The Effect of Audio-Visual Temporal and Spatial Consistency on the Pip-and-Pop Effect

Authors: Tang Xiaoyu, Cui Xinzhong, high sensitivity, Yuan Mengying, Gao Min

Date: 2022-06-14T00:00:00+00:00

Abstract

Multisensory integration follows spatial and temporal principles. Research has shown that the Pip-and-Pop effect arises from multisensory integration; does the Pip-and-Pop effect likewise adhere to spatial and temporal principles? The present study employed a dynamic visual search paradigm and investigated the influence of spatial congruency (Experiment 1) and temporal congruency (Experiment 2) on the Pip-and-Pop effect through two eye-tracking experiments. The results revealed: (1) The Pip-and-Pop effect was maximal when the visual target color change was accompanied by an ipsilateral auditory stimulus; no Pip-and-Pop effect was observed in the contralateral condition. (2) The Pip-and-Pop effect was maximal when auditory and visual stimuli were presented simultaneously, and the effect gradually diminished and eventually disappeared as the temporal interval between audiovisual stimuli increased. These findings demonstrate that audiovisual temporal and spatial congruency modulates the Pip-and-Pop effect, providing evidence for the role of multisensory integration in generating the Pip-and-Pop effect.

Full Text

The Impact of Audiovisual Spatial and Temporal Consistency on the Pip-and-Pop Effect

Tang Xiaoyu, Cui Xinzhong, Gao Min, Yuan Mengying
(School of Psychology, Liaoning Normal University; Liaoning Collaborative Innovation Center for Assessment and Cultivation of Healthy Personality in Children and Adolescents, Dalian 116029)

Abstract

Multisensory integration follows spatial and temporal principles. Since research suggests that the Pip-and-Pop effect arises from multisensory integration, does this effect also adhere to these same principles? Using a dynamic visual search paradigm, we conducted two eye-tracking experiments to examine the influence of spatial consistency (Experiment 1) and temporal consistency (Experiment 2) on the Pip-and-Pop effect. Results revealed: (1) The Pip-and-Pop effect was maximal when visual target color changes were accompanied by an ipsilateral auditory stimulus, while no effect was observed in the contralateral condition. (2) The effect was strongest when auditory and visual stimuli were presented simultaneously, gradually diminishing and eventually disappearing as the interval between stimuli increased. These findings demonstrate that audiovisual spatial and temporal consistency modulates the Pip-and-Pop effect, providing evidence for the role of multisensory integration in generating this phenomenon.

Keywords: Pip-and-Pop effect, audiovisual spatial consistency, audiovisual temporal consistency, dynamic visual search paradigm, eye movements

Introduction

In daily life, we constantly receive vast amounts of information from different sensory channels. Searching for objects in complex, dynamic environments is a time-consuming process. For instance, locating a friend in a crowd after becoming separated at a tourist attraction requires considerable time. However, if the friend waves while calling your name, search time is dramatically reduced. This scenario exemplifies a dynamic visual search process accompanied by auditory stimulation. Van der Burg et al. (2008) pioneered research on this phenomenon using a dynamic visual search paradigm in which participants searched for vertical or horizontal target line segments among distractors of various orientations while the segments continuously changed color between red and green. They discovered that presenting a pure tone simultaneously with target color changes improved search efficiency, a phenomenon termed the "Pip-and-Pop" effect.

The Pip-and-Pop effect has been attributed to multisensory integration (Chamberland et al., 2016; Van der Burg et al., 2008; Van der Burg et al., 2011). Multisensory integration (MSI) refers to the process by which information from different sensory modalities interacts and combines into a unified, coherent, and meaningful percept (Talsma et al., 2010; Tang et al., 2016; Van der Stoep et al., 2017). Van der Burg and colleagues found that synchronous auditory stimuli and visual target color changes produced early multisensory integration in the parieto-occipital region within 50 ms, which correlated significantly with the behavioral Pip-and-Pop effect—greater early integration corresponded to larger improvements in search efficiency. They proposed that early integration binds auditory and visual signals, making the visual target salient among continuously changing stimuli, thereby capturing attention more effectively and significantly

enhancing search efficiency compared to visual-only conditions. Thus, multisensory integration plays a crucial role in the Pip-and-Pop effect (Van der Burg et al., 2008; Van der Burg et al., 2011).

However, alternative explanations have emerged. Some researchers argue that audiovisual integration alone cannot fully explain the Pip-and-Pop effect, as auditory stimuli improve search efficiency even when no visual target is present. They propose that oddball stimuli (salient, low-probability stimuli appearing among high-probability stimuli) attract attention and cause the effect (Ngo & Spence, 2012; Tsai & Yeh, 2013; Zou et al., 2012). In visual search, an abruptly presented auditory stimulus constitutes an oddball stimulus, which more effectively captures attention (Chastain & Cheal, 1999; Escera et al., 1998; Escera et al., 2002). Research has shown that when visual targets and auditory oddball stimuli are presented simultaneously, subjective perceived time expands —a phenomenon called the "freezing effect" (Vroomen & de Gelder, 2000; Tse et al., 2004). Oddball stimuli across modalities can produce this effect (Ngo & Spence, 2010b; Ngo & Spence, 2012).

The debate continues regarding whether multisensory integration or oddball-driven attentional capture causes the Pip-and-Pop effect. Previous research indicates that multisensory integration follows two key principles: the spatial rule and the temporal rule. The spatial rule states that integration is maximal when information from different modalities originates from approximately the same spatial location (Spence, 2013; Van der Stoep et al., 2017). The temporal rule holds that integration is strongest when information from different modalities occurs at approximately the same time (Fister et al., 2016; Stevenson et al., 2012). Consequently, researchers have manipulated spatial and temporal consistency to investigate the underlying mechanism of the Pip-and-Pop effect (Van der Burg et al., 2008; Ngo & Spence, 2010a; Zou et al., 2012). If the effect follows these principles, multisensory integration likely contributes to its generation. If not, oddball-driven attentional capture may be the primary mechanism.

