When Vision Lies - Navigating Virtual Environments with Unreliable Visual Information

Eden Or*
Department of Biomedical Engineering
at Ben Gurion University of the Negev

Shachar Maidenbaum†
Department of Biomedical Engineering
and School of Brain sciences
at Ben Gurion University of the Negev

ABSTRACT

Humans typically utilize vision in a dominant role for navigation. However, what happens when vision becomes actively unreliable? Will it impair user performance, be suppressed, or be used advantageously? While such scenarios are rare in the real world, this question has important implications for multisensory integration in extended reality applications - e.g. virtual walls that a user sees but can walk through. We created virtual mazes which could be solved via audio or visual cues. We then manipulated the reliability of these sensory channels by including invisible walls which are not perceived but still blocked passage, and ghost walls which could be perceived but did not block participants. Participants navigated the exact same layouts under all conditions, and could solve these levels by ignoring the unreliable sensory modality and using only the other. Participants easily completed these mazes using vision-only, and with some difficulty via audition-only. Partially unreliable vision degraded performance, though still above audio-only demonstrating utilization of the unreliable visual cues. Mazes whose entire visual input was false degraded performance to the level of audio only, though participants subjectively reported it as easier then audio-only and did not close their eyes indicating that they still relied on vision. Testing a control in which visual information was both false and constantly moved, preventing its use as landmarks or optic flow, indeed caused participants to close their eyes, disregarding the false vision, but was accompanied by confounding nausea. In parallel, auditory incongruencies were easily suppressed across all unreliable auditory conditions. This demonstrates human attachment to visual information, even when mostly or completely false, and the ability to glean practical advantages from it unless it is completely stripped from usability. More broadly it lays a foundation for testing multisensory integration of sustained false sensory channels, and has implications for mixed reality design.

Keywords: Virtual Environment, Virtual Reality, Multisensory, Multimodal, Vision, Audition, Sensory Substitution, Navigation

1 INTRODUCTION

Spatial navigation through real and virtual environments relies on integrating sensory inputs from various sources [2], such as visual and auditory cues. These sensory inputs work together to create a coherent integrated perception of space, allowing us to understand our position, orientation, and movement within the environment [44]. In humans, the predominant mode of navigation in three-dimensional space is through the utilization of the visual sense [8, 35]. In virtual environments it plays an even greater role as other sensory channels are typically missing (e.g. taste and smell), or degraded (e.g. audio without echos) leaving vision as the main presentation channel. Many previous works tested the relationship between sensory channels during incongruencies, with emphasis on auditory and visual stimuli [10, 11, 14], but also in different modalities [4, 19, 26, 45, 46]. Those results are part of a broader research realm on psychophysical models of multisensory integration such as multisensory race models [32, 34, 37], bayesian integration and inference [1, 9, 17, 31, 43], and maximum likelihood estimation [39]. However, this body of work typically focused on momentary incongruencies (e.g. mismatched sources for a sound and image) or on degraded input (e.g. fog, noise) but what happens when these clashes go even further to being unreliable in a sustained fashion? What happens when our vision is lying to us as we navigate?

Why should we care about false visual information in navigation? In virtual environments, this is something that is already common - e.g. visual walls might be placed regardless of the boundaries of the real world, which the user might be able to just walk through, while the real world boundaries impose physical barriers that are not visible within the virtual environment. The guardian safety system deployed in many headsets [27], which brings up walls that were previously invisible when the user approaches predefined boundaries, is another common example of invisible walls. Beyond these common cases in virtual environments, such scenarios will be even more common in mixed reality applications where augmented objects will be added into the real world without necessarily having physical presence. Indeed, current mixed reality games and applications already frequently utilize augmented walls which the user is expected to treat as real despite not being barriers, augmented interface elements which visually block parts of the real world etc. Thus, we need to understand the basic science of how humans deal with these situations. How do we integrate clashing sensory information when one of our sensory channels is conveying false information?

There are three basic logical options for how the false information will be integrated (Figure 1) - (1) **Suppression** - The user might suppress this information and rely on the other correct sensory channels, which will lead to performance akin to not using the specific channel at all and relying just on a different channel. (2) **Interference** - The user might not be able to suppress the information, and integrating the false information into our multisensory preception will impair performance compared to not having that information at all and relying just on a different reliable channel. (3) **Enhancement** - a third intriguing option is that despite the sensory information being false users might be able to glean other useful information about their environment from it. In this case, we should see enhanced performance which will be better in comparison to not having input on the sensory channel and relying just on a reliable channel. If this is the case, it is especially interesting to understand what these advantageous aspects are, and to see if they can be disabled.

To explore this, we designed virtual mazes that could be successfully navigated using either audio or visual cues. We then adjusted the reliability of these sensory pathways by introducing invisible walls that participants couldn't perceive but still hindered progress, and ghost walls that participants could perceive but didn't obstruct their path. Each condition was repeated in identical layouts, allowing us to directly test the predictions of these three models by

*e-mail: edeo@post.bgu.ac.il
†e-mail:mshachar@bgu.ac.il

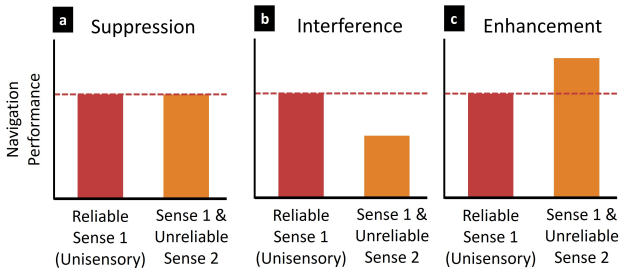

Figure 1: Integration models. The three potential models for integration of false sensory information from one unreliable sense with one other reliable channel (in orange), compared to performance using only the reliable channel (i.e. with no input from the unreliable channel - in red). Thus, the reliable sense with no input at all from the misleading sense sets the bar for performance on a task with unreliable sensory information (dotted bar). These include the unreliable chanel being (a) suppressed (i.e. performance is the same as without it), (b) interfering (i.e. performance is worse) or (c) enhancing (i.e. adding unreliable information improves performance)

comparing them to performance with only the reliable senses (See Figure 1) - will false sensory information be suppressed, enhance performance or interfere with it? And will there be differences between actual performance and the users subjective experience of their performance?

2 Related Work and Background

2.1 Studying spatial multisensory interactions in virtual environments

We are far from the first to test questions of multisensory integration and navigation with virtual environments. The advantages of virtual environments such as flexibility, simpler logistics, safety and relatively naturalistic task nature, have made them a key tool in recent years for research in general [3, 20, 40], and both navigation [5, 6, 12, 25] and multi-sensory integration [4, 7, 14, 31, 46] in particular. The ability to control and manipulate sensory cues and their separate locations, which are hard to separate in the real world, is especially useful for research into multi-sensory integration. Most multi-sensory integration research has focused on vision and audition, but recent work has gone beyond to other sensory channels, such as taste [29, 46], smell [38, 46] and haptics [19, 38], and even potential new augmented senses [15, 28, 30, 48]. For example, such research included developing a multisensory device which allows you to experience the seasons [38], a device which used smell and sound integration to explore visuospatial attention in VR [7] and work examining the changes in the perception of taste by food color [29].

