Journal of Experimental Psychology: General

Using Hearing and Vision for Motion Prediction, Motion Perception, and Localization

Yichen Yuan, Nathan Van der Stoep, and Surya Gayet

Online First Publication, January 27, 2025. https://dx.doi.org/10.1037/xge0001725

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001725

Using Hearing and Vision for Motion Prediction, Motion Perception, and Localization

Yichen Yuan, Nathan Van der Stoep, and Surya Gayet
Department of Experimental Psychology, Helmholtz Institute, Utrecht University

Predicting the location of moving objects in noisy environments is essential to everyday behavior, like when participating in traffic. Although many objects provide multisensory information, it remains unknown how humans use multisensory information to localize moving objects, and how this depends on expected sensory interference (e.g., occlusion). In four experiments, we systematically investigated localization for auditory, visual, and audiovisual targets (AV). Performance for audiovisual targets was compared to performance predicted by maximum likelihood estimation (MLE). In Experiment 1A, moving targets were occluded by an audiovisual occluder, and their final locations had to be inferred from target speed and occlusion duration. Participants relied exclusively on the visual component of the audiovisual target, even though the auditory component demonstrably provided useful location information when presented in isolation. In contrast, when a visual-only occluder was used in Experiment 1B, participants relied exclusively on the auditory component of the audiovisual target, even though the visual component demonstrably provided useful location information when presented in isolation. In Experiment 2, although localization estimates were in line with MLE predictions, no multisensory precision benefits were found when participants localized moving audiovisual target. In Experiment 3, a substantial multisensory benefit was found when participants localized static audiovisual target, showing near-MLE integration. In sum, observers use both hearing and vision when localizing static objects, but use only unisensory input when localizing moving objects and predicting motion under occlusion. Moreover, observers can flexibly prioritize one sense over the other, in anticipation of modality-specific interference.

Public Significance Statement

When crossing a busy street, with cars honking and engines rumbling, we must continuously keep track of what we see and hear. At the same time, our vision and hearing are often interrupted; our view of an incoming car might be obstructed by a bus, and the sound of construction works might obfuscate the sound of the engine. Surprisingly, we found that observers never combine vision and hearing to track occluded moving objects, always relying almost exclusively on a single sense. Instead, observers flexibly switch between senses in anticipation of sensory interruptions; using hearing or vision depending on (a) which sense will be interrupted and on (b) which sense conveys the most reliable information. Our study provides insights into how our senses work together to help us interact with objects in the world.

Keywords: multisensory perception, maximum likelihood estimation, audiovisual integration, motion perception, occlusion

Supplemental materials: https://doi.org/10.1037/xge0001725.supp

Aysecan Boduroglu served as action editor.

Yichen Yuan https://orcid.org/0000-0002-1000-7287
Nathan Van der Stoep https://orcid.org/0000-0002-0412-2078
Surya Gayet https://orcid.org/0000-0001-9728-1272

All data and all experiment, analysis, and visualization codes are available via the Open Science Framework at https://osf.io/54vms/. A preprint of this manuscript has been made publicly available and can be found via https://osf.io/preprints/socarxiv/uvzdh. Some of the data and ideas of this manuscript have been presented as a poster at International Multisensory Research Forum Conference (Brussels, Belgium) on June 28, 2023, and at System Vision Science Summer School (Tubingen, Germany) from August 14 to 24, 2023. All of the data and ideas of this article have been presented as a poster at the Dutch Society for brain and cognition Winter Conference (Egmond aan Zee, Netherlands) on December 14, 2013; at Vision Sciences Society Conference (St Pete Beach, United States of America) on May 18, 2024; and at Visual Neuroscience Summer School (Schloss Rauischholzhausen, Germany) from September 1 to 13, 2024. The study protocols were approved by the faculty ethics committee of Utrecht University (Number 21-0397). All participants signed informed consent for their participation. The authors have no conflicts of interest to disclose.

This work was supported by the China Scholarship Council (Grant 202206380011 to Yichen Yuan). The authors thank Eli Brenner for suggesting the analyses of shared variance.

Yichen Yuan played a lead role in data curation, formal analysis, funding acquisition, investigation, software, validation, visualization, and writing–original draft and an equal role in conceptualization, methodology, resources, and writing–review and editing. Nathan Van der Stoep played an equal role in

continued

Predicting the location of moving objects is essential in day-to-day behavior, for instance, when crossing a street or driving a car. Imagine that you are planning to cross a street, for example. When looking for incoming cars on either side of the road, your view of an approaching car might be temporally obstructed by a van parked on the near side of the road. In this case, vision and audition convey complementary information about the moving car, jointly supporting your prediction of the current location of the incoming car. Thus, hearing and vision both inform our behavior (crossing the street, or not).

Although the objects that we interact with in real life convey information through multiple senses, most previous studies on motion prediction focused solely on the visual modality (Battaglini & Mioni, 2019; Battaglini et al., 2018; Baurès et al., 2018, 2021; Erlikhman & Caplovitz, 2017; Flavell et al., 2018; Lugtigheid & Welchman, 2011). These studies have taught us how speed, motion duration, and attention affect visual motion prediction performance (and its neural mechanisms). The fact that real-world events convey information through multiple modalities is especially relevant if we consider the issue of sensory interruption. Sensory interruptions are unlikely to equally affect all senses and, in many cases, might affect one modality in particular. The van parked in front of us might block the view of the moving car but not its sound. In contrast, construction work or a conversation might make it harder to hear the car but does not obstruct our view of the car. Accordingly, it is reasonable to hypothesize that it would be beneficial to use both auditory and visual information when tracking objects, especially given the plausibility of sensory interruptions.

One paradigm used to study motion prediction in the laboratory is the prediction-motion (PM) task (e.g., DeLucia et al., 2016; Menceloglu & Song, 2023). In a typical PM task, a visual target moves toward an occluder that blocks a specific region on the screen. After the target is occluded, participants are required to predict the motion of the target and indicate when the moving target will reach the end of the occluder. Then the time interval between target disappearing and perceived target reappearing at the end of the occluder is measured as time-to-contact (Tresilian, 1995). This task requires participants not only to track the target in real-time but also to predict the future position of the target, based on a variety of motion cues, such as the speed, time, and space (for a recent review about motion prediction under occlusion, see Battaglini & Ghiani, 2021).

Despite being scarce, studies on audiovisual motion prediction yield inconsistent findings. While some studies found auditory information was not helpful when predicting the location of audiovisual targets (AV), others reported the opposite. Hofbauer et al. (2004), for instance, adopted a PM task and found lower localization variability for visual and audiovisual targets compared to auditory targets (A), while no significant difference in response variability was found between the unisensory visual and audiovisual targets. This shows that the additional auditory information did not increase the motion prediction performance in audiovisual condition compared to unisensory visual condition and participants relied primarily on visual information for motion prediction. Similar conclusions were drawn using a stereoscopic three-dimensional simulation (Dittrich & Noesselt, 2018) and realistic movie clips (Schiff & Oldak, 1990). Different results were observed in a study by Prime and Harris (2010), who modified the PM task by only occluding the visual component of the stimulus, but not the auditory component (i.e., using a visual-only occluder). They also introduced a spatial displacement between the auditory and visual components of the audiovisual stimuli (sound-ahead, sound-behind, sound-match). Participants were required to report the future position of the visual target, assuming it continued moving along with the unoccluded sound. The results showed that responses to visual stimuli were biased in the direction of the displaced auditory stimulus, suggesting that both auditory and visual information were used for localization prediction.

Thus, it remains unclear under what circumstances using multisensory information for motion prediction is more beneficial than using unisensory information. Different materials and methods might be one explanation for the inconsistent results (Dittrich & Noesselt, 2018). The stimuli used in PM tasks varied from simple flashes with sound clicks (Hofbauer et al., 2004), geometric stimuli with beeps (Chotsrisuparat et al., 2018; DeLucia et al., 2016; Prime & Harris, 2010), faces with voices (Lu et al., 2023), to complex three-dimensional scenes (Dittrich & Noesselt, 2018; Keshavarz et al., 2017; Wessels et al., 2023) and video clips (Schiff & Oldak, 1990). Experimental setups also varied from light-emitting diodes with speakers (Hofbauer et al., 2004), screens with speakers (Prime & Harris, 2010), projectors (Schiff & Oldak, 1990), to virtual reality (Keshavarz et al., 2017; Wessels et al., 2023), not to mention the differences in detailed parameters used in these experiments, such as target speed, tracking duration, and the occurrence and type of occlusion. The large variability of materials and methods used in different studies makes it difficult to compare the results across motion-tracking experiments and to compare these results with those of localization tasks, in which multisensory benefits have been reliably established (e.g., Freeman et al., 2018; Meijer et al., 2019). To understand how observers use multisensory input to localize moving objects, it seems imperative to systematically compare multisensory performance to unisensory performance—across localization tasks, motion-tracking tasks, and motion prediction tasks—while keeping the stimulus materials, methods, report task, and experimental setup constant.

The main goal of this study was to investigate under which circumstances information from multiple senses is used to predict the motion of audiovisual objects. We ask whether using both auditory and visual information provides a performance benefit compared to using unisensory information alone if multisensory inputs are available and useful. We consider that the extent to which observers rely on different senses can be context-dependent. Specifically, we investigated whether the presence and type of occlusion influence whether multisensory information is used to guide behavior. On the one hand, multisensory

conceptualization, methodology, project administration, resources, supervision, writing–original draft, and writing–review and editing. Surya Gayet played an equal role in conceptualization, methodology, project administration, resources, supervision, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Yichen Yuan, Department of Experimental Psychology, Helmholtz Institute, Utrecht University, Heidelberglaan 1, 3584 CS, Utrecht, The Netherlands. Email: y.yuan@uu.nl

objects inherently convey more information than unisensory objects, such that it would be more beneficial to use both modalities to guide behavior; on the other hand, simultaneously encoding and maintaining information in multiple modalities might require more resources, especially under occlusion, which might cause observers to rely on a single modality instead, depending on which senses are occluded/disrupted.

To systematically address these questions, we conducted a series of four experiments with the same experimental setup (same equipment and preparation before conducting the experiments, as well as the same logic of Matlab code for generating the condition and trial matrix) and stimuli, only differing in the presence or absence of an occluder, in the type of occluder (audiovisual or visual-only), and in the dynamics of the target (moving or static). We compared participants' localization performance (error and precision) for multisensory targets to that of the unisensory targets and maximum likelihood estimations (MLEs) of performance in the multisensory condition based on unisensory performance (Alais & Burr, 2004b). This MLE model assumes that if multisensory inputs are integrated, two unisensory inputs are weighted in a maximum likelihood fashion according to their reliability (variance), leading to a weighted average in localization response and a potential increase in precision for multisensory stimuli. Thus, by comparing the participant's performance in the audiovisual condition to the MLE prediction, we could test whether auditory and visual information are combined in an MLE fashion to benefit performance.

In Experiments 1 and 2, a moving target (either auditory, visual, or audiovisual) would appear on the left of the screen and move toward the right. Next, an audiovisual (Experiment 1A) or visual-only (Experiment 1B) occluder would unpredictably appear and disappear, and participants had to indicate where the target would have been when the occluder disappeared in horizontal space. In Experiment 2, there was no occluder, so participants were required to simply report where the moving target disappeared from the screen. In Experiment 3, the target did not move, but only briefly appeared as a static stimulus at the same endpoints as in Experiments 1 and 2. Participants had to report their location. In all four experiments, we measured the mean horizontal localization error and precision.

