Trans Audio Speech Lang. Proc.* 24, 1681–1693. DOI:10.1109/TASLP.2016.2577502.

Gelman, A., Goegebeur, Y., Tuerlinckx, F. and Van Mechelen, I. (2000). Diagnostic checks for discrete-data regression models using posterior predictive simulations, *J. R. Stat. Soc. Ser. C* 49, 247–268. DOI:10.1111/1467-9876.00190.

Greene, M. R. and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees, *Cogn. Psychol.* 58, 137–176. DOI:10.1016/j.cogpsych.2008.06.001.

Harris, R. W. and Reitz, M. L. (1985). Effects of room reverberation and noise on speech discrimination by the elderly, *Audiology* 24, 319–324. DOI:10.3109/00206098509078350.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.* 15, 1593–1623.

Judd, C. M., Westfall, J. and Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power, *Annu. Rev. Psychol.* 68, 601–625. DOI:10.1146/annurev-psych-122414-033702.

Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S. and Pardhan, S. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss, *Atten. Percept. Psychophys.* 78, 373–395. DOI:10.3758/s13414-015-1015-1.

Kopčo, N. and Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources, *J. Acoust. Soc. Am.* 130, 1530–1541. DOI:10.1121/1.3613705.

Kruschke, J. K. and Liddell, T. M. (2018). The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective, *Psychon. Bull. Rev.* 25, 178–206. DOI:10.3758/s13423-016-1221-4.

Lange, K., Kühn, S. and Filevich, E. (2015). “Just Another Tool for Online Studies” (JATOS): an easy solution for setup and management of web servers supporting online studies, *PLoS ONE* 10, e0130834. DOI:10.1371/journal.pone.0130834.

Lee, M. D. (2018). Bayesian methods in cognitive modeling, in: *Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience, Vol. 5*, T. Wixted (Ed.), pp. 37–84. John Wiley and Sons, Inc., New York, NY, USA.
Lewandowski, D., Kurowicka, D. and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method, *J. Multivar. Anal.* **100**, 1989–2001. DOI:10.1016/j.jmva.2009.04.008.

Makowski, D., Ben-Shachar, M. S. and Lüdecke, D. (2019). bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework, *J. Open Source Softw.* **4**, 1541. DOI:10.21105/joss.01541.

McCreery, A. and Calamia, P. (2006). Cross-modal perception of room acoustics, *J. Acoust. Soc. Am.* **120**, 3150. DOI:10.1121/1.4787814.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J. and Chait, M. (2021). An online headphone screening test based on dichotic pitch, *Behav. Res. Methods* **53**, 551–1562. DOI:10.3758/s13428-020-01514-0.

Moscatelli, A., Mezzetti, M. and Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: the generalized linear mixed model, *J. Vis.* **12**, 26. DOI:10.1167/12.11.26.

Oliva, A. and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition, *Prog. Brain Res.* **155**, 23–36. DOI:10.1016/S0079-6123(06)55002-2.

Pop, C. B. and Cabrera, D. (2005). Auditory room size perception for real rooms, in: *Proc. ACOUSTICS 2005*, Busselton, Western Australia, pp. 115–121.

Ritter III, K. A. and Chambers, T. L. (2021). Three-dimensional modeled environments versus 360 degree panoramas for mobile virtual reality training, *Virt. Real.* **26**, 571–581. DOI:10.1007/s10055-021-00502-9.

Ross, M. G. and Oliva, A. (2010). Estimating perception of scene layout properties from global image features, *J. Vis.* **10**, 2. DOI:10.1167/10.1.2.

Rouder, J. N. and Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection, *Psychon. Bull. Rev.* **12**, 573–604. DOI:10.3758/ BF03196750.

Salvatier, J., Wiecki, T. V. and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3, *PeerJ Comput. Sci.* **2**, e55. DOI:10.7717/peerj-cs.55.

Sandvad, J. (1999). Auditory perception of reverberant surroundings, *J. Acoust. Soc. Am.* **105**, 1193. DOI:10.1121/1.425625.

Schubert, T., Friedmann, F. and Regenbrecht, H. (2001). The experience of presence: factor analytic insights, *Presence (Camb.)* **10**, 266–281. DOI:10.1162/105474601300343603.

Schutte, M., Ewert, S. D. and Wiegrebe, L. (2019). The percept of reverberation is not affected by visual room impression in virtual environments, *J. Acoust. Soc. Am.* **145**, EL229. DOI:10.1121/1.5093642.

Te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A. and Konig, R. (2017). When size matters: advantages of weighted effect coding in observational studies, *Int. J. Public Health* **62**, 163–167. DOI:10.1007/s00038-016-0901-1.