2.2 Incongruence in multisensory interactions

One aspect of multisensory integration which is seeing recent interest is looking at the cases in which the sensory integration does not work well due to sensory signal degradation or clashes, typically in momentary time or location. Understanding the case in which integration is taxed, or even fails, can help us understand the underlying system of correct multisensory integration. This is especially true in VR where correctly synchronizing the different sensory channels is critical to avoid lags and feelings of nausea. Consequently, there is a growing body of research that explores these discrepancies and ways to utilize them. For some examples among many others, Finnegan et al. improved distance estimation using audio-visual incongruence [10], Kim and Lee investigated how different types of auditory-visual congruence affect VR experiences [14], and Weidner et al. focused on the impact of incongruence during eating in VR [46]. Our work here contributes to this research area, taking a step further by examining sustained clashes between sensory channels during navigation.

2.3 Auditory navigation in virtual mazes

As users typically navigate virtual environments mainly via visual cues, a key requirement for our work is to enable navigation using auditory sensory cues. The use of just spatial audio, such as the sound of bumping into walls or spatially localized cues for start and end positions, has previously been demonstrated to enable navigation [35, 41], but unfortunately such navigation is less efficient and leads to different mobility patterns then when navigating with vision [18, 24]. Another option is the use of echolocation, as demonstrated successfully in works such as [16, 30]. However navigation using echolocation requires significant training, and the processing power required for real time navigation is significant. Therefore we decided to integrate the EyeCane algorithm [22], a tool from the realms of sensory substitution originally developed for aiding blind users with obstacle detection and navigation by translating distance into frequency of auditory cues. The Eyecane has previously been used in both real and virtual setups [5, 22, 23], and has been shown to shift non-visual navigation patterns closer to visual ones [24]. This mapping is also similar to the auditory cues used in many modern vehicles for sensing obstacles while driving in ”reverse”. Thus, in our work here we combine spatial cues with the EyeCane’s distance-to-audio mapping to enable relatively fast and efficient navigation based on audio cues without vision.

3 Methods

3.1 Main experimental design - Experiment 1

To test how unreliable sensory information is perceived, integrated and treated by humans, we created a series of virtual mazes in which participants needed to navigate from a first person perspective. These virtual mazes could be perceived via vision and via audition and were designed such that we could disentangle the mazes presented in each sensory channel and the actual walls blocking the participant, thus manipulating the reliability of each sense. This led to two types of persistent clashes - “ghost” walls which could be perceived by a sensory channel despite not being there (3, can be thought of as a false-positive), and “invisible” walls which could not be perceived by a sensory channel despite being there (3, can be thought of as a false negative). Thus, a participant might see a wall but be able to walk through its location, or see an open corridor but still be blocked. We created 8 conditions (see examples of each in the Supp video) that can be divided into 2 main groups:

- **No clash conditions**:
  1. Visual-only: Navigation was with reliable visual input only, without sound for the walls.
  2. Auditory-only: Navigation was with reliable audio input only, with no visual input (black screen).
  3. Multisensory (audio and visual): Both visual and auditory channels were on and reliable.
- **Conditions with different degrees of unreliable inputs**:
  4. Invisible visual walls: Navigation was with reliable auditory information, but all walls were visually invisible. The maze was within a round arena to decrease the use of landmarks.
  5. Partial visual clash: The auditory sensory channel was fully reliable, while vision was reliable except for 2 ghost and 2 invisible walls.

6. Partial audio clash: The visual channel was fully reliable, while audio was reliable except for 2 ghost sound walls and 2 mute walls.
7. Full visual clash: The auditory sensory channel was fully reliable, while all visible walls were ghost walls and all real walls were invisible visually.
8. Full audio clash: The visual sensory channel was fully reliable, while all walls perceived via audio were ghost walls and all real walls were invisible to audition (mute).

Note that except for the changes mentioned above, the rest of the sensory input in the partial/full clash levels was normal (i.e., not degraded/blurred) - though as users did not know which was which, they had to suspect all walls. Users did not know if they were in a full or partial clash level, nor did they know how many false walls partial clash levels had.

Users were informed which sensory channels to trust at the start of each trial with only text instructions of the kind shown in fig 2d (e.g., "trust audio, vision could be misleading", "audio only"). Beyond this text, they did not receive an explanation of what "misleading" means and had to discover for themselves in the experiment. We used 4 different layouts (See Figure 4), to avoid learning effects. We had 8 conditions that each had to be performed in all 4 layouts, thus the users performed 32 trials during experiment 1. The orders of both conditions and layouts were pseudorandomized such that two consecutive trials would not share a condition or layout. (5). Our main measures were success rate, measured by reaching the end within a 2 minute time limit or giving up, time to completion of each trial, and the subjective difficulty ratings given by the participants to each type of trial. All of these parameters were logged.

3.2 Training and questionnaires

Prior to navigating the mazes, the participants filled a questionnaire that assessed their navigational abilities, and had a training session with three levels in order to acclimate themselves to the auditory feedback and the controls. Training levels covered multi-sensory, auditory-only and visual-only navigation but not clashes. At the end of the experiment participants filled a questionnaire regarding their experience, and rated the difficulty of each condition (audio-only, visual-only, invisible, visual-clash, auditory-clash, multi-sensory).

3.3 Experiment 2

Following the results of the first 18 participants in our main experiment, we performed a power calculation to determine the overall group size required to be properly statistically powered for an effect size of a Cohen's D of 0.8 in all of our comparisons, and the number of participants required to achieve gender-balance in our sample. Accordingly we then collected the rest of the results from 17 additional participants. These 17 participants performed the first experiment in an identical fashion to the first 18, however they also performed a second experiment afterwards which we added to test for the effects of a maze devoid of reliable vision. To do so we devised a stage in which both walls and floors were in constant motion in different directions, thus eliminating any consistent optic flow for navigation (See Video Supp). This experiment includes two trials in which the visual information was in constant motion while audition was reliable, and 2 trials in which vision was reliable but the audition was not (auditory cues were constantly rotating similar to the visual cues in the first control).