However, previous findings on spatial consistency have been inconsistent. Ngo and Spence (2010) found that auditory stimuli providing valid spatial information improved visual search efficiency compared to invalid cues, concluding that spatial consistency influenced the Pip-and-Pop effect and that greater spatial alignment facilitated stronger audiovisual integration (Ngo & Spence, 2010a). In contrast, Fleming et al. (2020) found no significant differences in behavioral or electrophysiological measures between spatially aligned and misaligned conditions, suggesting that spatial consistency is not a necessary condition for the Pip-and-Pop effect and that audiovisual integration does not play a substantial role. These discrepancies may stem from differences in spatial arrangement. Ngo and Spence (2010) randomly positioned stimuli on a $10\$\times10 invisible grid, while Fleming et al. (2020) reduced the search area, increased distractor orientations, and in
grids on each side for set size 36), allowing more explicit manipulation of spatial consistency.

Similarly, research on temporal consistency has yielded mixed results. Van der

Burg et al. (2008) used a dynamic visual search paradigm with eight Tone-Target-Interval (TTI) conditions and found that auditory stimuli facilitated visual search at TTIs of -100 ms, -50 ms, -25 ms, 0 ms, 25 ms, and 100 ms (where 0 ms indicates simultaneous presentation, negative values indicate auditory lead, and positive values indicate auditory lag). The facilitative effect was maximal at 0 ms, with a U-shaped relationship between TTI and reaction time—greater temporal consistency yielded faster search. They argued that synchronous auditory stimuli and those presented within the integration time window (-100 ms to 100 ms) enhanced visual target salience through crossmodal integration (Van der Burg et al., 2008). Conversely, Zou et al. (2012) compared 0 ms, -100 ms, and random auditory timing conditions, finding that all three facilitated visual search without significant differences between them. They concluded that the Pip-and-Pop effect resulted from a "freezing effect" caused by spatially uninformative auditory oddball stimuli, which captured attention regardless of timing rather than through audiovisual integration (Zou et al., 2012).

These studies used different set sizes, which may explain the inconsistent findings. Effects of temporal consistency were observed with set sizes of 48 and 60 (Van der Burg et al., 2008; Kösem & van Wassenhove, 2012) but not with set size 36 (Zou et al., 2012), suggesting that set size may moderate the influence of temporal consistency on the Pip-and-Pop effect. Similarly, set size may moderate spatial consistency effects (Ngo & Spence, 2010a; Fleming et al., 2020). With smaller set sizes and lower perceptual load, auditory stimuli may improve search efficiency regardless of spatial or temporal alignment. To clarify this issue, the present study introduced set size as a variable to investigate how spatial and temporal consistency affect the Pip-and-Pop effect under different perceptual loads using both behavioral and eye-tracking measures.

While behavioral reaction times can indicate whether auditory stimuli improve search efficiency, eye-tracking data reveal real-time information processing through saccades and fixations. Specifically, mean fixation number reflects perceptual load (more fixations indicate greater load and lower efficiency), mean saccade amplitude reflects the amount of information acquired before each saccade (larger amplitudes indicate higher efficiency), and mean fixation duration can quantify freezing effects to determine whether oddball-driven attentional capture contributes to the Pip-and-Pop effect. By applying eye-tracking technology to the dynamic visual search paradigm, the present study addressed two questions: (1) How does audiovisual spatial consistency affect the Pip-and-Pop effect when manipulating set size and auditory stimulus location? (2) How does audiovisual temporal consistency affect the effect when manipulating set size and the interval between auditory and visual stimuli? If multisensory integration underlies the Pip-and-Pop effect, it should follow both spatial and temporal principles—greater spatial and temporal consistency should produce larger effects. If oddball-driven attentional capture is the mechanism, abrupt auditory stimuli should affect visual search regardless of spatial or temporal alignment.

Experiment 1: The Influence of Audiovisual Spatial Consistency

Method

Participants We used G*Power toolbox 3.1.9.2 to calculate the required minimum sample size (Erdfelder et al., 2009; Faul et al., 2007). Based on previous research on auditory facilitation of visual search (Ngo & Spence, 2012), the effect size for audiovisual spatial consistency in repeated-measures ANOVA was $p^2 = 0.43$. Setting the Type I error probability (α) at 0.05 and statistical power (1-β) at 0.9, the minimum sample size was calculated as 24. To account for potential data loss from eye-tracking artifacts, we recruited 32 university students. Eight participants were excluded due to excessive blinking causing data loss, resulting in a final sample of 24 participants (5 males, 19 females) aged 19-28 years (M = 22.5, SD = 2.3). All participants were right-handed with normal hearing and normal or corrected-to-normal vision, and no color blindness. Participants received compensation upon completion.

Apparatus We recorded participants' right eye movements using an Eyelink 1000 Plus infrared eye-tracking system (SR Research, Canada) at a sampling rate of 1000 Hz. Visual stimuli were presented on a Dell P1914SF LCD monitor (19-inch viewable area, 1024$×$768 pixel resolution, 75 Hz refresh rate). Participants were seated 75 cm from the screen. Experimental procedures were programmed using Matlab 2016a with Psychtoolbox and Eyelink Toolbox.