To preface the results, we show that participants use both auditory and visual inputs when localizing static targets, but exclusively use unisensory information when localizing moving targets and predicting motion under occlusion. Which sense observers rely on depends on the modality of the occluder (i.e., which sensory input will be interrupted) and the reliability of the input sensory information.

Experiments 1A (Audiovisual Occluder) and 1B (Visual-Only Occluder)

In the first two experiments we investigated how participants use auditory and visual information when tracking audiovisual objects under occlusion. In Experiment 1A, both the auditory and the visual components of the target object were occluded. We considered two possible outcomes: (1) Participants might use both auditory and visual components to predict the upcoming target location because they jointly provide more information than any unisensory component alone (akin to multisensory benefits observed in typical localization tasks) and (2) alternatively, participants would rely exclusively on a single unisensory component to minimize attentional load during occlusion.

Experiment 1B was identical to Experiment 1A except that a visual-only occluder was used. In real-world situations, sensory occlusion is unlikely to equally affect all senses, and thus often affects one modality more than the other (e.g., the van parked in front of us might block the view of the incoming cars, but not its sound). We hypothesized that, in this situation, observers rely predominantly on the nonoccluded modality (the auditory component). This would show that observers can prioritize one sensory input over the other to account for expected interference while tracking moving objects.

Method

Transparency and Openness

All data and all experiment, analysis, and visualization codes are available via the Open Science Framework at https://osf.io/54vms/. The research was not preregistered. However, we used the same stimulus parameters, sample size, and analysis pipeline in all four sequentially conducted experiments, thus minimizing the chance that the current findings are dependent on accidental/opportunistic tuning of analysis settings.

Participants

Based on a sample size estimation performed in G*Power software (Faul et al., 2009), 34 participants are required for 80% power to observe a medium effect size (Cohen's $d = .5$) with a paired-samples t test ($\alpha = 0.05$) in a within-participants design. To include 34 participants in both experiments, a total of 78 participants were tested; 40 for Experiment 1A, and 38 for Experiment 1B. We defined subject-level inclusion criteria, aimed at asserting that all participants could perform the motion prediction task better than chance in all three conditions (auditory, visual, and audiovisual targets). Task performance was quantified as the correlation between the true target endpoints and the target endpoints reported by the participant. A significant positive correlation indicates that participants report a more rightward location when the target endpoint moved further rightward, showing that participants understood the instructions and were able to perform the task. Chance-level performance was established for each participant, by conducting a permutation test. This permutation test entailed shuffling the participant responses across trials, thereby decoupling the responded target location from the actual target location on that trial. Repeating this shuffling procedure 1,000 times, allowed us to construct a null distribution of correlations r (H0). This null distribution indicates the range of correlations we may observe when participants respond randomly while preserving the observed response variance and potential response biases of each specific participant. The null distribution then allows us to test how (un)likely the observed correlation is, as compared to the situation in which the participant responded randomly. Considering the standard α level of 5% (one-tailed p value), the observed correlation should lie above the upper boundary of the 95% percentile, indicating that the probability of obtaining a correlation at least as positive as that of the

observed data is less than 5% if this participant responded randomly. Thus, we excluded participants whose correlations lay below the upper boundary of the 95% percentile in any of the conditions (Deutsch et al., 2023; Holt & Sullivan, 2023). Based on this inclusion criterion, six participants were excluded in Experiment 1A and four participants were excluded in Experiment 1B. All participants reported normal or corrected-to-normal vision and normal hearing.

Data of 34 participants were included in the final analysis of each experiment (Experiment 1A: 30 participants reported their gender as female, four as male; $M_{age} = 24.50$ years, $SD = 4.64$, range = 20–44 years; Experiment 1B: 27 participants reported their gender as female, seven as male; $M_{age} = 24.47$ years, $SD = 3.89$, range = 19–40 years). Participants were asked to type their gender in a box, with the suggested options "Male" or "Female" listed above. Participants were asked to type in their age in another box. All participants signed informed consent and received money or course credits for their participation. The study protocols were approved by the faculty ethics committee of Utrecht University (Number 21-0397).

Apparatus and Stimuli

The experiments were conducted in a dimly lit lab and controlled using Matlab (2021b). Participants were seated with their head positioned in a chin rest, to keep their viewing distance at a fixed 40 cm in front of a 23-in. monitor (1,920 × 1,080; 60 Hz). The auditory stimuli were presented with two loudspeakers (Logitech Z150), placed along the vertical edges of the screen, with the cones of the speakers placed along the horizontal meridian of the screen. The setup was placed in the middle against the front wall of the room.

The auditory stimuli consisted of white noise as the target, 66 dB (A) on average, and pink noise as the occluder, 63 dB (A), of variable duration depending on the specific trial, generated in Matlab (2021b), sampled at 44.1 kHz, and quantized to 16 bits. Participants were able to adjust the volume to a subjectively comfortable level at the start of the tasks. The visual moving target stimuli consisted of uniform white noise (each pixel was randomly assigned to a value between black and white), with a 2D Gaussian aperture ($SD = 600$, amplitude = 0.05) on top of the white noise. Thus, the contrast of the white noise is highest at the center and decreases as it goes outward. Note that we heavily degraded the visual motion stimulus, to make the performance between the visual-only and the auditory-only conditions more similar. The white noise constituting the moving target stimulus was changed every five frames to ensure that participants tracked the visual object as a whole, rather than individual pixels. The visual occluder consisted of full-screen uniform white noise presented statically during the occlusion time.

Procedure

Both Experiments 1A and 1B included two tasks, a sound calibration task and a motion prediction task. All participants completed the calibration task before the motion prediction task. Each task was preceded by a practice session.

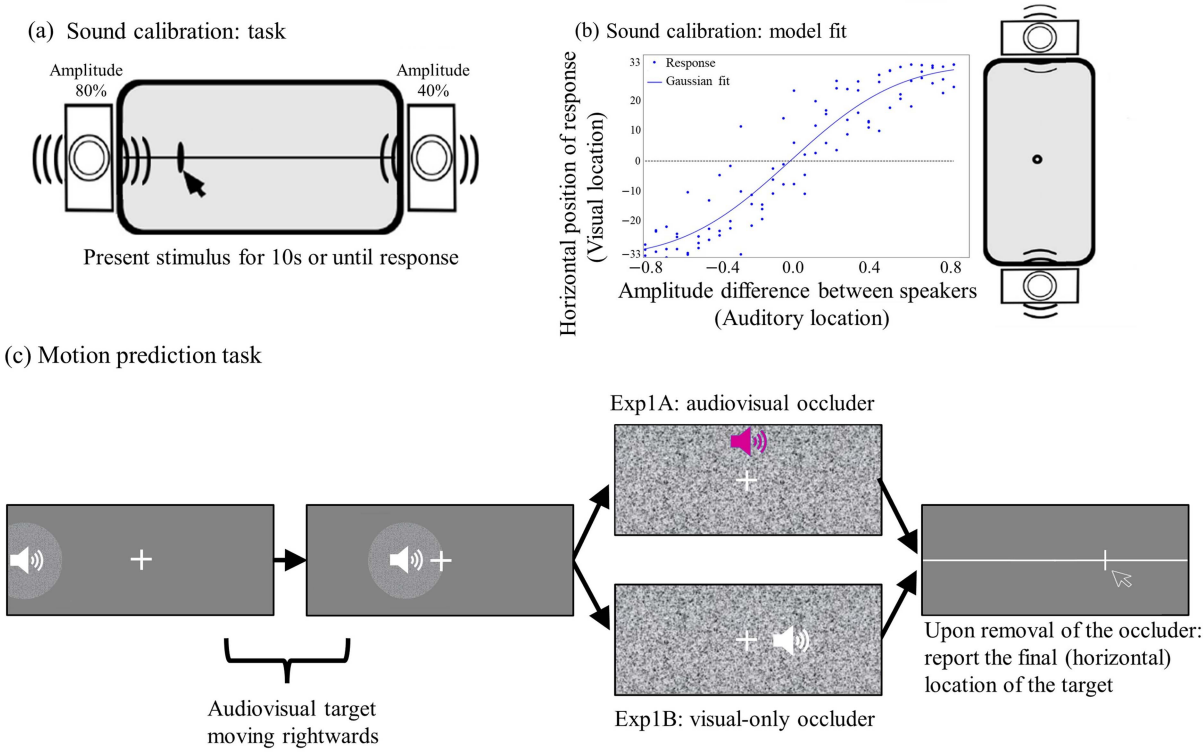
Sound Calibration Task. The goal of the calibration task was to relate the amplitude difference between speakers to the perceived horizontal location of the sound measured by participants' response. This way, we could map the sound balance levels of the left and right speakers to screen coordinates and present the auditory and visual stimulus at the same location in the motion prediction task for each individual participant. On each trial, auditory white noise, 66 dB (A), was played for 10 s or until participants provided a response. Participants reported where the sound originated from, by clicking on a horizontal line on the screen using a computer mouse (Figure 1a). We linearly changed the amplitude difference between the two speakers in 30 equal steps (amplitude difference from −0.8, i.e., amplitude of the left/right speaker equals 100%/20%, to 0.8, i.e., amplitude of the left/right speaker equals 20%/100%; Schut et al., 2018). Each step was presented three times, and trials were presented in random order. Each participant completed 90 trials in the calibration task.

The data were used to fit a cumulative Gaussian function and a linear function to the relation between amplitude difference and perceived sound location. The better model was chosen based on a cross-validation approach (Browne, 2000). Specifically, to have a substantial number of trials for the fitting procedure, 80 out of 90 trials were randomly selected and used to fit a cumulative Gaussian function and a linear function, separately. Then the fit (R^2) for the 10 left-out trials was calculated for both the cumulative Gaussian and linear fit. This random splitting of 80 and 10 trials for the cross-validation procedure was repeated 1,000 times. After 1,000 iterations, the model was chosen that yielded the higher R^2 on most iterations. The results of one example participant are shown in Figure 1b. Their responses were best captured by a cumulative Gaussian fit.

Motion Prediction Task. The goal of the motion prediction task was to investigate how participants use auditory and visual information when tracking audiovisual objects under audiovisual and visual-only occlusion. In Experiment 1A, each trial an auditory, visual, or audiovisual target moved horizontally at a constant speed (pseudorandomly chosen from 6, 9, or 12 degrees of visual angle [dva] per second) from the left edge of the screen to the right. At a varying timepoint (1 to 3 s, stimulus presentation times), the moving target was occluded by an audiovisual occluder. The audiovisual occluder was a full-screen visual white noise mask, accompanied by auditory pink noise, fully occluding the moving targets in both modalities. This approach is similar to Battaglini's visual extrapolation work (Battaglini et al., 2015, 2016, 2017). The visual part of the occluder was presented statically, always in front of the moving target. Importantly, the target was still implied to move behind the occluder but the target stimuli were absent. After a varying delay (1–3 s, stimulus prediction times), the occluder disappeared. Then a horizontal line was presented spanning from the left to right edge of the screen, and the cursor was depicted as a small vertical line appearing at the center of the horizontal midline. Participants moved the mouse horizontally to indicate where (on the horizontal line) the target should have been when the occluder disappeared. Feedback on response accuracy was not provided to avoid response strategies. After a response, the next trial began. The trial scheme is shown in Figure 1c (for the detailed participant instructions in Experiment 1A, see Supplemental Materials 1.1). Experiment 1B was identical to Experiment 1A except that the occluder was always visual-only (Figure 1c).