3.4 Auditory navigation

Navigating a virtual environment visually is commonplace. In order to enable participants to navigate via audition we utilized several types of audio cues. The main cue we used was a collision sound when participants collided with a wall. To facilitate faster and more efficient navigation we utilized the EyeCane algorithm [22] which

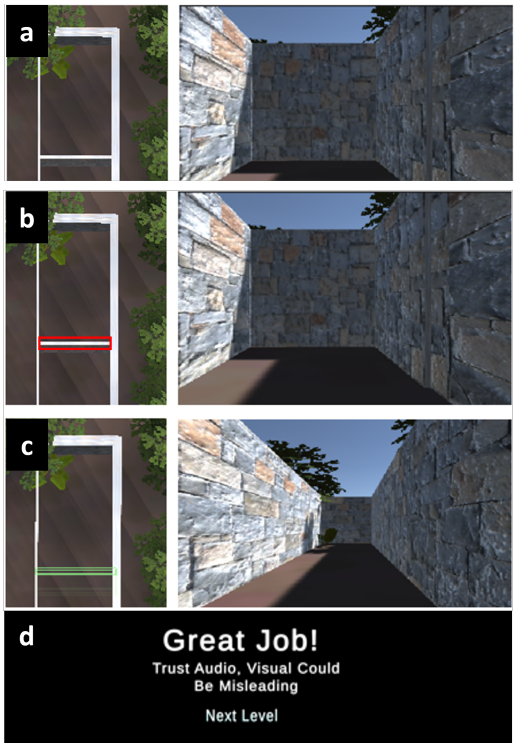

Figure 2: Walls and cues. a. An example of a regular wall from a top view (left) and from a first person view (right). b. An example of a ghost wall (red) from a top view (left) and from a first person (right). Note that from first person these both seem identical, however here the participant could walk through it. c. An example of an invisible wall (green) from a top view (left) and from a first person view (right). Note that here while the corridor looks open it is actually blocked by the invisible wall. Note that in all trials participants could only view from the first person. d. A between level scene congratulating a participant for completing the previous level and giving them the instructions for the next level - in this case the following level has reliable audio and unreliable vision.

transforms distance into sound frequency cues and has previously been used to enable virtual navigation [5, 24]. Finally, we added dedicated auditory cues for the start location and for the target which were activated by proximity. The selection of the specific sounds we used for each cue was determined through a preliminary experiment designed to identify the most effective audio cues for navigation.

3.5 Overall design

To summarize these sections, 35 participants signed their informed consent, filled out an entry survey, underwent a brief training session, and then performed the main experiment (3.1) and filled out a questionnaire. The last 17 participants also took part in experiment 2 (3.3). Finally participants were compensated for their time.

3.6 Experimental setup

To avoid contamination from other sensory channels (e.g. vestibular), and to provide compatibility with parallel neuroimaging in future stages of the study, we chose to utilize here desktop based virtual environments. Participants were seated comfortably in front of a standard desktop screen and used a standard keyboard. Controls included discrete turns of 45-degs (using left/right arrows), continuous forward/backward motion (using up/down arrow), thus lack of vision did not hamper moving in straight lines and were identical

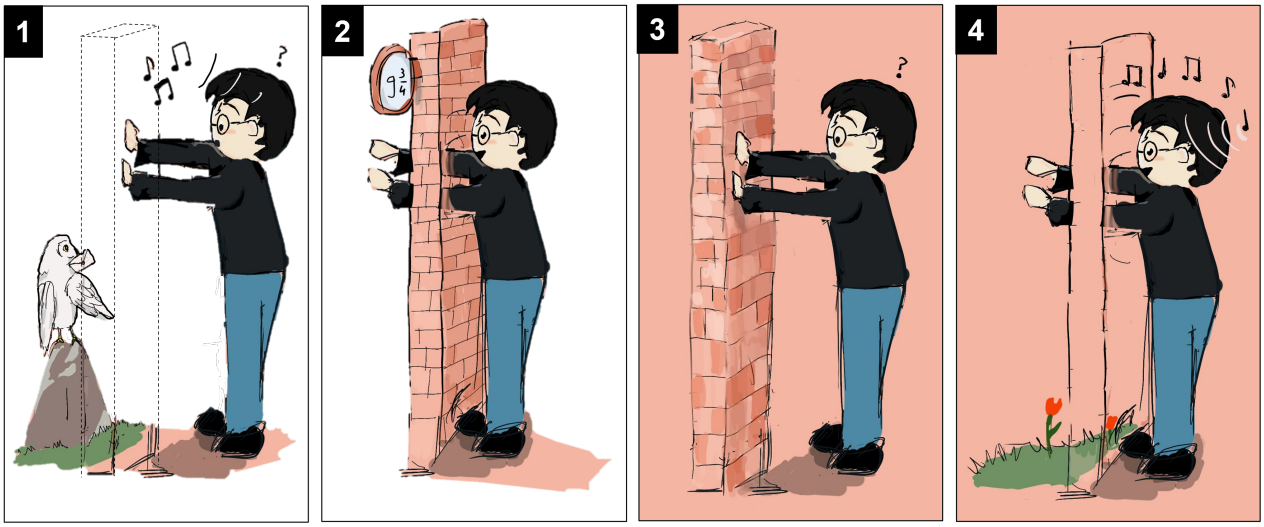

Figure 3: Illustration of the different types of clashes using different types of walls. 1. Invisible wall (false negative), which is audible but visually imperceptible, impeding progress during visual clash stages. 2. Ghost visual walls (false positive), do not produce any auditory cues and allow for uninhibited passage. 3. Mute wall (false negative), which isn't audible but is visible and will impede progress. 4. Ghost sound wall (false positive) - produces audio cues and allows passage.

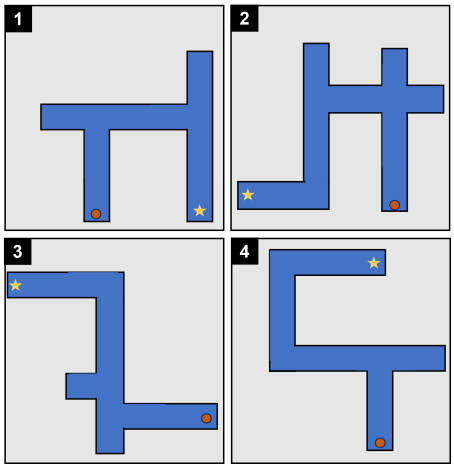

Figure 4: Schematic diagrams of the maze layouts from a top view. Stars mark the goal in each layout, and the orange circle mark the beginning locations of each trial.

for all conditions. Experiments were programmed using Unity 3D 2021.3.11f1, augmented with C# scripts.

3.7 Participants

35 participants (16 female, 19 male; aged 27.17±5.01 years) who reported normal levels of sight and hearing took part in experiment 1. The last 17 of these participants also completed experiment 2. Participants were compensated for their time by the standard rates table in our institute.

3.8 Ethics

This experiment was approved by the Ben Gurion University of the Negev IRB in accordance with the 1964 Helsinki Declaration and participants provided informed consent, and were informed that they could halt the experiment at any point.