Materials Visual stimuli consisted of multiple distractor line segments (oriented at 22.5°, 45°, 67.5°, 112.5°, 135°, and 157.5°) and one target segment (horizontal or vertical, counterbalanced across trials). All segments were either red (13.9 cd/m^2) or green (46.4 cd/m^2) presented on a black background (0.4 cd/m^2). Each segment measured 0.57° × 0.17°. Three set sizes were used: 36, 48, and 60 (Figure 1 [Figure 1: see original paper] illustrates set size 36). Visual stimuli were divided into two equal matrices presented on the left and right sides of the screen (for set size 36: 3.4° × 8.5° per side; set size 48: 5.1° × 8.5°; set size 60: 6.8° × 8.5°). The horizontal distance from the central fixation point to each side was 3.74°.

During the experiment, segments continuously changed color between red and green, with varying numbers changing on each cycle (set size 36: 1, 3, or 5 segments; set size 48: 1, 4, or 7 segments; set size 60: 1, 5, or 9 segments). Color change intervals were 50, 100, or 150 ms, randomly intermixed across nine intervals (one cycle) to prevent predictable patterns. Target color changes followed three rules: (1) The target changed color alone, only once per cycle, at an average frequency of 1.11 Hz (every 900 ms); (2) The interval before target color change was 150 ms, and after was 100 ms, ensuring clear audiovisual

integration within the ~$±$100 ms integration window; (3) The target never changed during the first three intervals of each cycle.

The auditory stimulus was a 1000 Hz pure tone (65 dB, 60 ms including 5 ms fade-in and 5 ms fade-out) presented through speakers located behind the screen on both sides. Auditory presentation had four conditions (see Figure 1): (1) No sound: no auditory stimulus; (2) Ipsilateral (valid spatial information): auditory stimulus presented on the same side as the visual target (right channel for right-side targets, left channel for left-side targets); (3) Contralateral (invalid spatial information): auditory stimulus presented opposite the visual target; (4) Bilateral (no spatial information): auditory stimulus presented through both channels regardless of target location, creating a central percept.

Procedure Experiment 1 used a 3 (set size: 36, 48, 60) × 4 (spatial consistency: no sound, ipsilateral, contralateral, bilateral) within-subjects design. Before the formal experiment, participants completed an auditory localization test to ensure they could discriminate sound direction. To minimize head movements and obtain accurate eye-tracking data, participants rested their forehead against a forehead rest and chin on a chinrest, then completed a nine-point calibration. Each trial began with a drift calibration point that remained until fixated, followed by a central fixation point for 1000 ms, then the search display. Participants searched for the target line segment and indicated whether it was vertical or horizontal by pressing keys (Z or M, counterbalanced across participants). If no response occurred after nine intervals, the stimulus cycle repeated until a response was made. Participants were instructed to ignore auditory stimuli.

The formal experiment comprised six blocks (two per set size), each containing 96 trials with randomly intermixed spatial consistency conditions, totaling 576 trials. Each block began with 20 practice trials. Participants could rest between blocks, and nine-point calibration was repeated before each block. The entire experiment lasted approximately 90 minutes.

Data Analysis Accuracy was defined as the proportion of correct trials. Reaction time (RT) was the interval between search display onset and participant response. Search efficiency reflected response speed—longer RTs indicated slower, less efficient search; shorter RTs indicated faster, more efficient search.

To investigate how auditory stimuli specifically affected search efficiency, we selected three eye-tracking measures based on previous research (Zou et al., 2012): (1) Mean fixation number: the number of fixations within the current interest area, reflecting perceptual load (more fixations indicate greater load and lower efficiency); (2) Mean saccade amplitude: the distance between successive fixations, with larger amplitudes indicating more information acquired per saccade and higher efficiency; (3) Mean fixation duration: the average duration of all fixations within the interest area, reflecting overall processing. Zou et al. (2012)

used this measure to quantify freezing effects, assessing whether auditory stimuli prolonged fixation durations.

We conducted 3 (set size) × 4 (spatial consistency) repeated-measures ANOVAs on all measures to examine the effects on the Pip-and-Pop effect.

Results

We excluded trials with incorrect responses, anticipatory responses (responses before the first target change), and excessively slow responses (no response after 10 target changes), which accounted for 5.5% of all trials.

Accuracy All participants achieved accuracy above 98%. ANOVA revealed no significant main effect of set size, $F(2, 46) = 0.74$, $p = 0.48$; no main effect of spatial consistency, $F(3, 69) = 0.25$, $p = 0.86$; and no significant interaction, $F(6, 138) = 0.66$, $p = 0.68$. Accuracy was high across all conditions.

Reaction Time ANOVA revealed a significant main effect of set size, $F(2, 46) = 5.34$, $p = 0.008$, $p^2 = 0.19$. RTs were significantly longer for set size 60 (3.5 s) than for set size 48 (3.3 s, $p = 0.03$) and set size 36 (3.2 s, $p = 0.004$), with no significant difference between the latter two ($p = 0.47$). The main effect of spatial consistency was also significant, $F(3, 69) = 11.86$, $p < 0.001$, $p^2 = 0.34$. RTs were shortest in the ipsilateral condition (3.07 s), significantly shorter than in the bilateral (3.27 s, $p = 0.006$), contralateral (3.50 s, $p < 0.001$), and no sound (3.43 s, $p < 0.001$) conditions. Bilateral RTs were also significantly shorter than contralateral ($p = 0.005$) and no sound ($p = 0.038$) RTs. Contralateral RTs were longest and did not differ significantly from no sound ($p = 0.35$). The set size × spatial consistency interaction was not significant, $F(6, 138) = 0.46$, $p = 0.84$ (see Figure 2a [Figure 2: see original paper]).