Thus, a one factor (target type: auditory, visual vs. audiovisual) within-participants design was adopted for both experiments. Moreover, the stimulus presentation times during which participants could see or hear the targets and the stimulus prediction times during which the targets were occluded were pseudorandomly chosen from

Figure 1
Procedure of Experiment 1



Note. (a) Schematic depiction of the calibration task. In each trial, white noise sound was played and participants were required to indicate where the sound originated from by clicking on the horizontal line on screen with a computer mouse. (b) An example of a cumulative Gaussian fit of the data from the sound calibration task. We plot the participant's reported sound locations (*y*-axis) as a function of the difference in sound amplitude between the left and right speakers (i.e., interaural level difference). The dashed horizontal line indicates the center of the screen (fixation). The blue dots are responses from individual trials, and the blue line shows the cumulative Gaussian fit of the data. (c) A schematic depiction of a trial in the motion prediction task in Experiments 1A and 1B (in the audiovisual target condition). Each trial an auditory, visual, or audiovisual target moved horizontally at a constant speed from left to right. At a varying timepoint, the moving target was occluded by an audiovisual (upper, Experiment 1A) or a visual-only occluder (lower, Experiment 1B). After an unpredictable delay, the occluder and the occluded stimuli disappeared and participants had to indicate where the target should have been when the occluder disappeared. Note that for illustration purposes, the visual stimuli are quite easy to see in the figure. In the real experiment, the contrast of the visual target is heavily reduced (following a Gaussian profile) and its effective diameter is larger than the vertical length of the screen. Exp = Experiment. The task showed in panels (a) and (b) of this figure is adapted from Schut et al. (2018). See the online article for the color version of this figure.

1 to 3 s, such that they were decorrelated. The speeds of the target objects were also varied so that participants could not know the location of the reappearing target solely based on the amount of time that elapsed (see Supplemental Materials 1.2 for more detailed information). Overall, each experiment lasted for 60 min. Trial type and speed were pseudorandomized. Both experiments consisted of nine blocks with 30 trials each, counterbalanced for the modality, speed, stimulus presentation times, and stimulus prediction times.

Data Analysis

Data analyses followed the same procedure in all experiments. First, to establish whether unisensory auditory and visual trials both contained usable localization information at the group level, we calculated the correlation between the true target endpoints and the target endpoints reported by the participant (akin to the analysis used to determine chance level performance in individual participants). Importantly, we included all participants in this analysis ($N = 40$ for Experiment 1A, $N = 38$ for Experiment 1B), following the same bootstrapping procedure used to define the chance level for individual participants. By including the entire sample, we aimed to test whether the population as a whole (rather than our included sample) could perform the task above the chance level. The group chance-level performance was established by shuffling each individual's responses across trials, computing a correlation for each participant, and averaging the correlation across participants for each permutation. After 1,000 permutations, we obtained a probability distribution of correlations for each condition at group level, given that responses were random (i.e., a null effect). Then, we compared whether the group mean correlations in the observed data were higher than the upper boundary of the 95% percentile in either condition. If that is the case, it would show that unisensory auditory and visual components

both contain usable localization information that observers can use to localize the moving target. We also compute *p* values quantifying how unlikely the observed group mean correlations were relative to the generated null distribution.

Second, localization error and precision were calculated separately for each participant in each experimental condition (auditory, visual and audiovisual targets). Localization error was calculated by subtracting the end position of targets (i.e., the position that the target would have been when the occluder was removed) from the response position of participants in dva. Positive dva indicated overestimation of the target location (rightward bias), while negative dva indicated underestimation of the target location (leftward bias). Precision was 1 divided by the standard deviation of the localization error (in dva). Larger precision indicated better performance. For all participants, responses within five pixels (0.18 dva) of the initial mouse location were considered as mistakes and thus discarded. Furthermore, for all participants, for each condition, trials with a localization error exceeding ±3 *SD* of the participant's mean localization error in each experimental condition were removed. Repeated measures analyses of variance (ANOVAs) were run separately for mean localization error and precision with the factor target type (auditory, visual vs. audiovisual). When necessary, the Greenhouse–Geisser correction was adopted to correct the degrees of freedom. Bonferroni corrections were used for post hoc pairwise comparisons.

To test whether auditory and visual information were integrated into the audiovisual target condition, we adopted an MLE model (Alais & Burr, 2004b; Ernst & Banks, 2002). This model generates a prediction of the results in the audiovisual condition using unisensory auditory and visual responses of participants, assuming multisensory integration happens. We compared the observed audiovisual responses to model predictions to see whether audiovisual inputs were integrated using MLE. Specifically, this model assumes that two unisensory channels are weighted in a maximum likelihood fashion according to their reliability (Equations 1 and 2). Thus, the more reliable the unisensory information is, the more the unisensory component of the audiovisual stimulus should contribute to the location estimation of the audiovisual stimuli. Accordingly, the model predicts that the mean localization error of the audiovisual stimuli should lie in between that of the unisensory auditory and visual stimuli. Another assumption made by this model is that the combined audiovisual variance should be smaller than the variance in the unisensory auditory and visual conditions (Equation 3). Thus, a significant increase in precision (1/*SD*) should be observed in the audiovisual condition compared to the unisensory conditions given that the sensory inputs are weighted in an MLE fashion. Paired *t* tests were used to test whether the localization error and precision observed in the audiovisual condition significantly differed from the prediction of the MLE model.

$$M_{\mathrm{AV\_prediction}} = \mathrm{w}_A M_A + \mathrm{w}_V M_V, \tag{1}$$

where

$$\mathrm{w}_A = \frac{\sigma_V^2}{\sigma_A^2 + \sigma_V^2}; \mathrm{w}_V = 1 - \mathrm{w}_A, \tag{2}$$

$$\sigma^2_{\mathrm{AV\_prediction}} = \frac{\sigma_A^2 \sigma_V^2}{\sigma_A^2 + \sigma_V^2} \leq \min(\sigma_V^2, \sigma_A^2). \tag{3}$$

To evaluate the extent to which participants used the available auditory and visual information to track audiovisual objects under occlusion, we used two distinct approaches, providing converging evidence. First, we quantified how well the performance in auditory and visual conditions could predict performance in the audiovisual condition, compared to how well performance in the audiovisual condition was predictive of itself. We did this by calculating group-level correlations between conditions in a split-half bootstrapping procedure (for more details, see Supplemental Materials 1.3). Second, we quantified how participants actually weighted the auditory and visual information based on the observed localization responses in the auditory, visual, and audiovisual conditions. We then compared these observed weights with the predicted weights (based on the MLE model), and with simulated weights, which describe the weights that would be observed if one modality did not contribute to behavior at all. Specifically, the predicted weights (based on the MLE model; Equation 2) described how ideal observers would weigh the auditory and visual components of audiovisual stimuli, going solely by the relative variances in the auditory and visual unisensory conditions. Put simply, these predicted weights reflect that more precise unisensory components of a multisensory stimulus should carry more weight than less precise unisensory components when optimally combined. Instead, the observed weights describe how much the auditory and visual components were actually weighed to obtain the observed localization in the audiovisual condition (Equation 4). Put simply, if localization in the audiovisual condition was closer to localization in one unisensory condition (e.g., vision) than the other, the former unisensory component apparently carried more weight. Finally, we also simulated the weights that would be observed if localization was driven by one modality alone (e.g., vision) and the other modality was fully ignored (Equation 5). Note that these simulated weights are not necessarily 0 or 1, because of measurement noise, which then also causes noise in the estimation of variances. To calculate the simulated weights, we first split each condition's data (i.e., data in unisensory auditory, unisensory visual, and audiovisual conditions) into two halves. To simulate—for instance—what the audiovisual responses would look like if they were based on the visual component alone, one half of the trials in the visual condition were used as the "vision only" audiovisual condition (and the other half were simply used as the visual condition). Using Equation 5, we could then calculate the weights of the auditory and visual components by generating the observed responses in the audiovisual condition, if participants relied solely on visual (or auditory) information. Intuitively, if participants relied only on visual information in the audiovisual trials, the simulated ("vision only") weights should be very similar to the observed weights that were calculated from the actual audiovisual trials. For all the weight calculations, weight values exceeding ±3 *SD* of the group mean weight in each experiment were removed. This exclusion led to one observed weight removed in Experiment 1B, one observed weight removed in Experiment 2, and two observed weights and one simulated weight removed in Experiment 3. Furthermore, considering that the weights were within the range of 0 and 1 theoretically (based on the MLE model), we delimitated the group mean weight value between 0 and 1. Individual weights were allowed to be outside of this range, however, to account for measurement noise (and thus for group mean weights of 0 and 1 to be possible).

$$w_{A\_observed} = \frac{M_{AV\_observed} - M_{V\_observed}}{M_{A\_observed} - M_{V\_observed}}; w_{V\_observed} = 1 - w_{A\_observed}, \quad (4)$$

$$w_{A\_simulated} = \frac{M_{V\_half2} - M_{V\_half1}}{M_{A\_half1} - M_{V\_half1}}; w_{V\_simulated} = 1 - w_{A\_simulated}. \quad (5)$$

Results

Usefulness of Localization Information

First, we set out to establish that the auditory and visual components of the moving target stimuli provided useable location information. To this end, we included all participants and tested whether the reported target location correlated positively with the actual (end) position of the target. As expected, in both Experiments 1A and 1B, correlations in auditory-only, visual-only, and audiovisual target conditions were higher than chance level (i.e., the upper boundary of the 95% percentile of the permuted distribution; Experiment 1A: all $r = .29–.40 > .02$; Experiment 1B: all $r = .37–.56 > .03$). In all conditions of Experiment 1, the observed group mean correlations were more than 23 *SDs* higher than the mean of the permuted distribution (the null; all $p < .001$). This establishes that both modalities—when presented in isolation—provided ample information for participants to perform the motion prediction task above chance level in both experiments.