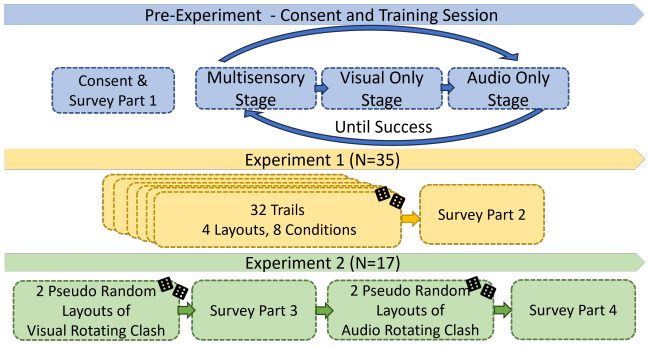

Figure 5: Overall paradigm diagram. Dice symbols denote that these levels were pseudorandomly ordered with a rule avoiding neighboring levels having the same layout or condition.

3.9 Analysis and statistics

Analysis for this experiment was done in python using mainly pandas, numpy, statsmodels, sklearn and scipy. Comparisons between trial durations were performed by averaging each participant's score across layouts so that each participant has a single score per condition, and then testing using a paired t-test the pairs of scores-per-condition on the participant level. Subjective ratings were compared across participants using paired t-tests as well. As success in trials was binary the distributions for success were not normal, so we used the a-parametric signed-rank test for these comparisons. Power calculations were performed with D=0.8. All results are reported as <value>±<standard deviation>. The chance levels for all results were corrected for multiple comparisons via the Bonferroni correction [47].

3.10 Reproducibility

The code with anonymized data is uploaded to OSF in the following link: https://osf.io/jbwd2/, in line with good practice in open and reproducible research.

4 RESULTS

4.1 Unisensory and reliable multisensory navigation

Our first step was to verify that participants could indeed navigate our mazes successfully in both the visual-only and auditory-only conditions as these conditions act as the bar for comparison for the unreliable sensory conditions. Participants indeed navigated the visual mazes with ease (success rate = 100%), completing them with an average time of 17.6±2.7s. The auditory-only stage proved more difficult (success rate = 61.4±23.7%) with an average completion time of 79.9±19.26s. The difference was significant in both success rate ($p<3.11E-11$, signed-rank test) and time ($p<9.67E-20$, paired-t test). (See Figure 6) Subjective wise, the audio-only condition was perceived as significantly more difficult as well ($p<2.84E-25$, paired t-test). For example, participant S22 reported that "As someone who relies on my vision in everyday life, relying on audio only is pretty challenging".

The multisensory condition was at ceiling (success=100%, time=18.3±3.4s) similar to the visual-only condition, and as both were at 100% performance, and with similar perceived difficulty ($p=0.71$) this unfortunately precluded testing for additive multisensory effects. Yet, it's important to note that there were still a portion of participants (28%) that perceived the multi-sensory condition as the easiest.

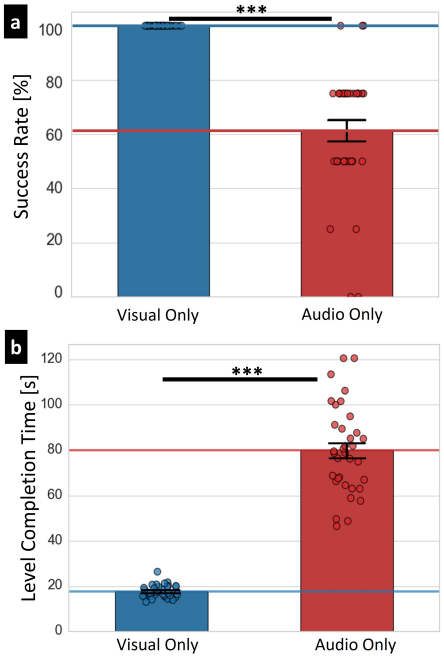

Figure 6: Navigation under reliable unisensory conditions, setting the bar for the rest of the levels. (a) Success level. (b) Time to completion. Each circle denotes the scores of a single participant, error bars are SE.

4.2 Performance during visual clashes

We then turned to testing performance when vision was partially unreliable, including two ghost and two invisible walls. Participants knew that their vision was unreliable but did not know the specific extent. Participants were able to successfully complete these levels with relative ease (success=91.4±17.0%, time=51.4±18.0s). While performance clearly degraded compared to the visual-only condition ($p<0.0005$, paired signed-rank test, $p<1.62E-13$, paired-t test, respectively) it was still perceived as significantly better than audio-only ($p<8.18E-07$, paired-t test) and indeed was significantly better in practice as well ($p<1.38E-05$, signed-rank test, $p<4.90E-11$, paired-t test, respectively). (See Figure 7)

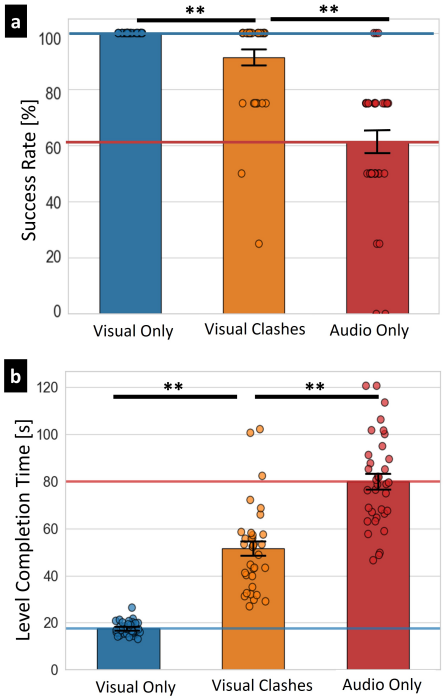

Figure 7: Navigation under the partial visual clash condition (a) Success level. (b) Time to completion. Each circle denotes the scores of a single participant, error bars are SE.

Next we analyzed the condition in which all walls were invisible. Here too we saw degraded performance (success=83.5±17.0%, time=61.54±18.7s) significantly worse then visual-only ($p<4.81E-05$, sign-rank test, $p<4.26E-16$, paired t-test), but also still significantly better than audio-only ($p<6.51E-05$, sign-rank test, $p<4.49E-06$, paired t-test) and with subjective experience matching objective - i.e. difficulty was perceived as harder then visual-only but easier then audio-only. Interestingly, participants reported still actively utilizing their vision and typically did not close their eyes or look away from the screen. (See Figure 8)

We then turned to the full clash condition in which the visual and auditory mazes were completely dissociated, decreasing visual reliability to minimum. Participants were still able to complete the levels, but only at a success rate that was similar to and slightly worse than that of the auditory condition (success=63.5±30.5%, time=77.6±26.3s, $p=0.78$, signed-rank test, $p=0.61$, paired t-test), however they reported that subjectively they found this level easier than audio-only ($p<8.18E-07$, paired t-test) and that they kept their eyes open throughout. (See Figure 9)

When comparing the experiences users reported that they prefer to have false vision then no vision at all. For example, participants reported that "I rely most on vision, even when it's not reliable it still helps me place myself" (S13); "without vision it's hardest because I couldn't see how much I was moving" (S17); "I prefer having a completely wrong visual maze to not seeing a maze at all" (S21); "When there was visual information, even if it was wrong, it felt much easier to me"(S22).