Eye Movement Data **Mean Fixation Number**: ANOVA revealed a significant main effect of set size, $F(2, 46) = 3.44$, $p = 0.04$, $p^2 = 0.13$. Fixations were significantly more numerous for set size 60 (14.20) than for set size 36 (13.11, $p = 0.01$), with set size 48 (13.57) not differing significantly from either (ps = 0.183 and 0.261). The main effect of spatial consistency was significant, $F(3, 69) = 7.45$, $p < 0.001$, $p^2 = 0.25$. Fixations were fewest in the ipsilateral condition (12.77), significantly fewer than in bilateral (13.45, $p = 0.02$), contralateral (14.16, $p = 0.002$), and no sound (14.12, $p = 0.003$) conditions. Bilateral fixations were also significantly fewer than contralateral ($p = 0.036$) and no sound ($p = 0.048$). Contralateral fixations were most numerous and did not differ from no sound ($p = 0.90$). The interaction was not significant, $F(6, 138) = 0.92$, $p = 0.48$ (see Figure 2b).

Mean Saccade Amplitude: ANOVA revealed a significant main effect of set size, $F(2, 46) = 63.31$, $p < 0.001$, $p^2 = 0.73$. Saccade amplitude was significantly larger for set size 60 (4.96°) than for set size 48 (4.50°, $p < 0.001$) and set size

36 (4.21°, $p < 0.001$), with set size 48 also larger than set size 36 ($p < 0.001$). The main effect of spatial consistency was significant, $F(3, 69) = 5.63$, $p = 0.002$, $p^2 = 0.20$. Amplitude was largest in the ipsilateral condition (4.64°), significantly greater than bilateral (4.55°, $p = 0.017$), contralateral (4.54°, $p = 0.028$), and no sound (4.49°, $p < 0.001$). Bilateral amplitude did not differ from contralateral ($p = 0.875$) or no sound ($p = 0.154$), and contralateral did not differ from no sound ($p = 0.156$). The set size × spatial consistency interaction was significant, $F(6, 138) = 5.61$, $p < 0.001$, $p^2 = 0.20$. Simple effects analysis revealed significant differences between spatial consistency conditions only for set size 60, $F(3, 21) = 13.05$, $p < 0.001$, $p^2 = 0.65$. For set size 60, ipsilateral amplitude (5.14°) was largest, significantly greater than contralateral (4.87°, $p < 0.001$) and no sound (4.80°, $p = 0.001$), but not significantly different from bilateral (5.03°, $p = 0.34$). Bilateral amplitude was also significantly greater than no sound ($p = 0.004$) and contralateral ($p = 0.018$). No sound amplitude was smallest and did not differ from contralateral ($p = 0.42$). No significant differences emerged for set sizes 48 or 36, $F(3, 21) = 2.37$, $p = 0.099$ and $F(3, 21) = 1.24$, $p = 0.32$, respectively (see Figure 2c).

Mean Fixation Duration: ANOVA revealed no significant main effect of set size, $F(2, 46) = 0.51$, $p = 0.60$; no main effect of spatial consistency, $F(3, 69) = 1.52$, $p = 0.22$; and no significant interaction, $F(6, 138) = 0.87$, $p = 0.52$ (see Figure 2d).

We further analyzed eye movement trajectories and heatmaps across four conditions (with visual targets in the same location) for set size 60 (see Figure 3 [Figure 3: see original paper]). Compared to the no sound baseline, ipsilateral conditions showed highest search efficiency with fewest fixations, followed by bilateral conditions. Although contralateral conditions had slightly fewer fixations than no sound, they showed substantially more fixations and lower efficiency than the other two conditions.

Discussion

Experiment 1 examined how audiovisual spatial consistency affects the Pip-and-Pop effect by manipulating set size and auditory stimulus location. Compared to the no sound baseline, ipsilateral conditions produced the shortest RTs, fewest mean fixations, and largest mean saccade amplitudes, indicating highest search efficiency and maximal Pip-and-Pop effect. Bilateral conditions also produced the effect, though weaker than ipsilateral. Contralateral conditions showed no significant differences from no sound across measures, indicating no Pip-and-Pop effect. These results demonstrate that spatial consistency influences the Pip-and-Pop effect—greater spatial alignment produces larger effects.

In ipsilateral conditions, when auditory stimuli originated from the target side, valid spatial information allowed participants to direct search to the target side immediately. Through few fixations and large saccades, participants rapidly located the target, yielding short search times and high efficiency—producing

maximal Pip-and-Pop effect. This suggests auditory stimuli may function as attentional cues. However, auditory stimuli did not immediately localize the target; they continued to enhance target salience throughout search, as evidenced by Pip-and-Pop effects in bilateral conditions lacking cue validity. Thus, both cueing and audiovisual integration likely contributed to improved search efficiency.

In contralateral conditions, when auditory stimuli originated from the non-target side, invalid spatial information caused participants to search extensively on the wrong side before returning to the target side. This involved numerous fixations and progressively smaller saccades during detailed search on the non-target side, creating search costs. Heatmap comparisons clearly revealed these costs in contralateral conditions (see Figure 3). Consequently, ipsilateral conditions were significantly more efficient than contralateral. Previous research shows that spatially informative auditory cues reduce target detection latency compared to non-informative cues (Perrott et al., 1990). Thus, spatially aligned auditory cues effectively direct spatial attention to visual targets, facilitating search. Additionally, studies demonstrate that multisensory interactions are strongest when stimuli from different modalities are spatially proximal and weaken with increasing spatial separation (Lewald et al., 2001; Slutsky & Recanzone, 2001; Soto-Faraco et al., 2003; Stein & Stanford, 2008). In summary, ipsilateral conditions likely produced maximal Pip-and-Pop effects through combined cueing effects and strong multisensory interactions from spatial alignment, while contralateral conditions' invalid cueing and reduced multisensory interactions (due to increased spatial distance) were insufficient to generate the effect.

