A cognitive temporal window supports flexible integration of multimodal events

By

Madison Lee

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

In

Psychology

December 17, 2022

Nashville, Tennessee

Approved:

Daniel T. Levin

Duane Watson

ACKNOWLEDGEMENTS

Thank you to Dan Levin for being an incredible mentor, I would not be where I am as a researcher without your invaluable guidance. Thank you to Duane Watson and Mark Wallace for agreeing to be on my committee and providing such helpful advice and feedback whenever we meet. Thank you to my friends and colleagues in the department. I feel so lucky to be surrounded by such insightful people. Everyone is so kind and generous with their time, always willing to have interesting and intellectual conversations. I have grown so much as a researcher with everyone's support. Finally, thank you to my family and to Travis Mason for being my support system throughout my time as a graduate student.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

Introduction

Effective perception and understanding of real-world events often requires the integration of auditory and visual information. For simple events, such as clapping hands, visual movements must be tightly linked with the sounds that emanate from them. This form of integration is often referred to as *multisensory integration*. Research in the field of neuroscience has established that this form of integration is associated with a multisensory temporal binding window of approximately +/- 250 ms where multisensory event information (i.e., a beep and flash) can be asynchronous yet perceptually bound and perceived as occurring simultaneously (Wallace & Stevenson, 2014). Such a window is necessary in part because the relationship between auditory and visual features of multisensory events is incompletely determined by simple timing features. For example, propagation delays both externally (because of differences in the speed of sound and light), and internally (because of between-modality variability in neural processing times) can create ambiguities in the precise timing of movements and the sounds they produce. Thus, it is useful for the perceptual system to treat as equivalent a range of relative timings (for review see Zhou, Cheung, & Chan, 2020).

While a window of +/- 250 ms has been established for multisensory events, there is an analogous issue for events that include auditory and visual components linked not by a common distal source, but rather because they are bound by mutually reinforcing forms of meaning. For example, the relationships between speech and speaker-produced movements (i.e., gestures and referred-to actions) are often characterized by temporal delays. However, it is not clear whether

the temporal delays in these multimodal relationships are meaningful and necessary for effective event perception and thus incorporated into the processing stream. As we will review below, some theories of event perception at least imply that specific inter-event timings are important for event perception, but on the other hand, a range of empirical phenomena suggest that fine grain event perception can be surprisingly insensitive to brief temporal disturbances of up to several seconds. Findings such as these suggest that temporal delays in inter-event relationships beyond the 250 ms multisensory integration window might be treated by many parts of the visual-cognitive system as equivalent. Thus, we test whether a larger *event-integration* window might extend for several seconds by assessing the degree to which cognitive processing is impacted by disturbances to temporal relationships between related visual events and speech.

Models of event perception such as Event Segmentation Theory (EST; Zacks et al., 2007) imply the necessity of temporal expectations for effective event perception. According to EST, perceivers continuously generate predictions and compare them to incoming information. Mismatches produce prediction errors that in turn induce the perception of event boundaries, and effective boundary segmentation is shown to be crucial for event understanding and learning (Kurby & Zacks, 2008; Flores et al., 2017). Importantly, predictions are often developed via one modality and confirmed via another modality. For example, an instructor might say "Now you need to subtract A from B" as they calculate a value on the whiteboard. Students following along will consequently predict A to be subtracted from B, and this prediction will be visually confirmed from the instructor's actions on the board. In this example, EST would posit the temporal relationship between the related visual and auditory events is consequential to prediction generation and subsequent perceptual processing and event understanding (Hommel et al., 2001; Zacks, 2020).

Supported by behavioral and neurophysiological evidence, the proposed predictive mechanism in EST does imply the importance of temporal information in event perception and down-stream cognitive processes. (Zacks et al., 2011; Reynolds et al., 2007). For example, Eisenberg, Zacks, & Flores (2018) propose that movie viewers who look to objects that actors are about to interact in the next second or so, do so to test very short-term predictions about upcoming events. Similarly, the narrative comprehension literature suggests that temporal information plays a fundamental role in memory representation as readers continued to generate temporally organized representations of events to facilitate comprehension and future prediction despite the narrative's lack of explicit temporal structure (Claus & Kelter, 2006). Further, research has demonstrated that participants extract regularities of temporal structure between visual events and use them to make temporal predictions. Events that fulfill a prediction consequently improve the perception and information processing of that event (Rohenkohl et al., 2012; Wiener & Kanai, 2016). For example, Graf et al (2007) found that participants projected the motion of point-light walkers forward during an occlusion, as indicated by priming for targets that matched the forward-projected configuration.

While most would agree that temporal expectations are at least in part the informational basis of event perception, the range of circumstances under which temporal information is utilized in event perception is unclear. Most of the research described induces participants to scrutinize events for small deviations in timing, either by specific task requirements, or by repeated presentations of dozens of events that parametrically vary in timing. When such repetition and demand to scrutinize events are lessened, precise temporal encoding may disappear. For example, Levin et al. (2022) observed that participants failed to report discontinuities in movies of short events where an edit was associated with action overlaps or

ellipses of up to 400 ms, especially when the events were not repeated with instructions to scrutinize them. Hymel et al. (2016) found evidence for nondiscrimination of temporal inconsistencies over even longer durations. Participants in those experiments viewed short movies in which a series of actions (such as grabbing a screwdriver and using it) were each depicted in brief shot. There was an average of 11 shots (and actions) per movie, and for some movies one of the actions was presented in reverse order (for example, the shot depicting use of the screwdriver was shown before the shot depicting grabbing it). The mean duration of the reversed shots was between 300 and 1000 ms, so the reversal lasted up to 2 seconds. Even so, participants had difficulty detecting the reversals when they were instructed to look for them and were unable to detect them when completing a distraction task. Participants never detected the reversals when they were asked to attend closely to the movies but were not specifically instructed to look for them.