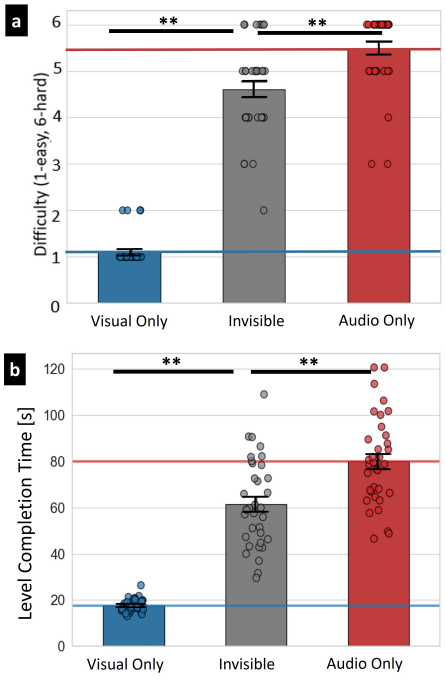

Figure 8: Navigation under the condition in which all walls were invisible (a) Success level. (b) Time to completion. Each circle denotes the scores of a single participant, error bars are SE.

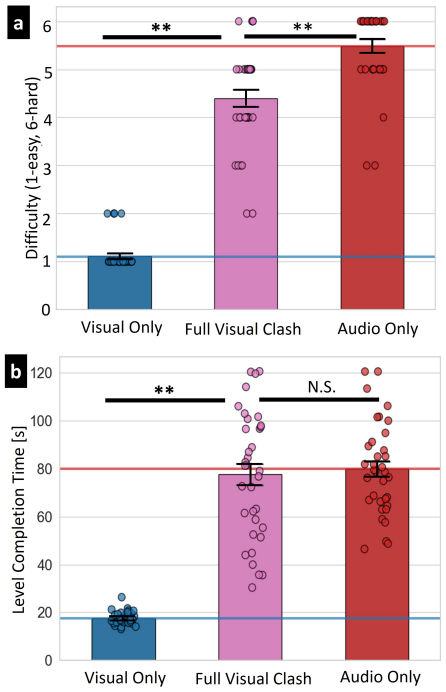

Figure 9: Navigation under the full visual clash condition compared to the two thresholds in terms of (a) subjective and (b) time. Each circle denotes the scores of a single participant, error bars are SE.

4.3 Destabilizing the unreliable visual information

Given these reports by users, we wondered if they would still persist in trying to utilize vision if we removed the ability to use it as landmarks / anchors for memory. To do so we caused the false visual information to be in constant motion (as described above in the methods section under "Experiment 2"). We found that this indeed caused participants to close their eyes or avert their gaze (experimenters noted this for at least 71% of them) and to shift to performing the task based on audio cues only, with performance matching that of the audio-only condition (See Figure 10). Participants reported discomfort and nausea before looking away - e.g. "I felt the need to look away" (S19), "the visual input wasn't relevant at all, therefore I closed my eyes or ignored the screen and once I relied on hearing it was easy" (S35).

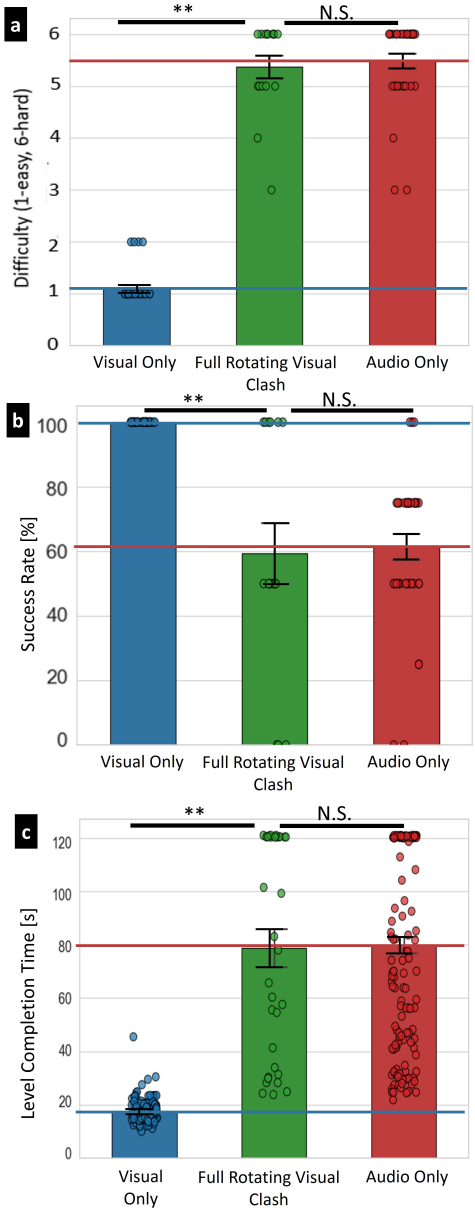

Figure 10: bar graph of control experiment compared to the two thresholds in terms of (a) subjective (b) success, and (c) time. Each circle denotes the scores of a single participant, error bars are SE.

4.4 Did similar effects emerge for false audition?

We next analyzed the effect of false auditory cues in the parallel conditions to the visual ones. Here participants were easily able to solve the mazes with partial audio clashes (success =100%, time = 19.06±4.24s) and full audio clashes (success = 99.28 ± 0.04%, time = 19.32±5.30s), in a manner similar to vision in success (p=1, p=0.317, signed-rank test, respectively), as shown in 11. This demonstrates that participants could easily ignore the auditory clashes. This was reflected also in their feedback - e.g. "I could easily ignore the audio and discard its misleading cues (S35)", "It didn't bother me at all" (S27), "when I knew the audio was misleading it was easy" (S28). On the other hand, in total, there was still a significant difference in perceived difficulty levels under this type of clash ($p<0.0004$, paired t-test) indicating that participants were aware of the auditory cues.

4.5 Learning effect

As participants perform 8 repetitions of each of the 4 layouts, there is a potential for learning effects throughout the experiment. To neutralize this, the condition order was randomized across participants, and thus learning effects should not bias our results. To test for learning effects we looked at each layout and compared the first half of the trials (first 4) each participant performed in it to the second half (final 4). We did not find learning effects for time (p=0.5, paired t-test) and only a trend for success (p=0.06, sign-rank test) indicating that failed trials were trending to be more common early in the experiment. However, given the randomization of condition order across participants this does not affect the rest of the results presented here.

4.6 Did participants close their eyes?

If the visual information is not useful, we would expect participants to simply close their eyes or look away to inhibit it. During the visual clash levels, only 28% reported that they closed their eyes or looked away, a result supported by the reports of the experimenters viewing them. In experiment 2 however, 71% participants reported closing their eyes or looking away. Anecdotally, the experimenters report that participants tended to look away or close their eyes during the audio-only condition as well though we did not explicitly track this.