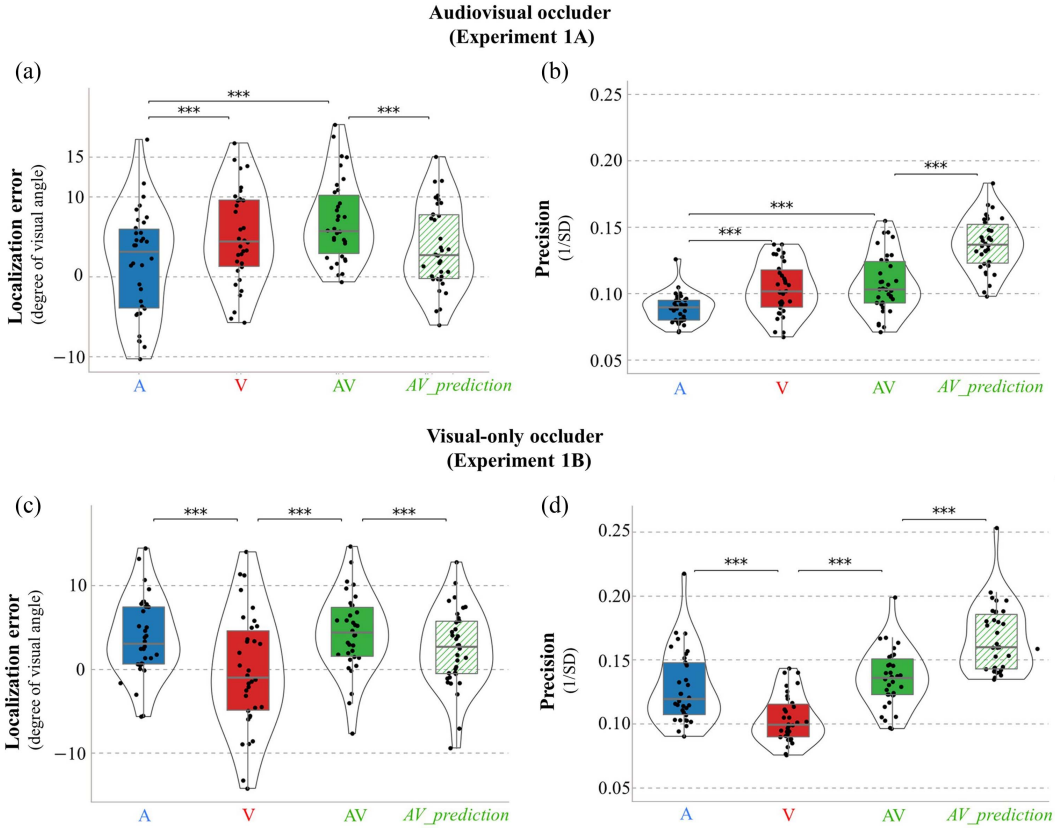
Localization Error and Precision

In Experiment 1A (audiovisual occluder), the repeated measures ANOVA on the localization error revealed a significant main effect of target type, $F(1.26, 41.53) = 21.95$, $p < .001$, $\eta_p^2 = .40$ (Figure 2a). Follow-up analyses showed that mean localization error for audiovisual targets was larger than that for auditory targets, $M = 7.00$ versus 1.72 dva, $SE = .88$ versus 1.13; $t(33) = 6.54$, $p < .001$, $d = .90$, but did not differ from visual targets (V), $M = 7.00$ versus 5.10 dva, $SE = .88$ versus 1.01; $t(33) = 2.36$, $p > .05$.

A repeated measures ANOVA performed on response precision revealed a significant main effect of target type, $F(1.54, 50.94) = 24.30$, $p < .001$, $\eta_p^2 = .42$ (Figure 2b). Follow-up analyses showed that audiovisual targets yielded more precise localization response than auditory targets, $M = .11$ versus .09, $SE = .004$ versus .002; $t(33) = 6.57$, $p < .001$, $d = 1.00$, but did not differ from visual

Figure 2

Results of Experiments 1A and 1B



Note. Panel (a) shows the localization error, and Panel (b) shows the response precision in the auditory (blue), visual (red), and audiovisual (green) conditions, and MLE model predictions (shaded green) in Experiment 1A (with an audiovisual occluder); Panels (c) and (d) depict the same for Experiment 1B (visual-only occluder). MLE = maximum likelihood estimation; A = auditory targets; V = visual targets; AV = audiovisual targets. See the online article for the color version of this figure.

*** $p < .001$.

targets, $M = .11$ versus $.10$, $SE = .004$ versus $.003$; $t(33) = 1.26$, $p > .05$.

An opposite pattern of results was found in Experiment 1B (visual-only occluder). Again, a repeated measures ANOVA performed on the localization error revealed a significant main effect of target type, $F(1.25, 41.26) = 19.02$, $p < .001$, $\eta_p^2 = .37$, see Figure 2c. Follow-up analyses showed that mean localization error for audiovisual targets was larger than that for visual targets, $M = 4.38$ versus $-.19$ dva, $SE = .80$ versus 1.19; $t(33) = 5.66$, $p < .01$, $d = .82$, but did not differ from auditory targets, $M = 4.38$ versus 3.80 dva, $SE = .80$ versus $.81$; $t(33) = .71$, $p > .05$.

A repeated measures ANOVA performed on response precision also revealed a significant main effect of target type, $F(1.57, 51.77) = 26.53$, $p < .001$, $\eta_p^2 = .45$ (Figure 2d). Follow-up analyses showed that audiovisual targets yielded more precise localization responses than visual targets, $M = .14$ versus $.10$, $SE = .004$ versus $.003$; $t(33) = 6.94$, $p < .001$, $d = 1.34$, but did not differ from auditory targets, $M = .14$ versus $.13$, $SE = .004$ versus $.005$; $t(33) = 1.54$, $p > .05$.

In sum, participants did not benefit from multisensory information compared to unisensory information while tracking objects under either audiovisual or visual-only occlusion. Instead, the data suggest that participants seemed to rely solely on the most informative unisensory signal given the upcoming occluder.

Maximum Likelihood Estimation

To test whether auditory and visual information were integrated according to MLE (i.e., weighted based on their relative reliability), we conducted a paired-samples *t* test between the performance in the audiovisual target condition and the MLE predictions that are based on the two unisensory conditions. The results showed that, in both Experiments 1A and 1B, audiovisual targets yielded worse localization error and worse response precision than expected by MLE (all $t > 7.94$, all $p < .001$, all $d > 1.36$; see Figure 2), showing that unisensory information was not combined in an MLE consistent manner (i.e., the auditory and visual components of the audiovisual targets were not weighted according to their relative reliability).

Utilization of Auditory and Visual Information

Considering that audiovisual targets did not provide any behavioral advantage over unisensory targets in terms of localization accuracy and precision, we next asked whether participants' performance fully relied on a single unisensory component, completely ignoring the other. To address this question, we used two different analysis approaches, yielding converging results.

First, we tested whether individual subjects' performance in the audiovisual condition was more similar to itself than to performance in the unisensory conditions (using a split-half bootstrapping procedure). We found that in both Experiments 1A and 1B, localization in the audiovisual target condition was distinguishable from localization in the two unisensory conditions. However, in Experiment 1A, precision in the audiovisual target condition was not distinguishable from localization in the two unisensory conditions. The observation that the audiovisual and visual conditions were distinguishable established that both components of the audiovisual targets affected behavior, despite not being combined in a way that yielded performance benefits (for more details, see Supplemental Materials 1.3).

Second, we compared the observed weights to the weights that would be predicted by MLE and to the (single modality simulated) weights that would be observed if one modality did not contribute to behavior at all. We found that in Experiment 1A, the reliance on visual input in the actual data of the audiovisual condition (observed visual weight: 70%) is more extreme than the MLE predictions (predicted visual weight: 60%); and less extreme than single visual modality simulations (simulated visual weight: 100%). In Experiment 1B, the reliance on auditory input in the actual data of the audiovisual condition (90%) is more extreme than the MLE predictions (60%) and less extreme than single auditory modality simulations (100%). This provides further evidence that both components of the multisensory stimuli contributed to behavioral reports; participants did not fully rely on one sense while completely ignoring the other (for more details, see Supplemental Materials 1.4).

In sum, these results suggest that although multisensory stimuli yielded no performance benefits compared to (i.e., the most reliable) unisensory stimulus, the other (less reliable) unisensory component was not completely discarded.

Interim Discussion

Overall, the results of Experiments 1A and 1B suggest that auditory and visual inputs were not integrated when tracking occluded audiovisual objects. Most importantly, even though unisensory auditory and visual inputs of the same object both contained useable location information, participants seemed to localize occluded audiovisual targets based on unisensory information alone. Specifically, participants relied exclusively on visual information under audiovisual occlusion, while they relied exclusively on auditory information under visual-only occlusion. Taken together, participants do not seem to benefit from audiovisual information when tracking occluded objects but flexibly switch between senses during tracking, to account for expected unimodal or multimodal interference.

Surprisingly, no multisensory benefits in performance were found in either Experiment 1A or 1B, suggesting that participants rely exclusively on one unisensory component when tracking audiovisual objects under occlusion. One possibility is that these findings are specific to occlusion; when multisensory stimuli are occluded, and thus need to be extrapolated for a period of time, only a single sensory component is used (probably the most reliable component). Alternatively, one could argue that observers always rely on a single unisensory component when tracking multimodal objects, irrespective of whether they are occluded or not. Experiment 2 aimed at distinguishing between these two possibilities, by following the same procedure as Experiment 1, but omitting the occluders. Thus, instead of a motion prediction task, participants now performed a motion perception (or tracking) task in Experiment 2.

Experiment 2 (No Occluder)

Method

Participants

Based on the same sample size estimation as Experiment 1B, a total of 36 participants were tested in Experiment 2. They all reported normal or corrected-to-normal vision and normal hearing. The inclusion criteria were identical to those of Experiment 1 and ensured that participants could perform the task above chance level

in all three conditions (auditory, visual, and audiovisual). This led to the exclusion of two participants. Thus, data of 34 participants (28 participants reported their gender as female, six as male; M_{age} = 24.35 years, SD = 4.72, range = 19–44 years) were included in the final analysis. All participants signed informed consent and received money or course credits for their participation.

Procedure

The experimental setup, stimuli, experimental design, timing, and task were identical to those of Experiment 1, except that no occluder was presented (Figure 3a). Here, participants could perceptually track the objects throughout the entire trial and indicate where the target was when it disappeared. The data analysis was identical to that of Experiment 1.

Results

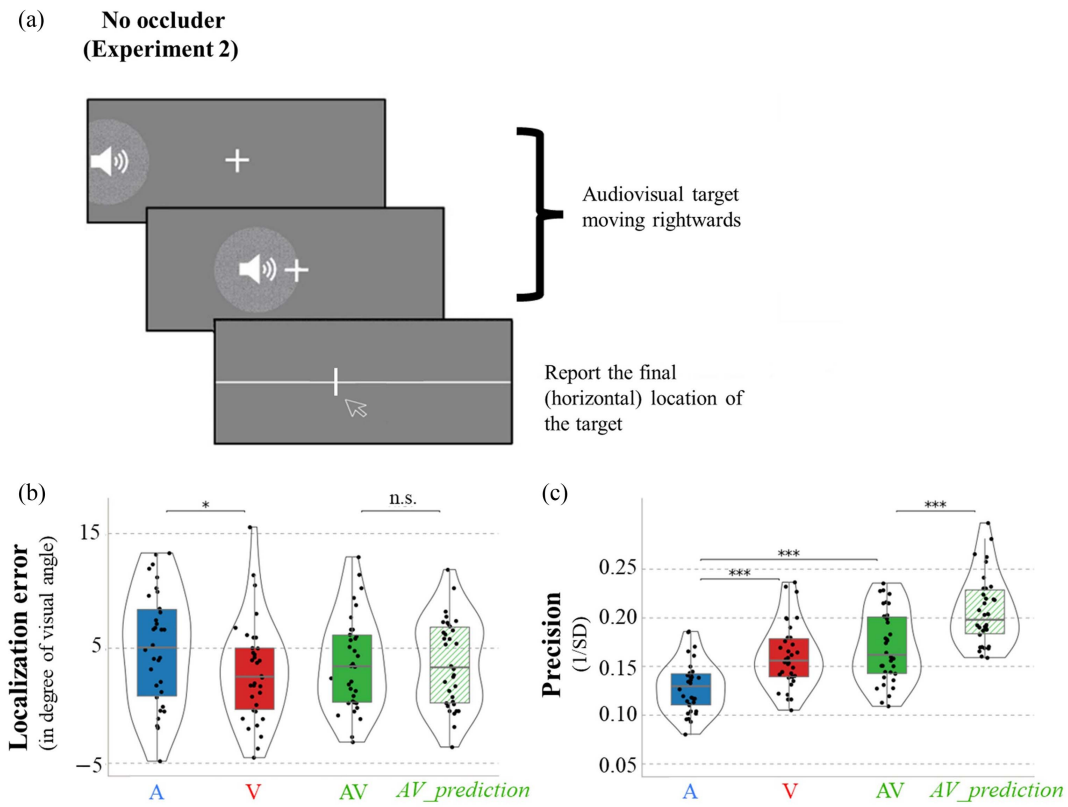
Usefulness of Localization Information

First, we set out to establish that the auditory and visual components of the moving target stimuli provided useable location information. To this end, we included all participants and tested whether the reported target location correlated positively with the actual (end) position of the target. As expected, correlations in auditory-only, visual-only, and audiovisual target conditions were higher than chance level (i.e., the upper boundary of the 95% percentile of the permuted distribution; all r = .59–.60 > .03). In all conditions, the observed group mean correlations were more than 49 *SDs* higher than the mean of the permuted distribution (the null; all $p < .001$). This establishes that both modalities—by themselves—provided ample information to perform the motion perception task.