Traer, J. and McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space, *Proc. Natl Acad. Sci. U. S. A.* **113**, E7856–E7865. DOI:10.1073/pnas.161252411.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. and Yau, C. (2021). Bayesian statistics and modelling, *Nat. Rev. Methods Primers* **1**, 16. DOI:10.1038/s43586-021-00017-2.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N. and Morey, R. D. (2018). Bayesian
Appendix: Statistical Model

A.1. Catch Trials

We begin by considering the participant responses to the catch trials. These trials, of which there were 20 per participant, were interspersed during the experimental session with the trials for the primary task. They were constructed to be sufficiently easy that any incorrect response would be indicative of a momentary lapse in attention or a more extended lack of diligence or misunderstanding of the task requirements. We assume that each participant has a ‘lapse rate’ ($\lambda$) that describes their propensity to lapse on a given trial — which thus affects their responses on the catch trials. We represented the data from the catch trials as the number of incorrect responses per participant ($y^w$), which we assume can be modelled as draws from a binomial distribution. The binomial distribution describes the expected number of ‘successes’ in a set of trials with binary outcomes, given a probability of ‘success’ on each trial. Under the assumption that participants respond randomly during a lapse (i.e., that they respond ‘match’ and ‘nonmatch’ with equal probability), we would expect that they would respond incorrectly on half of the trials in which they lapse — this is because there are two response alternatives in a catch trial and hence they have an equal chance of responding correctly or incorrectly when responding randomly. We can thus model the participant responses on the catch trials as:

$$y^w_i \sim \text{Binomial}(20, \frac{\lambda_i}{2})$$  \hspace{1cm} (1)

where $i$ indexes the participant, 20 is the number of catch trials, and $\lambda_i/2$ is the probability of incorrectly responding on a catch trial for a given participant.

We now consider the model for the participant lapse rates ($\lambda$). Given that this model must respect the constraint that a lapse rate cannot be less than zero or greater than one, we model a transformation of the lapse rate that converts it from the $(0, 1)$ interval to the $(-\infty, +\infty)$ interval rather than modelling the lapse rate directly. We use the transformation known as the logit, which is defined in its general form as:

$$\text{logit}(p) = \log\left(\frac{p}{1 - p}\right)$$  \hspace{1cm} (2)
Given that \( p/(1 - p) \) is the odds, the units of the logit scale are referred to as log-odds. The logit scale can be converted back to the original scale \textit{via} the inverse logit transformation, which is defined in its general form as:

\[
\text{logit}^{-1}(a) = \frac{1}{1 + \exp(-a)}
\]  

(3)

We model the lapse rates as:

\[
\lambda_i = \text{logit}^{-1}\left(\beta_0^w + \alpha_i^{pw}\right)
\]

(4)

This is a standard mixed effects model in which \( \beta_0^w \) describes the average logit lapse rate and \( \alpha_i^{pw} \) is a random effect that allows the logit lapse rate to vary across participants — with both parameters being unknown and are to be estimated from the data.

A.2. Presence Responses (IPQ)

Each participant completed the IPQ, in which they responded to a series of prompts relating to their experiences during their experimental session. We summarised each participant’s responses into a ‘presence’ score \( (y_i)^q \), which could range between 0 and 60. We assume that these presence scores can be modelled \textit{via} a binomial distribution, with the parameter \( \zeta \) describing the presence experienced by the participants as a proportion scale:

\[
y_i^q \sim \text{Binomial}(60, \zeta_i)
\]

(5)

Given that this proportion parameter is bounded to the \((0, 1)\) interval, we model the logit of the proportion rather than the proportion directly (see the above description of the catch trial data for background on this approach):

\[
\zeta_i = \text{logit}^{-1}\left(\beta_0^q + \alpha_i^{pq} + \beta_1^q x_i^e\right)
\]

(6)

which is a mixed-effects model with a between-subjects factor (delivery method). The parameter \( \beta_0^q \) describes the average logit presence across participants, and the \( \alpha_i^{pq} \) parameter is a random effect that allows the logit presence to vary across participants. The \( x_i^e \) term is a known categorical predictor; it has a value of \( \approx -0.58 \) for participants in the HMD delivery method condition and \( \approx +0.42 \) for participants in the web delivery method condition. The asymmetry in the magnitude of the predictor values is due a weighted effect coding approach that incorporates the unequal number of participants in the two experiments (44 in HMD and 60 in web). The associated coefficient, \( \beta_1^q \), thus represents the difference (on the logit scale) between the average presence for participants in the web and HMD delivery method conditions.