4.7 How did it feel to face the clashes?

Participants reported that ghost walls tended to be odd at first, but that they soon adapted to them - "at first it was weird, then I adapted" (S14); "Odd at first" (S17); "It was weird the first couple of times" (S18). Some even used fearful terminology for their initial experience - "I learned to put aside the fear" (S4); "At first it was scary but as we progressed through the levels it was less intimidating" (S19). More broadly, most participants described ghost walls as either "weird" or "confusing".

In contrast, the common description theme for invisible walls was "frustrating" or "surprising" - "At first it was surprising" (S19); "It kept surprising me throughout" (S14); "They were frustrating" (S23); "You just don't expect to hear the beeps" (S26); "It was surprising" (S27). Some users noted that this feeling was mitigated by getting the advance warning from the auditory cues "but at least I could prepare for it in advance" (S2).

Walking through auditory clash walls were reported as easy to ignore - "I just ignored it" (S2); "Didn't bother me" (S3) "I could see so i didn't really notice it" (S8); "I felt it was easier to ignore the sounds when i could rely on vision" (S19). We did not see a difference in description between mute walls and auditory ghost walls.

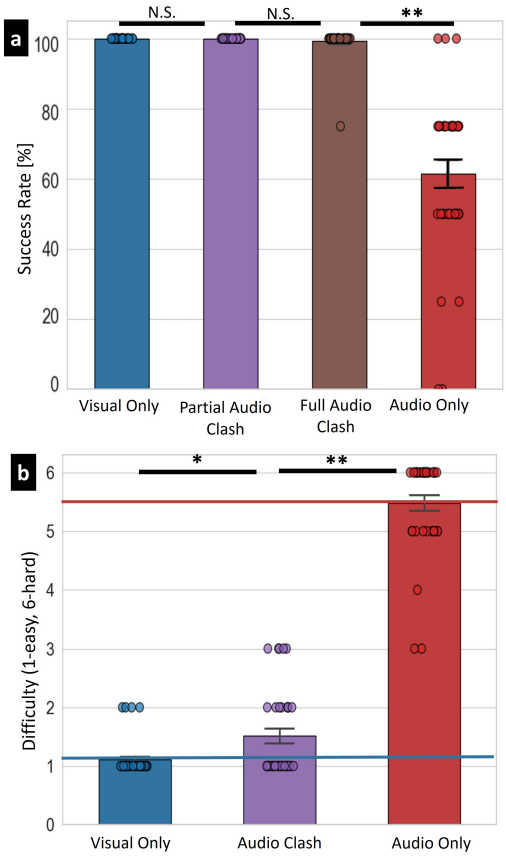

Figure 11: bar graph of auditory clashes compared to the two thresholds in terms of (a) success and (b) subjective.Each circle denotes the scores of a single participant, error bars are SE.

5 Discussion

Our results showed that participants could navigate the different conditions with different levels of success, time and subjective preference between the bars set by the visual-only and audio-only conditions. The results for the partial visual clash and the invisible levels support an **enhancing** model, showing that participants were able to make use of the visual information despite its degraded reliability. Further, while full visual clash performance matched audio-only, they perceived these levels as easier and reported relying on vision despite it being false. In experiment 2 we showed that a large enough visual clash will cause participants to abandon vision, closing their eyes and showing the same level of difficulty as compared to the audio condition matching a **suppression** model - although a combination of potential nausea/dizziness factor could be involved in some participants. Finally, these effects did not show up in auditory clashes, which matched the prediction of the **suppression** model. These results demonstrate the way humans cling to vision, insisting on relying upon it even when it is completely false and does not improve their performance, since they can wring spatial information such as landmarks and optic flow from it. Only when these are removed will participants shift from an enhancing model to a suppressing model.

5.1 Why are humans so attached to false vision? And how do we utilize it for enhanced performance?

Beyond its dominant role among our sensory channels, vision offers the ability to anchor our surroundings and act as a scaffolding for

other sensory information. Thus, a false wall may act as a comforting landmark reminding us which places we have already tried and which we haven't. An alternative option is the use of false vision for optic flow, to generate a sensation of moving within the virtual environments which is critical given the lack of idiothetic motion sensing. These two aspects are difficult to disentangle, as a scenario which disrupts one will typically disrupt the other. Future work may employ a paradigm with walls or other visual features randomly appearing for brief periods of time and vanishing, offering optic flow without the ability to use them as a landmark, versus an environment which moves with the user, offering landmarks without optic flow. A third option is more subjective - we may rely on false vision due to the dominant nature of our vision in navigation and our habit of relying on it [8, 35] - not using it does not seem natural to us, even though it is false. This option is reflected well in the full clash and participants' subjective quotes vs. their objective performance. Finally, a potential confound in our experiment 2 is that it induced feelings of dizziness and nausea - these are coupled strongly to the reliance on vision and were stated by many participants as the reason they looked away. This nausea might be induced exactly because of the loss of these visual properties, making it another aspect which is hard to disentangle and it may underlie and accompany this effect inherently.

5.2 Types of clashes

Another interesting aspect in facing visual clashes is the difference between facing invisible (false negative) and ghost walls (false positive). Participants reported greater difficulty with ghost walls. This may be due to experience - the greater prevalence of real life scenarios in which we must add a boundary to our spatial representation of an environment in comparison to the much rarer case of suppressing a visible wall. Another possible cause is the underlying spatial representation - it may be easier for us to enhance this representation with new features that might be missing in a sensory channel, while suppressing perceived features and subtracting them from the model may be more difficult.

Interestingly, the two types of clashes evoked a different emotional reaction. The experience of walking through a ghost wall was described by participants as not natural and even scary, and they reported reluctance to walk through it the first times, a reaction which decreased as they gained experience. In contrast, invisible walls evoked frustration - participants thought they could do something, and discovering that they couldn't was a frustrating surprise. Several participants explicitly noted that they did not get used to this and this reaction persevered also in the end of the experiment.

These reactions have several practical implications for XR design - they suggest that when the user first experiences walls in an XR experience they will by default treat them as real and avoid going through them. However if your experience breaks these illusions, they will have much less inhibition walking through future walls you render for them. The feeling of frustration evoked by completely invisible walls should also be noted as it can color the user's experience. This might be mitigated by giving the user an earlier warning that they are approaching such an invisible boundary (e.g. fading in a guardian wall [27] rather than snapping it up).

5.3 Is this effect unique to vision?

The conditions in which the auditory information was clashing did not lead to these effects. Participants easily rejected the false auditory inputs, no matter how much we increased the clash and despite relying on this same auditory information to solve neighboring conditions. However, we must note that a large part of the auditory information used in this experiment was not natural spatial audio but rather a sonification of distance which we do not typically use [22]. This may have made the auditory information easier to suppress, but note on the other hand that participants were easily able to suppress also the sounds of collisions with walls which are more naturalistic and spatial. Alternatively, this could be because of the dominance of vision over other sensory channels for sensory information [8]. Future work should test audition vs. other sensory channels such as haptics and vestibular to determine if this suppression ability is an effect of vision dominating or of the specific mapping.