Bilateral conditions also produced Pip-and-Pop effects. When auditory stimuli were presented through both channels, creating a spatially uninformative stimulus, synchronous presentation still facilitated visual target identification (Giard & Peronnet, 1999; Ngo & Spence, 2012; Yang et al., 2014). This facilitation likely reflects multisensory performance improvement effects, similar to the redundant signals effect (RSE) (Mishler & Neider, 2016; Mishler & Neider, 2018), where responses to multisensory stimuli are faster and more accurate than to unisensory stimuli (van den Brink et al., 2014; Van der Stoep et al., 2015). Research also shows that integrated multisensory stimuli attract attention more effectively than unisensory stimuli (Lunn et al., 2019). Furthermore, Zou et al. (2012) found that synchronous auditory stimuli increased fixations on the target side compared to the non-target side, suggesting participants could more efficiently reject the non-target side and rapidly shift gaze to the target side. Although bilateral auditory stimuli provided no valid spatial information, the integrative facilitation and enhanced attentional capture from dual-channel stimulation improved search efficiency, producing Pip-and-Pop effects.

In conclusion, the Pip-and-Pop effect emerges through reduced fixation numbers and increased saccade amplitude, thereby shortening search time. Experiment 1 demonstrates that audiovisual spatial consistency influences the Pip-and-Pop effect—greater spatial alignment produces larger effects.

Experiment 2: The Influence of Audiovisual Temporal Consistency

Experiment 2 investigated whether the Pip-and-Pop effect follows the temporal principle—that is, whether temporal consistency modulates the effect.

Method

Participants Using G*Power 3.1.9.2 (Erdfelder et al., 2009; Faul et al., 2007) with an effect size of $p^2 = 0.53$ for temporal consistency (Van der Burg et al., 2008), $\alpha = 0.05$, and power = 0.9, the minimum sample size was 24. We recruited 35 university students, excluding 8 due to excessive blinking, yielding a final sample of 27 participants (5 males, 22 females) aged 18-25 years (M = 20.7, SD = 2.4). All were right-handed with normal hearing and normal or corrected-to-normal vision, and no color blindness. Participants received compensation upon completion.

Materials Based on Experiment 1 results showing significant differences in mean saccade amplitude only between set size 48 and the other sizes, with no differences in RT, fixation number, or fixation duration, we eliminated set size 48. Second, to ensure auditory stimuli coincided with visual distractor changes in asynchronous conditions, we used a single 100 ms interval between color changes (instead of 50, 100, and 150 ms in Experiment 1).

Auditory stimuli were identical to Experiment 1. Tone-Target-Intervals (TTI) had six conditions: (1) No sound; (2) TTI = -200 ms: auditory stimulus preceded target color change by 200 ms; (3) TTI = -100 ms: preceded by 100 ms; (4) TTI = 0 ms: simultaneous; (5) TTI = 100 ms: followed by 100 ms; (6) TTI = 200 ms: followed by 200 ms.

Procedure Experiment 2 used a 2 (set size: 36, 60) × 6 (temporal consistency: no sound, -200 ms, -100 ms, 0 ms, 100 ms, 200 ms) within-subjects design. The procedure matched Experiment 1 (see Figure 4a [Figure 4: see original paper]). The formal experiment comprised 12 blocks, with each combination of set size and temporal consistency presented in separate blocks containing 48 trials each, totaling 576 trials. Each block began with 12 practice trials. Participants could rest between blocks, and the experiment lasted approximately 90 minutes.

Data Analysis We used the same measures as Experiment 1 and conducted 2 (set size) × 6 (TTI) repeated-measures ANOVAs to examine effects on the Pip-and-Pop effect.

Results

Data exclusion criteria matched Experiment 1, resulting in exclusion of 2% of trials.

Accuracy All participants achieved accuracy above 98%. ANOVA revealed no significant main effect of set size, $F(1, 26) = 0.49$, $p = 0.49$; no main effect of TTI, $F(5, 130) = 0.89$, $p = 0.49$; and no significant interaction, $F(5, 130) = 1.64$, $p = 0.15$. Accuracy was high across all conditions.

Reaction Time ANOVA revealed a significant main effect of set size, $F(1, 26) = 102.81$, $p < 0.001$, $p^2 = 0.80$, with longer RTs for set size 60 (4.61 s) than set size 36 (2.82 s). The main effect of TTI was significant, $F(5, 130) = 8.91$, $p < 0.001$, $p^2 = 0.26$. RTs were shortest at TTI = 0 ms (2.73 s), significantly shorter than all other conditions (ps < 0.001). RTs at TTI = -100 ms (3.26 s, $p = 0.001$) and TTI = 100 ms (3.44 s, $p = 0.021$) were significantly shorter than no sound (4.40 s). RTs at TTI = -200 ms (3.85 s, $p = 0.174$) and TTI = 200 ms (4.60 s, $p = 0.744$) did not differ significantly from no sound. RTs at -200 ms were significantly longer than at -100 ms ($p = 0.001$), and RTs at 200 ms were significantly longer than at 100 ms ($p = 0.001$).

The set size × TTI interaction was significant, $F(5, 130) = 3.96$, $p = 0.002$, $p^2 = 0.13$. Simple effects analysis showed significant TTI effects for both set sizes. For set size 60, $F(5, 22) = 17.26$, $p < 0.001$, $p^2 = 0.80$, RTs were shortest at 0 ms (3.35 s), significantly shorter than all other conditions (ps < 0.001). RTs at -100 ms (3.96 s, $p = 0.002$) and 100 ms (4.22 s, $p = 0.02$) were significantly shorter than no sound (5.65 s). RTs at -200 ms (4.77 s, $p = 0.15$) and 200 ms (5.74 s, $p = 0.921$) did not differ from no sound. RTs at -200 ms were significantly longer than at -100 ms ($p = 0.001$), and RTs at 200 ms were significantly longer than at 100 ms ($p < 0.001$).