A.3. Congruence Task Trials

The primary task in the experiment was to judge the perceived congruence between the auditory and visual stimulation on a given trial. There were 100
of such trials per participant, formed from the pairwise combination of 10 auditory locations and 10 visual locations, where each trial produced a binary outcome where the audiovisual stimulus was (coded as 1) or was not (coded as 0) judged to be congruent (a match). With 104 participants included in the analysis, we represent this data as a three-dimensional array \((y)\) that is indexed by the coordinates \(i\) (participants; \(i \in \{1, \ldots, 104\}\)), \(j\) (auditory location; \(j \in \{1, \ldots, 10\}\)), and \(k\) (visual location; \(k \in \{1, \ldots, 10\}\)).

Given that each trial produces a binary outcome and our assumption that the response on each trial is independent of the responses to the other trials, we model the responses via a Bernoulli distribution in which the probability of a ‘success’ (i.e., a 1) is given by the parameter \(p\):

\[
y_{ijk} \sim \text{Bernoulli}(p_{ijk})
\]  

The probability of a match response on a given trial \((p_{ijk})\) is affected by both the details of a given congruence task trial and the propensity for a participant to respond independently of the task requirements (i.e., to lapse). We thus use the estimate of each participant’s lapse rate \((\lambda_i)\) to shift and scale the probability of a match response that is based on the congruence task requirements \((p^t)\):

\[
p_{ijk} = (\lambda_i/2) + (1 - \lambda_i)p^t_{ijk}
\]

These lapse rate components allow match and nonmatch responses to still be expected even when task-related considerations make them highly unlikely. For example, even when we might be almost certain of a nonmatch response based on the characteristics of a particular trial (that is, \(p^t_{ijk} \approx 0\)), we could still accommodate the occurrence of a match response from a participant lapse because the probability of a match response \((p_{ijk})\) is then \(\approx \lambda_i/2\).

Because the probability of a match response based on congruence task requirements \((p^t)\) is bounded to the \((0, 1)\) interval, we model the logit of the probability rather than the probability directly (see the above description of the catch trial data for background on this approach). We will construct a mixed-effects model for \(p^t\) that has a within-subjects factor (congruence), a between-subjects factor (delivery method), and random effects for participants, auditory locations, and visual locations. Because this model involves a large number of terms, we begin our description with the simplest model and explain each elaboration until we reach the final model that was used in the analysis — we hope that this aids in the interpretability of the model.

We begin with the simple model:

\[
p^t_{ijk} = \logit^{-1}(\beta_0)
\]

Here, the parameter \(\beta_0\) describes the average log-odds of a match response across all participants, auditory locations, and visual locations.
To accommodate our key question of interest, we expand the model to capture differences between congruent trials and incongruent trials:

\[ p_{ijk}^l = \logit^{-1}(\beta_0 + \beta_1 x^c_{jk}) \]  

(10)

Here, \( x^c \) is a categorical predictor that has a value of +0.9 on congruent trials (where \( j = k \)) and a value of −0.1 on incongruent trials (where \( j \neq k \)); the asymmetry in the magnitude of the contrast coefficients is due a weighted effect coding approach (Te Grotenhuis et al., 2017) that incorporates the differing number of congruent (10 per participant) and incongruent (90 per participant) trials. The associated coefficient, \( \beta_1 \), thus represents the change in the average log-odds of a match response for congruent trials relative to incongruent trials.

Because our study involves two delivery methods (HMD and web), we would also like to be able to measure the influence of this between-subjects factor on participant responses. We first expand the model to:

\[ p_{ijk}^l = \logit^{-1}(\beta_0 + \beta_1 x^c_{jk} + \beta_2 x^e_i) \]  

(11)

Here, \( x^e \) is a categorical predictor that has a value of \( \approx -0.58 \) for participants in the HMD delivery method condition and \( \approx +0.42 \) for participants in the web delivery method condition; similar to \( x^c \), the asymmetry in the magnitude of the predictor values is due to the unequal number of participants in the two delivery methods (44 in HMD and 60 in web). The associated coefficient, \( \beta_2 \), represents the change in the average log-odds of a match response between the two delivery methods. We then further expand the model to:

\[ p_{ijk}^l = \logit^{-1}(\beta_0 + \beta_1 x^c_{jk} + \beta_2 x^e_i + \beta_3 x^c x^e_{ijk}) \]  

(12)

The additional predictor \( (x^c x^e) \) is formed from the multiplication of the congruence \( (x^c) \) and delivery method \( (x^e) \) predictors. The associated coefficient, \( \beta_3 \), represents the difference between delivery methods in the difference in the average log-odds of a match response on congruent and incongruent trials.