Localization Error and Precision

A repeated measures ANOVA performed on the localization error revealed a significant main effect of target type, $F(1.18, 38.91) = 3.90$, $p < .05$, $\eta_p^2 = .11$; see Figure 3b. Follow-up analyses showed that localization error for auditory targets was larger than that for visual targets, $M = 4.96$ versus 3.00 dva, $SE = .85$ versus .75; $t(33) = 2.75$, $p < .05$, $d = .44$, but there was no significant difference between the audiovisual condition and either the auditory, $M = 3.68$ versus 4.96 dva, $SE = .70$ versus .85; $t(33) = -1.79$, $p > .05$, or

Figure 3
Procedure and Results of Experiment 2 (No Occluder)



Note. (a) Schematic depiction of a trial: On each trial, an auditory, visual, or audiovisual target moved horizontally at one of three constant speeds from left to right. After varying delays, the target disappeared and participants had to indicate where the target was when it disappeared. Localization error (b) and response precision (c) in auditory (blue), visual (red), and audiovisual (green) conditions, and MLE model prediction (shaded green). MLE = maximum likelihood estimation; A = auditory targets; V = visual targets; AV = audiovisual targets; n.s. = not significant. See the online article for the color version of this figure.
* $p < .05$. *** $p < .001$.

visual condition, $M = 3.68$ versus 3.00 dva, $SE = .70$ versus .75; $t(33) = .96$, $p > .05$.

A repeated measures ANOVA performed on response precision revealed a significant main effect of target type, $F(1.81, 59.79) = 33.44$, $p < .001$, $\eta_p^2 = .50$ (Figure 3c). Follow-up analyses showed that audiovisual targets yielded more precise localization responses than auditory targets, $M = .17$ versus .13, $SE = .006$ versus .004; $t(33) = 7.77$, $p < .001$, $d = 1.28$, but did not differ from visual targets, $M = .17$ versus .16, $SEs = .006$; $t(33) = 1.67$, $p > .05$. Visual targets also yielded more precise localization responses than auditory targets, $M = .16$ versus .13, $SE = .006$ versus .004; $t(33) = 6.10$, $p < .001$, $d = 1.00$.

Maximum Likelihood Estimation

To test whether auditory and visual information were combined in an MLE manner (i.e., weighted based on their relative reliability), we conducted a paired-samples *t* test between the performance in the audiovisual target condition and the MLE predictions that are derived from the two unisensory conditions. The localization error in the audiovisual condition did not differ from the MLE prediction, $M = 3.68$ versus 3.73 dva, $SE = .70$ versus .65; $t(33) = -.19$, $p > .05$. However, audiovisual precision was worse than the MLE prediction, $M = .17$ versus .21, $SE = .006$; $t(33) = -8.21$, $p < .001$, $d = -1.41$ (see Figure 3). These results show that localization accuracy was in line with MLE predictions, yet no multisensory benefit was found for precision, suggesting that unisensory information were not fully combined in an MLE consistent manner.

Utilization of Auditory and Visual Information

Considering audiovisual targets did not provide any performance benefits over unisensory targets, we further tested whether participants' performance fully relied on a single unisensory component, completely ignoring the other. First, we tested whether individual participants' performance in the audiovisual condition was more similar to itself than to performance in the unisensory conditions (using the same split-half bootstrapping procedure). We found that for both localization error and precision, behavior in the audiovisual target condition could be distinguished from that in the two unisensory conditions. This established that both components of the multisensory stimuli affected behavior, despite not being combined in a way that yielded performance benefits (for more details, see Supplemental Materials 1.3).

Second, we compared the observed weights to the predicted weights and to the (single modality simulated) weights that would be observed if only one modality (i.e., vision) contributed to behavior. We found that the reliance on visual input in the actual data of the audiovisual condition (observed visual weight: 70%) is more extreme than the MLE predictions (predicted visual weight: 60%) and less extreme than single visual modality simulations (simulated visual weight: 80%). This provides further evidence that both components of the multisensory stimuli contributed to behavior. That is, participants did not fully rely on one sense while completely ignoring the other (for more details, see Supplemental Materials 1.4).

In sum, these results show that although multisensory stimuli yielded no precision benefits compared to (the more reliable) unisensory visual stimulus, the other (less reliable) unisensory auditory component was not completely discarded.

Interim Discussion

To summarize, when participants localized moving targets (without occlusion), no multisensory benefits were found for precision compared to unisensory visual precision, suggesting that auditory and visual inputs were not integrated. Most importantly, even though unisensory auditory and visual inputs of the same object both contained useable location information, participants seemed to localize moving audiovisual targets exclusively based on unisensory visual information.

In the motion-tracking tasks of Experiments 1A, 1B, and 2, no clear multisensory benefits were observed. This stands in stark contrast with the multisensory benefits reported in the literature for static stimuli (e.g., Alais & Burr, 2004b; Freeman et al., 2018; Meijer et al., 2019). Thus, to ensure that our current stimuli and setup were suitable for picking up such multisensory benefits, we next conducted a localization task with the same auditory, visual, and audiovisual stimuli, but now using static targets. Experiment 3 was identical to Experiment 2, except that targets were now presented statically for 200 ms at the ending positions of the moving stimuli in Experiment 2.

Experiment 3 (Static Targets)

Method

Participants

Based on the same sample size estimation as Experiment 2, a total of 55 participants were tested in Experiment 3. They all reported normal or corrected-to-normal vision and normal hearing.

The inclusion criteria were identical to those of Experiment 2 and ensured that participants could perform the task above chance level in all three conditions (auditory, visual, and audiovisual). This led to an exclusion of 21 participants. This larger number of exclusions was presumably due to the task difficulty, which was followed from the very brief presentation duration (200 ms; as compared to multiple seconds in Experiments 1 and 2), and the low stimulus contrast, which was identical to that of Experiments 1 and 2. Thus, the data of 34 participants (30 participants reported their gender as female, four as male; $M_{age} = 22.85$ years, $SD = 2.93$, range = 18–30 years) were included in the final analysis. All participants signed informed consent and received money or course credits for their participation.

Procedure

The experimental setup, stimuli, and experimental design were identical to those of Experiment 2, except that a localization task was adopted. Here, on each trial, an auditory, visual, or audiovisual target (a very brief auditory and/or visual transient) was presented statically for 200 ms at the end positions of the corresponding moving targets in Experiment 2. Participants were required to indicate the location of the target. Data analysis was identical to that of Experiment 2.

Results

Usefulness of Localization Information

First, we set out to establish that the auditory and visual components of the moving target stimuli provided useable location information. To this end, we included all participants, and tested whether the reported target location correlated positively with the actual (end) position of the target. As expected, correlations in all three conditions were higher than chance level (i.e., the upper boundary of the 95% percentile of the permuted distribution; all $r = .24$–$.62 > .02$). In all conditions, the observed group mean correlations were more than 23 *SDs* higher than the mean of the permuted distribution (the null; all $p < .001$). This establishes that both modalities—by themselves—provided ample information to perform the localization task.

Localization Error and Precision

No significant effects were found with a repeated measures ANOVA performed on the localization error (all $p > .05$; Figure 4a). For precision, a main effect of target type was found, $F(1.69, 55.90) = 21.11$, $p < .001$, $\eta_p^2 = .39$; Figure 4b. Follow-up analyses showed that audiovisual targets yielded more precise localization responses than both auditory targets, $M = .14$ versus $.11$, $SE = .005$ versus $.006$; $t(33) = 4.29$, $p < .001$, $d = .77$, and visual targets, $M = .14$ versus $.10$, $SE = .005$ versus $.003$; $t(33) = 6.37$, $p < .001$, $d = 1.14$, while no significant differences between the auditory and visual conditions were found, $M = .11$ versus $.10$, $SE = .006$ versus $.003$; $t(33) = 2.08$, $p > .05$. These results confirmed a significant multisensory performance benefit.

Maximum Likelihood Estimation

To test whether auditory and visual information were combined in an MLE fashion, we conducted a paired-samples *t* test between the performance in the audiovisual condition and the MLE prediction. Localization error in the audiovisual condition did not differ from the MLE prediction, $M = .92$ versus 1.16 dva, $SE = .37$ versus $.34$; $t(33) = -.86$, $p > .05$. However, despite the audiovisual precision benefit, audiovisual targets yielded worse precision than expected by MLE, $M = .14$ versus $.16$, $SE = .005$ versus $.006$; $t(33) = -4.91$, $p < .001$, $d = -.84$; Figure 4, suggesting participants combined the auditory and visual information in a near-MLE fashion.

Utilization of Auditory and Visual Information

Considering that a substantial multisensory precision benefit was found, participants relied on both auditory and visual information. To allow for a direct comparison with Experiments 1 and 2, however, we still analyzed the extent to which participants' performance relied on both auditory and visual information. First, we found that for both localization error and precision, behavioral responses in the audiovisual target condition could be distinguished from trials in the two unisensory conditions. This provides further evidence that both components of the multisensory stimuli were used to perform the task (for more details, see Supplemental Materials 1.3).