5.4 Comparison to blind navigation patterns

When humans navigate without vision they tend to display certain movement patterns - for example, thigmotaxis, in which humans stay close to the surface of a guiding wall and avoid open spaces [42]. Our environment was a series of corridors which limits the ability to test how these patterns were adopted by our participants, but previous work with the EyeCane sonification approach has shown such that the ability for distal perception enables participants to leave the shelter of the walls in favor of the middle of corridors [22].

5.5 Comparison to other sonification tools

We chose to implement the virtual EyeCane approach due to its simplicity, the speed in which users master it [22], and the similarity in concept to the audio cues modern cars use to indicate obstacles while driving in reverse. This approach works well for simple environments such as corridors, but is less suitable for complex and rich environments where more nuanced tools such as [36] would work better. Indeed, if the scenario requires understanding the content of the scene rather than just its shape, the user would benefit more from tools utilizing image recognition [13], or sensory substitution of the whole scene [21]

5.6 Applications

Our main motivation for this research is approaching the basic science of multisensory integration from a new direction, enabling us to better understand who our senses interact in general and in virtual environments in particular. Understanding multisensory integration is critical for designing proper experiences in virtual and mixed reality, enhancing presence [14, 38] and decreasing nausea and discomfort [33]. However, multisensory clashes hold a wide application space in virtual and especially in mixed reality as well, as in these environments they occur "naturally" much more frequently then in the real world. This is especially true for cases in which we see something, be it an interface element or game object, that has no tangible physical presence, and in cases where the virtual additions obscure real world obstacles and indeed might even make the user believe the space in front of them is clear. Such clashes are already commonplace in gaming, exercise and productivity software, and as these tools become more ubiquitous will only grow, especially in the context of mixed reality interfaces and objects.

6 Limitations

6.1 Does virtual navigation on a desktop simulate real navigation?

To avoid contamination from other sensory channels (e.g. vestibular), and to provide compatibility with parallel neuroimaging in planned future stages of the study, we utilized desktop based virtual environments. This is obviously far from the rich naturalistic multisensory experience of navigating in the real world, and even far from what is currently offered by modern HMDs in which the idiothetic cues merge with the visual to lead to increased immersion. Future work will need to address this gap both in terms of testing these idiothetic cues, and of testing sensory clashes not one-vs-one but rather all-vs-one. We predict that performance of idiothetic + auditory will be much closer to the performance level offered by vision, though still not as powerful.

6.2 How do our results fit with models of multisensory integration?

Understanding visual, and other types of sensory clashes is crucial in several approaches for studying multi-sensory integration, which investigates how sensory inputs are integrated in our brain. Methods such as Bayesian inference and Maximum Likelihood Estimation [39] rely on modeling sensory conflicts as a form of edge case testing [9, 17]. Therefore, gaining a comprehensive understanding of these scenarios is important for advancing our knowledge of multi-sensory integration. However these theories typically focus on millisecond incongruency resolution and on edge cases. There is a need to extend them to longer periods of time, which our paradigm offers a way of testing in the future with additional levels of clashes to extend these models to sustained naturalistic scenarios.

6.3 Eye tracking

In our experiment, we tracked whether participants closed their eyes using documentation from the experimenter who observed the participants, along with their self-reporting. While the resolution of a single answer per trial was enough to answer the questions we posed here, the finer time and spatial resolution offered by eye tracking holds a great potential for future work in this field. Utilizing eye trackers would provide a more objective method to track when and where participants were looking, and see how this varied within trials and which parts of the virtual scene they were using - for example, were they latching onto specific visual features or screen locations as landmarks? Did they alternate opening and closing their eyes or did they keep looking away once they shut them?

6.4 Disentangling nausea from fully unreliable vision

In Experiment 2, we reported that 71% of the participants closed their eyes when exposed to false visual information in constant motion (3.3). This behavior could be influenced by two main factors: willingly abandoning visual feedback due to an inability to trust it, or experiencing dizziness at varying levels. Distinguishing between these factors is challenging since nausea and dizziness could inherently result from exposure to misleading visuals, serving as a mechanism leading to suppression. Therefore, assessing the relative influence of each factor is challenging. Future work may consider employing the paradigm mentioned in 5.1, involving randomly appearing walls, to further explore the relationship between them.

7 Conclusion and Future Work

Our results demonstrate how attached humans are to vision - including insisting on relying upon it even when they actively know that it is false and does not contribute to their behavioral performance supporting a role predicted by an enhancing model. We also found that degrading the false visual information's ability to serve as an indicator of optic flow or as landmarks breaks this effect and shifts performance to be in line with a suppression model, though this is confounded by accompanying nausea in some participants. Future work will extend this paradigm to additional senses, testing haptics via dedicated interfaces and idiothetic cues via 6DOF HMDs and focus both on the basic science of multisensory integration, and on the practical implications for XR design of these interactions.

Acknowledgments

This research was supported by ISF grant No. 1322/22. We would like to thank our participants for participating in this experiment.

References

[1] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *Josa a*, 20(7):1391–1397, 2003.

[2] A. Berthoz and I. Viaud-Delmon. Multisensory integration in spatial orientation. *Current opinion in neurobiology*, 9(6):708–712, 1999.

[3] C. J. Bohil, B. Alicea, and F. A. Biocca. Virtual reality in neuroscience research and therapy. *Nature reviews neuroscience*, 12(12):752–762, 2011.

[4] R. Byrne, J. Marshall, and F. F. Mueller. Ar fighter: Using hmds to create vertigo play experiences. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pp. 45–57, 2018.

[5] D.-R. Chebat, S. Maidenbaum, and A. Amedi. Navigation using sensory substitution in real and virtual mazes. *PloS one*, 10(6):e0126307, 2015.

[6] N. Diersch and T. Wolbers. The potential of virtual reality for spatial navigation research across the adult lifespan. *Journal of Experimental Biology*, 222(Suppl_1):jeb187252, 2019.

[7] N. Dozio, E. Maggioni, D. Pittera, A. Gallace, and M. Obrist. May i smell your attention: Exploration of smell and sound for visuospatial attention in virtual reality. *Frontiers in psychology*, 12:671470, 2021.

[8] A. D. Ekstrom. Why vision is important to how we navigate. *Hippocampus*, 25(6):731–735, 2015.

[9] M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

[10] D. J. Finnegan, E. O'Neill, and M. J. Proulx. Compensating for distance compression in audiovisual virtual environments using incongruence. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 200–212, 2016.

[11] P. Gao, K. Matsumoto, T. Narumi, and M. Hirose. Visual-auditory redirection: Multimodal integration of incongruent visual and auditory cues for redirected walking. In *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 639–648. IEEE, 2020.