For set size 36, $F(5, 22) = 8.30$, $p < 0.001$, $p^2 = 0.65$, RTs were again shortest at 0 ms (2.11 s), significantly shorter than all other conditions (ps < 0.001). RTs at -100 ms (2.57 s) were significantly shorter than no sound (3.15 s, $p = 0.01$), while RTs at 100 ms (2.67 s) did not differ significantly from no sound ($p = 0.092$). RTs at -200 ms (2.93 s, $p = 0.372$) and 200 ms (3.47 s, $p = 0.489$) did not differ from no sound. RTs at -200 ms were significantly longer than at -100 ms ($p = 0.018$), and RTs at 200 ms were significantly longer than at 100 ms ($p = 0.013$) (see Figure 5a [Figure 5: see original paper]).

Eye Movement Data **Mean Fixation Number**: ANOVA revealed a significant main effect of set size, $F(1, 26) = 113.85$, $p < 0.001$, $p^2 = 0.81$, with more fixations for set size 60 (16.88) than set size 36 (10.31). The main effect of TTI was significant, $F(5, 130) = 11.47$, $p < 0.001$, $p^2 = 0.31$. Fixations were fewest at 0 ms (9.85), significantly fewer than all other conditions (ps < 0.001). Fixations at -100 ms (11.93, $p < 0.001$) and 100 ms (12.65, $p = 0.006$) were significantly fewer than no sound (16.46). Fixations at -200 ms (14.03, $p = 0.082$)

and 200 ms (16.64, $p = 0.926$) did not differ from no sound. Fixations at -200 ms were significantly more numerous than at -100 ms ($p < 0.001$), and fixations at 200 ms were significantly more numerous than at 100 ms ($p < 0.001$).

The set size × TTI interaction was significant, $F(5, 130) = 5.11$, $p < 0.001$, $p^2 = 0.16$. Simple effects analysis showed significant TTI effects for both set sizes. For set size 60, $F(5, 22) = 19.33$, $p < 0.001$, $p^2 = 0.82$, fixations were fewest at 0 ms (12.02), significantly fewer than all other conditions (ps < 0.001). Fixations at -100 ms (14.43, $p = 0.001$) and 100 ms (15.56, $p = 0.01$) were significantly fewer than no sound (20.90). Fixations at -200 ms (17.27, $p = 0.079$) and 200 ms (20.10, $p = 0.943$) did not differ from no sound. Fixations at -200 ms were significantly more numerous than at -100 ms ($p < 0.001$), and fixations at 200 ms were significantly more numerous than at 100 ms ($p < 0.001$).

For set size 36, $F(5, 22) = 10.19$, $p < 0.001$, $p^2 = 0.70$, fixations were fewest at 0 ms (7.68), significantly fewer than all other conditions (ps < 0.001). Fixations at -100 ms (9.44, $p = 0.002$) and 100 ms (9.75, $p = 0.019$) were significantly fewer than no sound (12.02). Fixations at -200 ms (10.79, $p = 0.18$) and 200 ms (12.19, $p = 0.902$) did not differ from no sound. Fixations at -200 ms were significantly more numerous than at -100 ms ($p = 0.018$), and fixations at 200 ms were significantly more numerous than at 100 ms ($p = 0.006$) (see Figure 5b).

Mean Saccade Amplitude: ANOVA revealed a significant main effect of set size, $F(1, 26) = 9.01$, $p = 0.006$, $p^2 = 0.26$, with larger amplitudes for set size 60 (4.59°) than set size 36 (4.37°). The main effect of TTI was significant, $F(5, 130) = 4.67$, $p = 0.001$, $p^2 = 0.15$. Amplitude was largest at 0 ms (4.78°), significantly larger than all other conditions (ps < 0.021). Amplitudes at -100 ms (4.49°, $p = 0.032$) and 100 ms (4.50°, $p = 0.003$) were significantly larger than no sound (4.25°). Amplitudes at -200 ms (4.43°, $p = 0.055$) and 200 ms (4.45°, $p = 0.151$) did not differ from no sound. Amplitudes at -200 ms and -100 ms did not differ significantly ($p = 0.545$), nor did amplitudes at 200 ms and 100 ms ($p = 0.727$). The set size × TTI interaction was not significant, $F(5, 130) = 2.13$, $p = 0.066$ (see Figure 5c).

Mean Fixation Duration: ANOVA revealed a significant main effect of TTI, $F(5, 130) = 4.39$, $p = 0.001$, $p^2 = 0.14$. Fixation duration was longest at 0 ms (245.04 ms), significantly longer than all other conditions (ps < 0.026). Durations at 100 ms (237.03 ms, $p = 0.031$) and -100 ms (236.11 ms, $p = 0.036$) were significantly longer than no sound (228.14 ms). Durations at 200 ms (237.25 ms, $p = 0.061$) and -200 ms (234.59 ms, $p = 0.187$) did not differ from no sound. Durations at 200 ms and 100 ms did not differ ($p = 0.909$), nor did -200 ms and -100 ms ($p = 0.691$). The main effect of set size was not significant, $F(1, 26) = 1.42$, $p = 0.25$, and the interaction was not significant, $F(5, 130) = 1.26$, $p = 0.29$ (see Figure 5d).

Discussion

Experiment 2 examined how audiovisual temporal consistency affects the Pip-and-Pop effect by manipulating set size and the interval between auditory and visual stimuli. Compared to the no sound baseline, RTs were significantly shorter, mean fixations decreased, mean saccade amplitude increased, and mean fixation duration lengthened at TTIs of 0 ms, -100 ms, and 100 ms. No significant differences from baseline were found at -200 ms or 200 ms. This indicates that Pip-and-Pop effects occurred within the -100 ms to 100 ms window, with maximal effect at 0 ms, while no effect was observed at $\pm$200 ms. These results demonstrate that greater temporal consistency between audiovisual stimuli produces larger Pip-and-Pop effects.