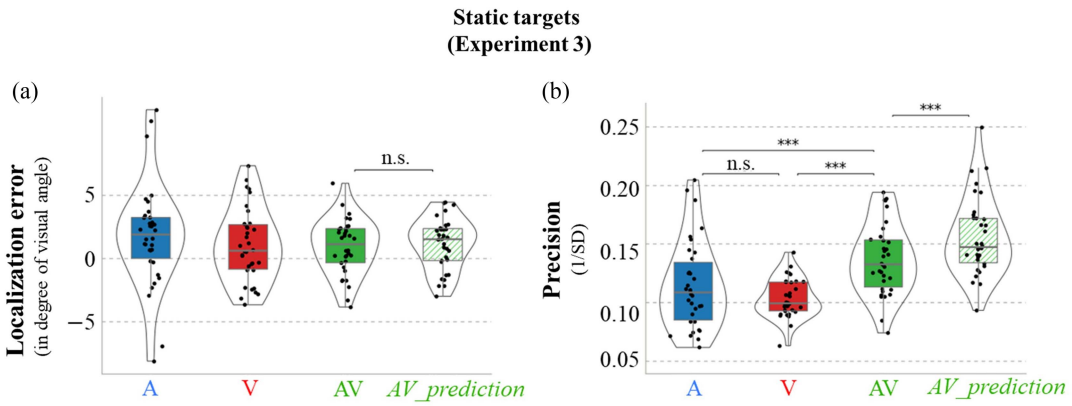
Second, we again compared the observed weights to the predicted weights and single modality simulated weights. We found that the reliance on visual input in the actual data of the audiovisual condition (observed visual weight: 70%) is more extreme than the MLE predictions (predicted visual weight: 50%) and less extreme than single visual modality simulations (simulated visual weight: 100%). This shows that both components of the multisensory stimuli were used (for more details, see Supplemental Materials 1.4).

In sum, these results confirm that participants used information from both modalities when localizing static audiovisual objects.

Comparison Across the Four Experiments

The extent to which information from different senses is used and combined to guide behavior seemed to differ substantially between experiments (e.g., the presence and type of occluder, and stimulus motion). To statistically compare the findings obtained under these

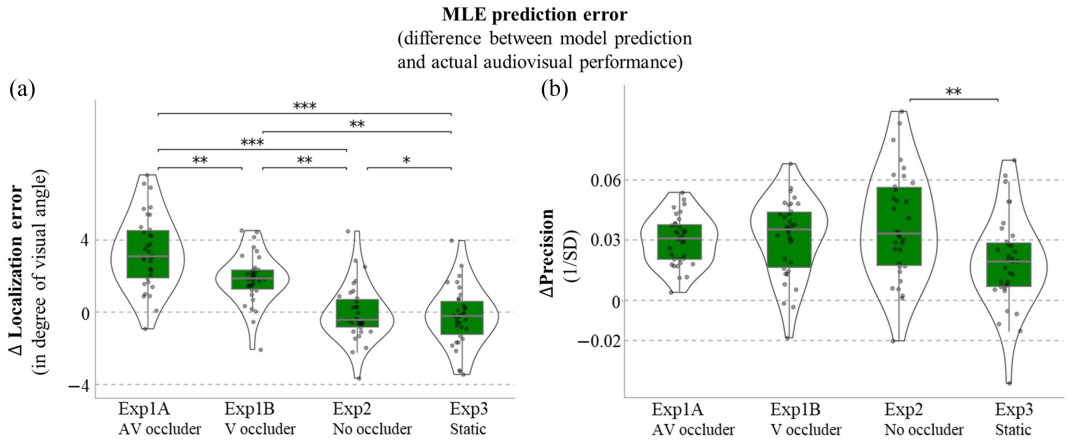
Figure 4
Results of Experiment 3 (Static Targets)



Note. Localization error (a) and response precision (b) in auditory (blue), visual (red), and audiovisual (green) conditions, and model prediction (shaded green). A = auditory targets; V = visual targets; AV = audiovisual targets; n.s. = not significant. See the online article for the color version of this figure.
*** $p < .001$.

Figure 5

Across-Experiment Comparison of MLE Prediction Error for Localization Error (a) and Precision (b)



Note. MLE = maximum likelihood estimation; Exp = Experiment; V = visual targets; AV = audiovisual targets; n.s. = not significant. See the online article for the color version of this figure.
** $p < .01$. *** $p < .001$.

different types of situations, we subtracted the MLE prediction from the performance in the audiovisual conditions for both localization error and precision to obtain MLE prediction errors. For visualization, we flipped the sign for precision, such that most of the data values were larger than 0 (see Figure 5). We conducted a one-way ANOVA (Experiments: 1A, 1B, 2, 3) to compare MLE prediction errors across experiments. The ANOVA performed on localization error revealed a significant main effect of experiment, $F(3, 132) = 35.58$, $p < .001$, $\eta_p^2 = .45$. Follow-up analysis showed that the MLE prediction error was the largest for Experiment 1A (audiovisual occluder; all $t > 3.65$, all $p < .01$, all $d > .88$, see Figure 5a). The MLE prediction error was also larger for Experiment 1B (visual occluder) than Experiments 2 (no occluder) and 3 (static targets; all $t > 4.75$, all $p < .001$, all $d > 1.15$), while no differences were found between Experiments 2 and 3, $M = -.05$ versus $-.24$, $t(33) = .48$, $p > .05$.

A one-way ANOVA on the MLE prediction error for precision also revealed a significant main effect of Experiments, $F(3, 132) = 4.45$, $p < .01$, $\eta_p^2 = .09$. Follow-up analysis showed that the MLE prediction error was smaller for Experiment 3 (static target) than for Experiment 2 (no occluder), $M = .02$ versus $.04$, $t(33) = -3.59$, $p < .01$, see Figure 5b, while no differences were found for other post hoc comparisons (all $|t| < 2.36$, all $p > .05$).

Overall, these results suggest that performance is better predicted by MLE for the static target (i.e., without motion). The model might fail to capture multisensory performance in particular when the target is moving.

Interim Discussion

Overall, the results of Experiment 3 show that participants integrate the auditory and visual components when localizing a static audiovisual target object. In terms of localization error, the performance in the audiovisual condition followed MLE predictions. Importantly, in terms of precision, a substantial multisensory benefit was found (albeit smaller than model predictions), thus showing near-MLE performance.

General Discussion

In the present study, we investigated how observers weigh the input from different senses when localizing brief static, moving, and occluded multisensory stimuli. To this aim, we conducted a series of four experiments with the same general stimulus and task properties. We compared localization performance for unimodal (auditory and visual) and multimodal (audiovisual) stimuli while observers were tracking audio-visually occluded (Experiment 1A), visually occluded (Experiment 1B), and nonoccluded (Experiment 2) moving objects, and when localizing static objects (Experiment 3). When tracking occluded objects and tracking objects without occlusion, participants relied almost exclusively on a single sensory modality (there was no multisensory benefit). Which sense participants relied on depended on the modality of the occluder (i.e., expected interference) and the relative reliability of the sensory input. Instead, when locating static audiovisual objects, participants used both auditory and visual information. A substantial multisensory precision benefit was found only when locating static audiovisual objects. For static stimuli, localization performance was in line with MLE model predictions, where auditory and visual information were weighted according to their respective reliabilities. Together, these results show near-MLE audiovisual integration for locating static, but not for moving occluded or unoccluded objects.

The present study had a number of key strengths. First, our experimental setup uniquely allowed us to compare multisensory benefits across a range of different task contexts. Second, we established that participants could use both the unisensory components and the combined multisensory components to perform all tasks above the chance level. Thus, we ensured that the unisensory components of multisensory stimuli were both useful for the task at hand, yet participants did not (optimally) use it in some multisensory contexts.

Finally, our bootstrapping approach allowed us to explore what the results would look like if participants relied only on one sense and totally ignored the other. This analysis confirmed that participants did not totally ignore one sense, as it affected behavior, even if it did not yield performance benefits.

In Experiment 1A, we found that participants localized audiovisual moving targets exclusively based on visual information that was available prior to audiovisual occlusion. In contrast, in Experiment 1B, participants relied exclusively on the still available auditory information when localizing audiovisual moving objects during visual-only occlusion. Why would participants rely exclusively on one sense, considering that the auditory and visual modalities both provided substantial location information when presented in isolation? We consider two explanations. One possibility is that tracking occluded objects requires a large amount of attention resources, which is an effortful process. The behavioral performance benefit of using a secondary sense might not outweigh the added cost of the additional attention resources. As a result, participants may single out one sense to localize the occluded object. Another possible explanation is that visual and auditory motion prediction might rely on different underlying modality-specific processes, which are not easily combined to generate unified behavioral responses. For instance, studies have shown that visual motion prediction tasks could involve mental extrapolation or visual imagery (DeLucia & Liddell, 1998; Schiff & Oldak, 1990), while auditory motion prediction tasks could involve a timing mechanism and spatial representations where participants count the elapsed time to track objects under occlusion (DeLucia et al., 2016). In our study, observers adjusted their location predictions to the speed of the object prior to occlusion (also see Supplemental Materials 1.2), also when tracking occluded auditory-only objects under occlusion. It seems therefore unlikely that observers counted the passage of time to track occluded auditory objects. In our view, the most plausible explanation as to why participants rely exclusively on one modality when localizing occluded objects is to minimize the usage of attention resources.

This leads us to inquire under what circumstances observers rely on visual information or on auditory information for localizing occluded audiovisual objects. Experiment 1 demonstrates that which senses participants relied on for localizing occluded audiovisual objects is context-dependent. Specifically, in Experiment 1A where we used an audiovisual occluder, participants could expect both the visual and auditory components of the target object to be occluded. In that case, participants relied exclusively on visual information to locate the occluded object. One explanation for this is that observers have a tendency to rely more on visual information, even when visual and auditory components are equally informative (so-called "visual dominance," Alais & Burr, 2004b; Colavita, 1974; Spence, 2009). Another explanation is that, in this case, the visual component provided more reliable location information than the auditory component (based on performance in the unimodal condition). Importantly, the pattern of results was reversed in Experiment 1B where we used a visual-only occluder. Here, participants could expect the auditory component of the stimulus to be unaffected by the occluder, and observers relied exclusively on the auditory information. Note that the only difference between Experiment 1A and 1B was the nature of the occluder (audiovisual or visual-only). These findings show that, in parallel to a (potentially obligatory) visual dominance effect, participants can flexibly choose which sense to rely on, to account for expected interference.

These results show that auditory and visual components of a multisensory stimuli are weighted differently depending on expected (modality-specific) interference and sensory reliability. In the present study, other factors that may also affect the weighting are task instructions and response modality. In a study done by Prime and Harris (2010), participants performed an audiovisual motion prediction task with a visual-only occluder, akin to our Experiment 1B. In contrast to our findings, where participants relied exclusively on the auditory component, they found that participants relied predominantly on the visual component. One possible explanation is that in their task, participants were explicitly instructed to predict the future position of the visual component of the audiovisual stimulus, which could have instilled a visual bias in the task interpretation. Previous studies have shown that prestimulus attention to vision can increase visual precision and thereby the weight of visual inputs (Badde et al., 2020; Ferrari & Noppeney, 2021). In the present study, participants were not given any modality-specific instructions; they were required to predict the future position of the audiovisual object. Together, these studies suggest that the instructions could affect the weighting of audiovisual information for tracking occluded objects. It should be noted that in both Prime and Harris' study and ours, participants were required to report the location of the target in the visual modality (by indicating a location on the screen). The nature of the report task could cause participants to rely more on visual information. Future studies could also adopt a nonvisual (auditory) report task to investigate the influence of response modality; for instance, by reporting the position in the auditory domain by adjusting the sound location or by making head-saccades in a fully darkened room.