The model described thus far assumes that the probability of a match response based on the congruence task requirements is identical for each participant in the two between-subject conditions. To allow for variation across participants, we introduce participant random effect terms into the model:

\[ p_{ijk}^l = \logit^{-1}(\beta_0 + \alpha_p^p + (\beta_1 + \alpha_p^{p x c}) x^c_{jk} + \beta_2 x^e_i + \beta_3 x^c x^e_{ijk}) \]  

(13)

Here, the additional terms \( \alpha_p^p \) and \( \alpha_p^{p x c} \) allow participants to vary in their overall propensity to respond ‘match’ (\( \alpha_p^p \)) and in the change in the average log-odds of a match response for congruent relative to incongruent trials (\( \alpha_p^{p x c} \)). This model assumes that neither the specific auditory stimulus that
is delivered nor the specific visual stimulus that is presented on a particular trial affect the probability of a match response based on the congruence task requirements. We thus include additional random effects (Judd et al., 2017) to account for such stimulus-related variation (often referred to as 'item effects'; e.g., Rouder and Lu, 2005).

\[
p_{ijk}^t = \logit^{-1}\left( \left( \beta_0 + \alpha_i^p + \alpha_j^a + \alpha_k^v \right) + \left( \beta_1 + \alpha_i^{p \times c} \right) x_{jk}^c + \beta_2 x_i^e + \beta_3 x_{ijk}^{c \times e} \right) \tag{14}
\]

Here, the additional terms allow the overall log-odds of a match response to vary with different auditory locations across all visual locations (\(\alpha^a\)) and with different visual locations across all auditory locations (\(\alpha^v\)). In our final elaboration of the model, we allow for these item effects to vary across participants:

\[
p_{ijk}^t = \logit^{-1}\left( \left( \beta_0 + \alpha_i^p + \alpha_j^a + \alpha_k^v + \alpha_{ij}^{p \times a} + \alpha_{ik}^{p \times v} \right) + \left( \beta_1 + \alpha_i^{p \times c} \right) x_{jk}^c + \beta_2 x_i^e + \beta_3 x_{ijk}^{c \times e} \right) \tag{15}
\]

### A.4. Statistical Approach

We used a Bayesian framework (Lee, 2018; van de Schoot et al., 2021; Wagenmakers et al., 2018) to determine the posterior probability distribution of the statistical model parameters given our observed data. As our primary goal is to estimate the values of the parameters, rather than to perform formal hypothesis testing or model comparison (Calin-Jageman and Cumming, 2019), reporting summaries of the posterior distributions is our primary method of communicating the outcomes of the study (Kruschke and Liddell, 2018). We use the median as the measure of centrality and credible intervals (equal-tailed intervals, calculated using quantiles; Makowski et al., 2019) as the measure of uncertainty when summarising the posterior distributions.

The Bayesian framework requires the specification of prior distributions for the model parameters. We adopted the strategy of providing weakly informative priors, where the informative aspect is on the approximate scale of the parameter values. The priors for the fixed effects \(\beta_{[0,\ldots,3]}\) were each given by a normal distribution with a mean of 0 and a standard deviation of 1. The prior for \(\beta_0^w\) (the intercept of the logit of the lapse rate) was given by a normal distribution with a mean of logit(0.125) and a standard deviation of 0.75. The prior for \(\beta_0^q\) (the intercept of the logit of the presence proportion) was given by a normal distribution with a mean of 0 and a standard deviation of 1.5 and the prior for \(\beta_1^q\) was given by a normal distribution with a mean of 0 and a standard deviation of 1. The participant random effects were assumed to be drawn from a multivariate normal distribution:

\[
\begin{bmatrix}
\alpha_{i}^{pw} \\
\alpha_{i}^{pq} \\
\alpha_{i}^{p} \\
\alpha_{i}^{p \times c}
\end{bmatrix}
\sim \mathcal{N}_4\left(\begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix}, \Sigma^p\right)
\tag{16}
\]
where $\Sigma^p$ is a $4 \times 4$ covariance matrix. This structure allows for correlations to be observed between the participant random effects. The participant random effects covariance matrix ($\Sigma^p$) was constructed via a set of half-normal distributions for the standard deviations and an LKJ distribution (Lewandowski et al., 2009) for the correlations. The standard deviation for the half-normal priors was 1 for the participant intercept and congruence random effects, 0.5 for the participant lapse rate random effect, and 1 for the participant presence random effect. The shape parameter of the LKJ distribution ($\eta$) was given a value of 2, which applies a weak prior against strong absolute correlations. The item effects and their interactions with participants were each assumed to be drawn from zero-centred normal distributions, where the standard deviation parameter was given a prior of a half-normal distribution with a standard deviation of 1.