The relationship between RT and TTI followed a U-shaped function: smaller absolute TTI values yielded shorter RTs. The effect occurring within the -100 ms to 100 ms window aligns with the known audiovisual integration temporal window (Lewald et al., 2001; Lewald & Guski, 2003). Previous research shows that auditory stimuli synchronous with visual targets enhance perceived visual intensity (Stein et al., 1996), suggesting that the 0 ms condition may strengthen visual target perception, producing maximal Pip-and-Pop effect. At $\pm$200 ms, outside the integration window, auditory stimuli could not effectively integrate with visual targets, yielding no effect.

Mean fixation number showed the same U-shaped relationship with TTI: smaller absolute TTI values produced fewer fixations. Mean saccade amplitude and fixation duration results also showed optimization only at 0 ms, -100 ms, and 100 ms. Such oculomotor optimization has been attributed to the freezing effect, where abrupt auditory stimuli "freeze" eye movements, delaying saccades and prolonging mean fixation duration (Zou et al., 2012). This allows participants to search across larger spatial extents with fewer fixations, more precisely guiding subsequent fixations to the target and accelerating localization. This freezing-induced prolongation can improve search efficiency both within the current region and by facilitating rapid rejection of target-absent regions (Zou et al., 2012). However, our finding that fixation duration was longest at 0 ms suggests that the facilitative effect is not entirely attributable to freezing. Overall, these results demonstrate that the Pip-and-Pop effect follows the temporal consistency principle—greater temporal alignment produces larger effects.

General Discussion

Using a dynamic visual search paradigm, two experiments investigated how audiovisual spatial and temporal consistency influence the Pip-and-Pop effect. Experiment 1 found maximal effects in ipsilateral conditions, weaker effects in bilateral conditions, and no effect in contralateral conditions. Experiment 2 found maximal effects with simultaneous presentation, diminishing as the interval between stimuli increased. Together, these findings demonstrate that

greater audiovisual consistency in space and time produces larger Pip-and-Pop effects.

Perceptual Load as a Moderator

Both experiments revealed effects of set size on the Pip-and-Pop effect. Experiment 1 found a set size × spatial consistency interaction for mean saccade amplitude, while Experiment 2 found set size × temporal consistency interactions for RT and fixation number. Previous research indicates that increasing distractor number elevates cognitive burden during single-target detection (Pluta et al., 2011). Thus, larger set sizes increased perceptual load, and both experiments showed that larger set sizes produced greater perceptual load and larger Pip-and-Pop effects.

Experiment 1 found that spatial consistency increased saccade amplitude only for set size 60. According to perceptual load theory, high-load conditions require full allocation of attentional resources (Lavie, 2005). In our study, set size 60 demanded greater attentional investment than smaller set sizes, producing maximal Pip-and-Pop effects.

Experiment 2 found reduced auditory facilitation for set size 36 in RT and fixation number measures. For set size 36, significant facilitative effects occurred only at 0 ms and -100 ms, with no effect at 100 ms. Zou et al. (2012), using set size 36, found facilitative effects regardless of temporal alignment, with no significant differences between timing conditions. This may reflect lower perceptual load with smaller set sizes. Research shows that auditory facilitation is stronger when searching for complex versus simple visual stimuli (Knoeferle et al., 2016). With set size 36, the visual search task was relatively simple, allowing rapid target detection based on color changes alone. With set size 60, only synchronous auditory stimuli integrated with visual targets could attract sufficient attention for rapid detection. Together, these results indicate that perceptual load moderates the influence of both spatial and temporal consistency on the Pip-and-Pop effect.

Mechanisms Underlying the Pip-and-Pop Effect

Our findings that audiovisual spatial and temporal consistency affect the Pip-and-Pop effect suggest that the phenomenon follows both spatial and temporal principles of multisensory integration. This provides evidence that multisensory integration at least partially contributes to the effect. However, alternative explanations based on oddball-driven attentional capture remain. Ngo et al. (2010) found comparable facilitation from visual and auditory cues, suggesting that any cue making the visual target an oddball stimulus can produce the effect (Ngo & Spence, 2010b).

Could our Pip-and-Pop effects be explained by oddball-driven attentional capture? First, all conditions except the no sound baseline used bimodal stimulation, so even if simultaneous auditory cues made visual targets oddballs

that captured attention, we could not exclude audiovisual integration contributions. Second, our auditory cues included both spatially informative and uninformative conditions, whereas previous oddball studies used only uninformative cues. Third, our dynamic visual search paradigm differed from previous oddball studies. In our task, synchronous auditory cues made visual targets "pop out" from dynamic distractors. In contrast, oddball studies used sequential search paradigms with brief target presentations, where synchronous cues produced freezing effects that allowed short-duration search (Ngo & Spence, 2012). Fourth, oddball-driven attentional capture better explains freezing effects—oddball stimuli prolong subjective time (Tse et al., 2004). Only Experiment 2 showed freezing effects (prolonged fixation duration), while Experiment 1 did not, possibly because spatially informative cues allowed oculomotor control without requiring freezing.

For these four reasons, we propose that multisensory integration contributes partially to the Pip-and-Pop effect, but we cannot fully exclude or confirm the role of oddball-driven attentional capture. The precise mechanism requires further investigation.

Conclusion

Audiovisual spatial and temporal consistency modulates the Pip-and-Pop effect. Greater spatial and temporal alignment produces larger effects. These findings provide evidence that multisensory integration contributes to the Pip-and-Pop effect. Future research should directly compare explicit audiovisual integration conditions with oddball-driven attentional capture conditions within the dynamic visual search paradigm to further elucidate the underlying mechanism.