When observers had to locate moving objects without occlusion (Experiment 2), patterns of localization responses in the audiovisual condition were in line with MLE predictions. Auditory and visual information were not combined optimally, however, as observers relied mainly on the visual component of the audiovisual target, resulting in no multisensory precision benefit. As discussed above, in the case of occlusion, participants might rely exclusively on one sensory modality because extrapolating location information from two sensory modalities is too resource costly. This statement stays true in the case of motion without occlusion. Participants might still extrapolate the future location of the target based on speed. Thus, it is possible that tracking a moving object (in a straight line) engages motion extrapolation, even in the absence of occlusion. And therefore, even in the absence of an occluder, observers use only a single modality to localize a moving object, to minimize the usage of attention resources. An alternative reason for why we observed no multisensory benefit when observers tracked moving objects is that the moving visual stimulus is much more reliable than the auditory stimulus, as reflected in the difference in precision. The additional location information provided by the auditory information did not outweigh the additional resource cost of tracking the auditory component of the audiovisual stimulus. Previous studies focusing on multisensory motion perception have also shown that the visual motion typically dominates over other modalities (Soto-Faraco et al., 2004) and bimodal motion yields only a small improvement in the threshold of detecting motion direction compared to unimodal motion (Alais & Burr, 2004a), which is consistent with our results and the latter interpretation. However, the visual motion dominance does not necessarily mean that no multisensory benefits occur in motion perception. Studies have found selective integration of audiovisual moving targets when objects are approaching but not when

receding in depth (Cappe et al., 2009, 2012). In our Experiment 1 and 2, with horizontal motion, participants exclusively relied on one sense, resulting in no multisensory integration. It seems that audiovisual information is not optimally combined when both sensory inputs are moving horizontally.

How is it possible that, in Experiments 1 and 2 where unisensory auditory and visual information contained ample localization information, participants nonetheless exclusively used unisensory information to localize a moving object, resulting in no clear multisensory benefits? One possibility that we considered was that our experimental setup (stimuli, task, etc.) was somehow not suited to reliably establish multisensory performance benefits. To address this possibility, we conducted Experiment 3, where we required participants to localize static target objects for which multisensory benefits have been reliably established (Alais & Burr, 2004b; Freeman et al., 2018; Meijer et al., 2019). With static target objects, we found substantial multisensory benefits in precision, consistent with previous studies. Importantly, these multisensory benefits were found using the same behavioral setup (same equipment and preparation before conducting the experiments, as well as the same logic of Matlab code for generating the condition and trial matrix) and stimuli as in Experiments 1A, 1B, and 2, in which no such multisensory benefits were found. By doing so, we established that the absence of multisensory benefits in Experiments 1A, 1B, and 2 must be attributed to the differences in task design, such as the presence and type of occluder, and stimulus motion or difference in the perceptual task (motion prediction vs. localizing moving stimuli vs. localizing static stimuli) and are unlikely to be the result of the basic stimulus properties or experimental setup. It remains an open question why moving stimuli would evoke no (or less) multisensory benefits in precision in the context of our experiments.

One possible explanation for the lack of multisensory benefits for motion stimuli in Experiments 1 and 2 is that a multisensory benefit in precision can only be found when the localization of unisensory auditory and visual conditions are similar, as proved by Otto et al. (2013). In Experiments 1A and 1B, the localization error for auditory and visual unimodal targets differed significantly. This could cause participants to prioritize the location that the more reliable sensory input indicates, at the expense of the less reliable sensory input which adds only limited location information, thus attenuating any multisensory precision benefits. This could also explain previous work, in which performance in the audiovisual condition was limited by performance in the visual condition (Hofbauer et al., 2004). Similarly, in Experiment 2, the localization error for auditory and visual unimodal targets differed significantly; and again no multisensory precision benefit was found. Following this line of reasoning, only in Experiment 3, in which the localization error for auditory and visual targets did not differ, did we find a multisensory benefit for precision. This is in line with the Bayesian causal inference model (Körding et al., 2007; Rohe & Noppeney, 2015) that multisensory enhancement is greatest when auditory and visual stimuli occur closely in space (within a spatial integration window) and is likely to come from a common source (Holmes & Spence, 2005; Otto et al., 2013).

Regardless of there being multisensory benefits or not, precision in the audiovisual condition was always worse than MLE predictions across all four experiments, which we took as evidence that integration of the auditory and visual components was close to MLE at best (and nonexistent at worst). Does the model truly predict the audiovisual conditions differently across experiments? To systematically test the differences in MLE prediction error, we compared the differences between actual performance in audiovisual conditions and predictions made by MLE at group level and found that the MLE could capture the performance better in the context of localization than in the context of motion perception and motion prediction. This result is consistent with our findings that we only found substantial multisensory integration when localizing static stimuli. An alternative possibility, however, is that model does not adequately describe the data pattern in the case of moving stimuli. To test this, we correlated the actual performance in audiovisual conditions with predictions made by MLE across participants (see Supplemental Materials 1.5). We showed that MLE did capture relevant variance in the data, even in the experiments with moving targets, although it did so more poorly. This suggests that, overall, the MLE model was successful in capturing relevant aspects of the data. Thus, the finding that participants' performance in the audiovisual condition was less than predicted by the MLE shows that participants indeed performed less well than if they optimally combined information from both modalities.

Not all senses convey the same amount of information to guide target localization. Accordingly, in Experiment 1A, vision conveyed more information than hearing, and in Experiment 1B, the opposite was the case. While observers successfully took this into account, by weighing the more reliable sense more than the less reliable sense, they consistently appeared to overdo it. That is, the more reliable sense was weighed even more strongly than predicted by MLE predictions (i.e., more strongly than they should, going by the relative variances of the two unimodal conditions) to the extent that one sense seemed not to contribute to multisensory performance at all. One possibility is that the weighing of senses inherently occurs in an all-or-none fashion, selecting the most informative modality and discarding the other. A number of analyses, however, argue against this possibility. First, the weights were more extreme than predicted but were typically less extreme than 1:0. Second, split-half correlation analyses showed that, typically, audiovisual conditions yielded behavior that was distinguishable from auditory-only and visual-only conditions (even in the absence of a performance benefit), thus the less reliable sense was not completely ignored and at least processed to some extent.

Thus, another interesting possibility is that the MLE prediction, describing how to best combine the two modalities, is too optimistic. One key assumption of the MLE model is that all the variance in the two unimodal conditions can be combined, to generate an audiovisual response of reduced variance. This requires that the variance in the two unimodal conditions is unique to that condition and independent from that of the other condition. Some variance, however, might be shared between the unimodal conditions. Such shared variance could include, for instance, noise that is introduced after the stimuli have been shown (e.g., behavioral response noise) or task-specific noise that similarly affects performance in both unimodal conditions (e.g., a tendency to overshoot, a bias toward the center of the screen, or attractive/repulsive effects from the response in the previous trial; Otto & Mamassian, 2012; Otto et al., 2013). The more variance is shared by the two unimodal conditions, the more optimistic the (traditional) MLE predictions are, and the more weights predicted by the MLE are less extreme than the actual optimal weighing of the auditory and visual

components. In the present case, these considerations could imply that (a) participants actually weighed the auditory and visual components of the audiovisual moving targets optimally in Experiments 1 and 2 and that (b) static audiovisual targets in Experiment 3 were actually optimally combined (but MLE predictions were too optimistic).

To assess to what extent our conclusion would change, depending on the amount of shared variance in our experiments, we generated "adjusted" MLE predictions for precision, in which we gradually increased the proportion of shared variance from 0% (no shared variance) to 100% (all of the variances in one condition is shared with the other conditions). For Experiments 1 and 2, we found that, even assuming that very substantial amounts of the observed variance were shared between conditions (i.e., up to 80% of the observed variance), the conclusions remained that hearing and vision were not optimally combined. Conversely, in Experiment 3, it might be the case that hearing and vision were in fact optimally combined (although this was not the case according to traditional MLE predictions), if we assume that at least 60% of the observed variance was shared between conditions. Similarly, we generated "adjusted" MLE predictions for weight. We found that the weights predicted by the adjusted MLE (although more extreme than those predicted by the traditional MLE) never reached the observed weight in Experiments 1A, 2, and 3, irrespective of the assumed level of shared variance. In Experiment 1B, the weights predicted by the adjusted MLE reached the observed weights if 92% or more of shared variance was assumed between conditions. These latter results suggest that participants tend to favor the most reliable component of a multimodal stimulus more than an ideal observer would. A detailed description of these analyses and results is provided in Supplemental Materials 1.6. Because we could not quantify the amount of shared variance in the current experimental task, it remains unknown whether participants in Experiment 3 (static targets) actually weighed hearing and vision optimally. However, these simulations do show that combining hearing and vision when localizing moving targets (Experiments 1 and 2) yields negligible benefits. More generally, we make the case that future implementations of the MLE model may benefit from accounting for shared variance (e.g., by adding an experimental condition that isolates the amount of shared noise or by estimating shared noise as an additional free parameter).

The comparisons of performance in the audiovisual conditions with (adjusted or traditional) MLE predictions allow us to determine whether or not observers optimally combined the two unimodal components for the localization of a multisensory target. Whether or not observers combined the two unimodal components at all was instead determined by the presence or absence of an advantage in the audiovisual condition compared to the best performing unimodal condition. Thus, our conclusions that observers only rely on a single modality in Experiments 1 and 2, partly rely on a null effect (although converging evidence is obtained from our analysis of weights, and our correlational analyses in Supplemental Materials 1.2 and 1.3). Because nonsignificant differences between conditions cannot be interpreted as evidence for the absence of a difference, we additionally performed Bayesian analyses testing whether performance in the audiovisual condition was better than that in the unimodal conditions (for details, see Supplemental Materials 1.7). Generally, these analyses confirmed our key results; localization error in Experiments 1 and 2 did not benefit from combining hearing and vision (yielding substantial evidence for the null); for precision, however, these analyses yielded inconclusive evidence for the null. Most importantly, in all four experiments, our conclusion that participants relied almost exclusively on unisensory information (Experiments 1A, 1B, and 2) or combined auditory and visual information (Experiment 3) follows not only from null effects in a single ANOVA but are based on converging evidence from multiple analyses (ANOVAs on localization error and precision, correlation between unimodal and multimodal conditions, and weight analyses). Together, these results imply that observers reap negligible benefits from combining hearing and vision when tracking moving stimuli.

To conclude, the current results suggest that human observers use both hearing and vision when localizing static objects, but use only unisensory input when localizing moving objects and when predicting motion under occlusion, perhaps to minimize attentional load. Moreover, observers can flexibly prioritize one sense over the other, in anticipation of modality-specific interference.

Constraints on Generality

Participants across four experiments in the present study were all recruited from the Netherlands. Most of the participants were students from Utrecht University, who are within a young age range (18–44 years). Considering the effects of age on multisensory integration (de Dieuleveult et al., 2017; Mozolic et al., 2012), we do acknowledge that our results might not necessarily generalize to older age groups. In particular, studies have shown that older adults tend to use all available sensory information to perform tasks at hand, thereby benefiting more from multisensory stimuli (compared to unisensory stimuli) than younger controls. For older adults, this comes at the cost of having difficulties properly weighing information from different sensory modalities, especially when sensory information is unreliable (e.g., when sensory information is distracted, disrupted, or taken away).