[12] L. Hejtmanek, M. Starrett, E. Ferrer, and A. D. Ekstrom. How much of what we learn in virtual reality transfers to real-world navigation? *Multisensory Research*, 33(4-5):479–503, 2020.

[13] B. D. Jain, S. M. Thakur, and K. Suresh. Visual assistance for blind using image processing. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0499–0503. IEEE, 2018.

[14] H. Kim and I.-K. Lee. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2080–2090, 2022.

[15] R. K. Kiryakova, S. Aston, U. R. Beierholm, and M. Nardini. Bayesian transfer in a complex spatial localization task. *Journal of Vision*, 20(6):17–17, 2020.

[16] A. J. Kolarik, A. C. Scarfe, B. C. Moore, and S. Pardhan. Blindness enhances auditory obstacle circumvention: Assessing echolocation, sensory substitution, and visual-based navigation. *PloS one*, 12(4):e0175750, 2017.

[17] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams. Causal inference in multisensory perception. *PLoS one*, 2(9):e943, 2007.

[18] O. Lahav, H. Gedalevitz, S. Battersby, D. Brown, L. Evett, and P. Merritt. Virtual environment navigation with look-around mode to explore new real spaces by people who are blind. *Disability and rehabilitation*, 40(9):1072–1084, 2018.

[19] M. Li, S. Sareh, G. Xu, M. B. Ridzuan, S. Luo, J. Xie, H. Wurdemann, and K. Althoefer. Evaluation of pseudo-haptic interactions with soft objects in virtual environments. *PLoS One*, 11(6):e0157681, 2016.

[20] J. Lin, J. Cronjé, I. Käthner, P. Pauli, and M. E. Latoschik. Measuring interpersonal trust towards virtual humans with a virtual maze paradigm. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2401–2411, 2023.

[21] S. Maidenbaum, G. Buchs, S. Abboud, O. Lavi-Rotbain, and A. Amedi. Perception of graphical virtual environments by blind users via sensory substitution. *PloS one*, 11(2):e0147501, 2016.

[22] S. Maidenbaum, S. Hanassy, S. Abboud, G. Buchs, D.-R. Chebat, S. Levy-Tzedek, and A. Amedi. The "eyecane", a new electronic travel aid for the blind: Technology, behavior & swift learning. *Restorative neurology and neuroscience*, 32(6):813–824, 2014.

[23] S. Maidenbaum, S. Levy-Tzedek, D.-R. Chebat, and A. Amedi. Increasing accessibility to the blind of virtual environments, using a virtual mobility aid based on the" eyecane": Feasibility study. *PloS one*, 8(8):e72555, 2013.

[24] S. Maidenbaum, S. Levy-Tzedek, D. R. Chebat, R. Namer-Furstenberg, and A. Amedi. The effect of extended sensory range via the eyecane sensory substitution device on the characteristics of visionless virtual navigation. *Multisensory research*, 27(5-6):379–397, 2014.

[25] S. Maidenbaum, A. Patel, T. Gedankien, and J. Jacobs. The effect of navigational aids on spatial memory in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 644–645. IEEE, 2020.

[26] J. Marshall, S. Benford, R. Byrne, and P. Tennent. Sensory alignment in immersive entertainment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

[27] Meta. Guardian system: Oculus developers, 2022.

[28] M. Nardini. Merging familiar and new senses to perceive and act in space. *Cognitive Processing*, 22(Suppl 1):69–75, 2021.

[29] T. Narumi. Multi-sensorial virtual reality and augmented human food interaction. In *Proceedings of the 1st workshop on multi-sensorial approaches to human-food interaction*, pp. 1–6, 2016.

[30] J. Negen, L.-A. Bird, H. Slater, L. Thaler, and M. Nardini. A new sensory skill shows automaticity and integration features in multisensory interactions. *BioRxiv*, pp. 2021–01, 2021.

[31] J. Negen, L. Wen, L. Thaler, and M. Nardini. Bayes-like integration of a new sensory skill with vision. *Scientific Reports*, 8(1):1–12, 2018.

[32] P. A. Neil, C. Chee-Ruiter, C. Scheier, D. J. Lewkowicz, and S. Shimojo. Development of multisensory spatial integration and perception in humans. *Developmental science*, 9(5):454–464, 2006.

[33] C. M. Oman. Sensory conflict in motion sickness: an observer theory approach. *Pictorial communication in virtual and real environments*, pp. 362–376, 1991.

[34] T. U. Otto and P. Mamassian. Multisensory decisions: the test of a race model, its logic, and power. *Multisensory Research*, 30(1):1–24, 2017.

[35] A. Pasqualotto and T. Esenkaya. Sensory substitution: the spatial updating of auditory scenes "mimics" the spatial updating of visual scenes. *Frontiers in Behavioral Neuroscience*, 10:79, 2016.

[36] G. Presti, D. Ahmetovic, M. Ducci, C. Bernareggi, L. Ludovico, A. Baratè, F. Avanzini, and S. Mascetti. Watchout: Obstacle sonification for people with visual impairment or blindness. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 402–413, 2019.

[37] D. H. Raab. Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, 1962.

[38] N. Ranasinghe, P. Jain, N. Thi Ngoc Tram, K. C. R. Koh, D. Tolley, S. Karwita, L. Lien-Ya, Y. Liangkun, K. Shamaiah, C. Eason Wai Tung, et al. Season traveller: Multisensory narration for enhancing the virtual reality experience. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018.

[39] N. W. Roach, J. Heron, and P. V. McGraw. Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B: biological sciences*, 273(1598):2159–2168, 2006.

[40] D. Russell. *Implementing Augmented Reality Into Immersive Virtual Learning Environments*. IGI Global, 2020.

[41] J. Sánchez, M. Sáenz, A. Pascual-Leone, and L. Merabet. Navigation for the blind through audio-based virtual environments. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pp. 3409–3414. 2010.

[42] V. R. Schinazi, T. Thrash, and D.-R. Chebat. Spatial navigation by congenitally blind individuals. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(1):37–58, 2016.

[43] L. Shams. Early integration and bayesian causal inference in multisensory perception. 2012.

[44] B. E. Stein, T. R. Stanford, and B. A. Rowland. Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience*, 15(8):520–535, 2014.

[45] P. Tennent, J. Marshall, B. Walker, P. Brundell, and S. Benford. The challenges of visual-kinaesthetic experience. In *Proceedings of the 2017 conference on designing interactive systems*, pp. 1265–1276, 2017.

[46] F. Weidner, J. E. Maier, and W. Broll. Eating, smelling, and seeing: Investigating multisensory integration and (in) congruent stimuli while eating in vr. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2423–2433, 2023.

[47] E. W. Weisstein. Bonferroni correction. *https://mathworld. wolfram. com/*, 2004.

[48] D. Wolf, M. Rietzler, L. Hnatek, and E. Rukzio. Face/on: Multi-modal haptic feedback for head-mounted displays in virtual reality. *IEEE transactions on visualization and computer graphics*, 25(11):3169–3177, 2019.