References

Chamberland, C., Hodgetts, H. M., Vallières, B. R., Vachon, F., & Tremblay, S. (2016). Pip and Pop: When auditory alarms facilitate visual change detection in dynamic settings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 284-288.

Chastain, G., & Cheal, M. (1999). Time course of attention effects with abrupt-onset and offset single- and multiple-element precues. *The American Journal of Psychology, 112*(3), 411-436.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: *Zeitschrift für Psychologie, 217*(3), 108-124.

Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of Cognitive Neuroscience, 10*(5), 590-604.

Escera, C., Corral, M. J., & Yago, E. (2002). An electrophysiological and behavioral investigation of involuntary attention towards auditory frequency, duration and intensity changes. *Cognitive Brain Research, 14*(3), 325-332.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences.* Behavior Research Methods, 39*(2), 175-191.

Fister, J. K., Stevenson, R. A., Nidiffer, A. R., Barnett, Z. P., & Wallace, M. T. (2016). Stimulus intensity modulates multisensory temporal processing. *Neuropsychologia, 88*, 92-100.

Fleming, J. T., Noyce, A. L., & Shinn-Cunningham, B. G. (2020). Audio-visual spatial alignment improves integration in the presence of a competing audio-visual stimulus. *Neuropsychologia, 146*, 107530.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*(5), 473-490.

Kösem, A., & van Wassenhove, V. (2012). Temporal structure in audiovisual sensory selection. *PLoS One, 7*(7), e40936.

Knoeferle, K. M., Knoeferle, P., Velasco, C., & Spence, C. (2016). Multisensory brand search: How the meaning of sounds guides consumers' visual attention. *Journal of Experimental Psychology: Applied, 22*(2), 196-210.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences, 9*(2), 75-82.

Lewald, J., Ehrenstein, W. H., & Guski, R. (2001). Spatio-temporal constraints for auditory-visual integration. *Behavioural Brain Research, 121*(1-2), 69-79.

Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research, 16*(3), 468-478.

Lunn, J., Sjoblom, A., Ward, J., Soto-Faraco, S., & Forster, S. (2019). Multisensory enhancement of attention depends on whether you are already paying attention. *Cognition, 187*, 38-49.

Mishler, A., & Neider, M. (2016). Evidence for the Redundant Signals Effect in Detection of Categorical Targets. *Journal of Vision, 16*(12), 1024.

Mishler, A. D., & Neider, M. B. (2018). Redundancy gain for categorical targets depends on display configuration and duration. *Visual Cognition, 26*(6), 393-404.

Ngo, M. K., & Spence, C. (2010a). Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. *Attention, Perception, & Psychophysics, 72*(6), 1654-1665.

Ngo, M. K., & Spence, C. (2010b). Crossmodal facilitation of masked visual target identification. *Attention, Perception, & Psychophysics, 72*(7), 1938-1947.

Ngo, M. K., & Spence, C. (2012). Facilitating masked visual target identification with auditory oddball stimuli. *Experimental Brain Research, 221*(2), 129-136.

Pariyadath, V., & Eagleman, D. (2007). The effect of predictability on subjective duration. *PLoS One, 2*(11), e1264.

Perrott, D. R., Saberi, K., Brown, K., & Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. *Perception & Psychophysics, 48*(3), 214-226.

Pluta, S. R., Rowland, B. A., Stanford, T. R., & Stein, B. E. (2011). Alterations to multisensory and unisensory integration by stimulus competition. *Journal of Neurophysiology, 106*(6), 3091-3101.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport, 12*(1), 7-10.

Soto-Faraco, S., Kingstone, A., & Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia, 41*(13), 1847-1862.

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences, 1296*, 31-49.

Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience, 8*(6), 497-506.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(4), 255-266.

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., & Wallace, M. T. (2012). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. *Experimental Brain Research, 219*(1), 121-137.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400-410.

Tang, X., Wu, J., & Shen, Y. (2016). The interactions of multisensory integration with endogenous and exogenous attention. *Neuroscience & Biobehavioral Reviews, 61*, 208-224.

Tsai, Y. Y., & Yeh, S. L. (2013). Freezing effect in tactile perception: Sound facilitates tactile identification by enhancing intensity but not duration. *Journal of Experimental Psychology: Human Perception and Performance, 39*(4), 925-932.

Tse, P. U., Intriligator, J., Rivest, J., & Cavanagh, P. (2004). Attention and the subjective expansion of time. *Perception & Psychophysics, 66*(7), 1171-1189.

Van der Brink, R. L., Cohen, M. X., van der Burg, E., Talsma, D., Vissers, M. E., & Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cerebral Cortex, 24*(8), 2169-2177.

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and Pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053-1065.

Van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55*(3), 1208-1218.

Van der Stoep, N., Van der Stigchel, S., & Nijboer, T. C. (2015). Exogenous spatial attention decreases audiovisual integration. *Attention, Perception, & Psychophysics, 77*(2), 464-482.

Van der Stoep, N., Van der Stigchel, S., Nijboer, T. C. W., & Spence, C. (2017). Visually induced inhibition of return affects the integration of auditory and visual information. *Perception, 46*(1), 6-17.

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance, 26*(5), 1583-1590.

Yang, W., Chu, B., Yang, J., Yu, Y., Wu, J., & Yu, S. (2014). Elevated audio-visual temporal interaction in patients with migraine without aura. *Journal of Headache & Pain, 15*(1), 44.

Zou, H., Müller, H. J., & Shi, Z. (2012). Non-spatial sounds regulate eye movements and enhance visual search. *Journal of Vision, 12*(5), 123-129.

Note: Figure translations are in progress. See original paper for figures.