Moreover, we acknowledge that our results might not necessarily generalize to complex stimuli and real-life situations. Our data were acquired based on our specific task in the lab. The horizontal linear motion we used may limit the amount of integration. We acknowledge that motion in real life is not always linear, and the motion prediction depends on many more factors than it does in this reductionist lab study. Thus, our results may not be able to directly predict everyday behavior. We acknowledge that more studies need to be done to extend our findings to different setups and with different samples.

Apart from the potential lack of generalization to other age groups and complex real-life situations, as described above, it should be stressed that we study basic multisensory processes that are not expected to substantially differ between university students and the general population (e.g., in terms of gender, ethnic background, and education level). Because basic multisensory integration findings have been found to converge across mammal species (e.g., rats, cats, monkeys, humans), we do not expect such fundamental principles to substantially differ between groups of the same species (i.e., student and nonstudent populations). Thus, although the general population may generally do worse in the task (e.g., higher localization error in all conditions), the difference we observed in the unisensory and multisensory conditions will likely remain.

References

Alais, D., & Burr, D. (2004a). No direction-specific bimodal facilitation for audiovisual motion detection. *Cognitive Brain Research*, *19*(2), 185–194. https://doi.org/10.1016/j.cogbrainres.2003.11.011

Alais, D., & Burr, D. (2004b). 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

Badde, S., Navarro, K. T., & Landy, M. S. (2020). Modality-specific attention attenuates visual-tactile integration and recalibration effects by reducing prior expectations of a common source for vision and touch. *Cognition*, *197*, Article 104170. https://doi.org/10.1016/j.cognition.2019.104170

Battaglini, L., Campana, G., Camilleri, R., & Casco, C. (2015). Probing the involvement of the earliest levels of cortical processing in motion extrapolation with rapid forms of visual motion priming and adaptation. *Attention, Perception, & Psychophysics*, *77*(2), 603–612. https://doi.org/10.3758/s13414-014-0795-z

Battaglini, L., Casco, C., Isaacs, B. R., Bridges, D., & Ganis, G. (2017). Electrophysiological correlates of motion extrapolation: An investigation on the CNV. *Neuropsychologia*, *95*, 86–93. https://doi.org/10.1016/j.neuropsychologia.2016.12.019

Battaglini, L., Contemori, G., Maniglia, M., & Casco, C. (2016). Fast moving texture has opposite effects on the perceived speed of visible and occluded object trajectories. *Acta Psychologica*, *170*, 206–214. https://doi.org/10.1016/j.actpsy.2016.08.007

Battaglini, L., & Ghiani, A. (2021). Motion behind occluder: Amodal perception and visual motion extrapolation. *Visual Cognition*, *29*(8), 475–499. https://doi.org/10.1080/13506285.2021.1943094

Battaglini, L., Maniglia, M., Konishi, M., Contemori, G., Coccaro, A., & Casco, C. (2018). Fast random motion biases judgments of visible and occluded motion speed. *Vision Research*, *150*, 38–43. https://doi.org/10.1016/j.visres.2018.08.001

Battaglini, L., & Mioni, G. (2019). The effect of symbolic meaning of speed on time to contact. *Acta Psychologica*, *199*, Article 102921. https://doi.org/10.1016/j.actpsy.2019.102921

Baurès, R., Fourteau, M., Thébault, S., Gazard, C., Pasquio, L., Meneghini, G., Perrin, J., Rosito, M., Durand, J. B., & Roux, F. E. (2021). Time-to-contact perception in the brain. *Journal of Neuroscience Research*, *99*(2), 455–466. https://doi.org/10.1002/jnr.24740

Baurès, R., Maquestiaux, F., DeLucia, P. R., Defer, A., & Prigent, E. (2018). Availability of attention affects time-to-contact estimation. *Experimental Brain Research*, *236*(7), 1971–1984. https://doi.org/10.1007/s00221-018-5273-8

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Cappe, C., Thelen, A., Romei, V., Thut, G., & Murray, M. M. (2012). Looming signals reveal synergistic principles of multisensory integration. *The Journal of Neuroscience*, *32*(4), 1171–1182. https://doi.org/10.1523/JNEUROSCI.5517-11.2012

Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2009). Selective integration of auditory-visual looming cues by humans. *Neuropsychologia*, *47*(4), 1045–1052. https://doi.org/10.1016/j.neuropsychologia.2008.11.003

Chotsrisuparat, C., Koning, A., Jacobs, R., & van Lier, R. (2018). Effects of auditory patterns on judged displacements of an occluded moving object. *Multisensory Research*, *31*(7), 623–643. https://doi.org/10.1163/22134808-18001294

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*(2), 409–412. https://doi.org/10.3758/BF03203962

de Dieuleveult, A. L., Siemonsma, P. C., van Erp, J. B., & Brouwer, A. M. (2017). Effects of aging in multisensory integration: A systematic review. *Frontiers in Aging Neuroscience*, *9*, Article 80. https://doi.org/10.3389/fnagi.2017.00080

DeLucia, P. R., & Liddell, G. W. (1998). Cognitive motion extrapolation and cognitive clocking in prediction motion task. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 901–914. https://doi.org/10.1037/0096-1523.24.3.901

DeLucia, P. R., Preddy, D., & Oberfeld, D. (2016). Audiovisual integration of time-to-contact information for approaching objects. *Multisensory Research*, *29*(4–5), 365–395. https://doi.org/10.1163/22134808-00002520

Deutsch, P., Czoschke, S., Fischer, C., Kaiser, J., & Bledowski, C. (2023). Decoding of working memory contents in auditory cortex is not distractor-resistant. *The Journal of Neuroscience*, *43*(18), 3284–3293. https://doi.org/10.1523/JNEUROSCI.1890-22.2023

Dittrich, S., & Noesselt, T. (2018). Temporal audiovisual motion prediction in 2D- vs. 3D-environments. *Frontiers in Psychology*, *9*, Article 368. https://doi.org/10.3389/fpsyg.2018.00368

Erlikhman, G., & Caplovitz, G. P. (2017). Decoding information about dynamically occluded objects in visual cortex. *Neuroimage*, *146*, 778–788. https://doi.org/10.1016/j.neuroimage.2016.09.024

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferrari, A., & Noppeney, U. (2021). Attention controls multisensory perception via two distinct mechanisms at different levels of the cortical hierarchy. *PLOS Biology*, *19*(11), Article e3001465. https://doi.org/10.1371/journal.pbio.3001465

Flavell, J. C., Barrett, B. T., Buckley, J. G., Harris, J. M., Scally, A. J., Beebe, N. B., Cruickshank, A. G., & Bennett, S. J. (2018). Temporal estimation in prediction motion tasks is biased by a moving destination. *Journal of Vision*, *18*(2), Article 5. https://doi.org/10.1167/18.2.5

Freeman, L. C. A., Wood, K. C., & Bizley, J. K. (2018). Multisensory stimuli improve relative localisation judgments compared to unisensory auditory or visual stimuli. *The Journal of the Acoustical Society of America*, *143*(6), EL516–EL522. https://doi.org/10.1121/1.5042759

Hofbauer, M., Wuerger, S. M., Meyer, G. F., Roehrbein, F., Schill, K., & Zetzsche, C. (2004). Catching audiovisual mice: Predicting the arrival time of auditory-visual motion signals. *Cognitive, Affective & Behavioral Neuroscience*, *4*(2), 241–250. https://doi.org/10.3758/CABN.4.2.241

Holmes, N. P., & Spence, C. (2005). Multisensory integration: Space, time and superadditivity. *Current Biology*, *15*(18), R762–R764. https://doi.org/10.1016/j.cub.2005.08.058

Holt, C. A., & Sullivan, S. P. (2023). Permutation tests for experimental data. *Experimental Economics*, *26*(4), 775–812. https://doi.org/10.1007/s10683-023-09799-6

Keshavarz, B., Campos, J. L., DeLucia, P. R., & Oberfeld, D. (2017). Estimating the relative weights of visual and auditory tau versus heuristic-based cues for time-to-contact judgments in realistic, familiar scenes by older and younger adults. *Attention, Perception, & Psychophysics*, *79*(3), 929–944. https://doi.org/10.3758/s13414-016-1270-9

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLOS ONE*, *2*(9), Article e943. https://doi.org/10.1371/journal.pone.0000943

Lu, F., Li, Y., Yang, J., Wang, A., & Zhang, M. (2023). Auditory affective content facilitates time-to-contact estimation of visual affective targets. *Frontiers in Psychology*, *14*, Article 1105824. https://doi.org/10.3389/fpsyg.2023.1105824

Lugtigheid, A. J., & Welchman, A. E. (2011). Evaluating methods to measure time-to-contact. *Vision Research*, *51*(20), 2234–2241. https://doi.org/10.1016/j.visres.2011.08.019

Meijer, D., Veselič, S., Calafiore, C., & Noppeney, U. (2019). Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex*, *119*, 74–88. https://doi.org/10.1016/j.cortex.2019.03.026

Menceloglu, M., & Song, J. H. (2023). Motion duration is overestimated behind an occluder in action and perception tasks. *Journal of Vision*, *23*(5), Article 11. https://doi.org/10.1167/jov.23.5.11

Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., & Laurienti, P. J. (2012). Multisensory integration and aging. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 81–392). CRC Press, Taylor and Francis.

Otto, T. U., Dassy, B., & Mamassian, P. (2013). Principles of multisensory behavior. *The Journal of Neuroscience*, *33*(17), 7463–7474. https://doi.org/10.1523/JNEUROSCI.4678-12.2013

Otto, T. U., & Mamassian, P. (2012). Noise and correlations in parallel perceptual decision making. *Current Biology*, *22*(15), 1391–1396. https://doi.org/10.1016/j.cub.2012.05.031

Prime, S. L., & Harris, L. R. (2010). Predicting the position of moving audiovisual stimuli. *Experimental Brain Research*, *203*(2), 249–260. https://doi.org/10.1007/s00221-010-2224-4

Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLOS Biology*, *13*(2), Article e1002073. https://doi.org/10.1371/journal.pbio.1002073

Schiff, W., & Oldak, R. (1990). Accuracy of judging time to arrival: Effects of modality, trajectory, and gender. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 303–316. https://doi.org/10.1037/0096-1523.16.2.303

Schut, M. J., Van der Stoep, N., & Van der Stigchel, S. (2018). Auditory spatial attention is encoded in a retinotopic reference frame across eye-movements. *PLOS ONE*, *13*(8), Article e0202414. https://doi.org/10.1371/journal.pone.0202414

Soto-Faraco, S., Spence, C., Lloyd, D., & Kingstone, A. (2004). Moving multisensory research along: Motion perception across sensory modalities. *Current Directions in Psychological Science*, *13*(1), 29–32. https://doi.org/10.1111/j.0963-7214.2004.01301008.x

Spence, C. (2009). Explaining the Colavita visual dominance effect. *Progress in Brain Research*, *176*, 245–258. https://doi.org/10.1016/S0079-6123(09)17615-X

Tresilian, J. R. (1995). Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Perception & Psychophysics*, *57*(2), 231–245. https://doi.org/10.3758/BF03206510

Wessels, M., Zähme, C., & Oberfeld, D. (2023). Auditory information improves time-to-collision estimation for accelerating vehicles. *Current Psychology*, *42*, 23195–23205. https://doi.org/10.1007/s12144-022-03375-6

Received June 3, 2024
Revision received October 22, 2024
Accepted November 27, 2024 ■