Findings such as these suggest that temporal information within events, and even the temporal sequence between short events, may not be represented by default. Broadly, these findings may be consistent with the longstanding idea of a "psychological present" consisting of a 2-3 second window in time within which incoming sensory information is automatically and pre-semantically integrated (Pöppel, 2009). In other words, conscious activity is segmented into 2-3 sec windows and perceptual information in that window is integrated independent of the content being processed. Pöppel (2009) summarizes empirical work from movement control, spontaneous speech, and auditory and visual processing that all provide results consistent with this hypothesis. More recently, Fairhall (2014) presented participants with 13-second silent movie clips. The clips were divided into chunks, varying in duration from less than one second to the full 13 seconds. Then, each chunk was divided into 1/4 second intervals and the intervals

were scrambled. Participants rated how difficult the scrambled clips were to follow. Video clips that were scrambled within 2-second chunks were reportedly easy to follow, but as chunk duration increased, participant's difficulty rating increased considerably. Fairhall (2014) concluded that the increased difficulty derived from the fact scrambling within the larger chunks displaced action information beyond the psychological present.

Combining the Fairhall (2014) finding, the temporal nondiscrimination findings, and the idea of pre-semantic integration implies that events within the temporal present are integrated automatically, but without default inclusion of much temporal or sequential information. On this view, fine grain event perception seems to be surprisingly insensitive to brief temporal disturbances of up to several seconds, which may place important limits on the nature of the predictive processing posited by EST, at least for fine grained events. However, other data reviewed above clearly demonstrate that participants can generate temporal predictions in some cases, and some evidence in support of temporal window hypotheses does assume that precise predictions are possible within the time-range of the window (Pöppel, 2009). It is therefore possible that evidence for an event-integration window characterized by nondiscrimination of temporal information is generated only from limited situations. For example, it is possible that just like evidence for temporal discrimination is limited to situations involving very high levels of scrutiny, evidence of nondiscrimination is limited to situations characterized by shallow processing and low levels of attention and effort.

Current Study and Hypotheses

To test for evidence of an event-integration window in multimodal perceptual processing in situations involving relatively deep processing, we disrupted the temporal relationship

between a person's actions and the speech describing those actions for instructional videos that participants knew they would be tested on. We did this by displacing the audio channel of screen-captured instructional videos forward or backward 0, 1, 3, or 7 seconds relative to the video. So, a one-second forward displacement would entail moving the instructor's speech one second earlier relative to the actions they are producing, and a one-second backward displacement would move the speech to be one-second later than the actions. It is important to note that these displacements disrupt only the conceptual relationship between on-screen movements (represented primarily by movements of the instructor's cursor, their typing, and their menu selections) and the instructor's speech which was not visibly produced because the instructor's face could not be seen – only their computer screen was visible. We purposely chose a setting where the displacements would not produce multimodal perceptual mismatches between sounds and movements that produced the sounds, for example by disrupting the synchrony between an instructor's lip movements and their speech. Although this setting does not include information that is often available to perceivers, it is an extremely common learning setting that viewers find comfortable, as evidenced by the literally billions of views these videos receive (Jaeger, Little, & Levin, 2021).

To better understand temporal information in multimodal events, we coded 12 minutes of screen-captured instructional video for the frequency and timing of predictive relationships between audio and channels (Figure 1). The delay duration between the prediction-generating action from one modality and the subsequent action from the other modality was coded. For example, an instructor says, "Now let's create a new Excel sheet" (prediction-generating action), and 22 frames after "sheet" was verbalized the instructor clicks the icon representing a new Excel sheet. In a separate example, an instructor says, "You need to then press sort", and 4

frames after "sort" was verbalized the instructor presses the "Sort" button. The analysis shown in Figure 1 suggests that most intra-event information falls within the range of one second (~30 video frames).

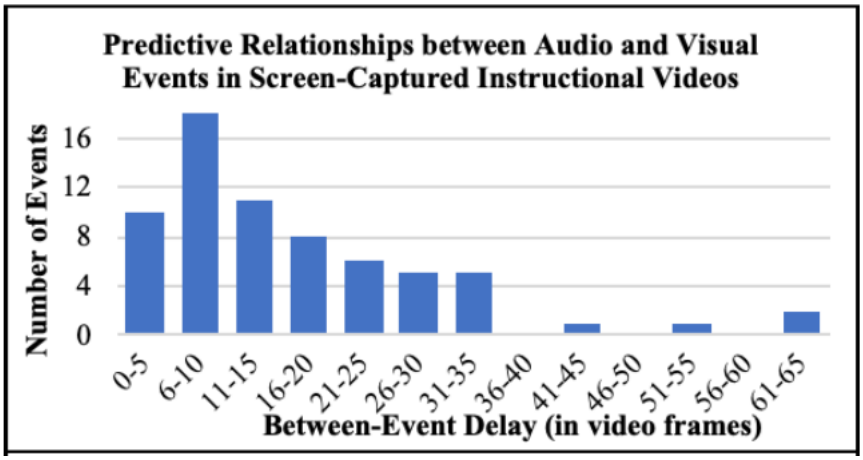

Figure 1. Distribution of between-event delay durations (in video frames) between the audio and video channel of an event. Six different screen-captured instructional videos were coded.

In the current experiments, participants watched eight screen-captured instructional videos, each varying in disruption levels. After each video participants completed measures of event segmentation, learning, disruption awareness, segmentation uncertainty, and perceived workload. If temporal expectations are the informational basis of event perception, we would expect temporally disrupting multimodal relationships would have negative consequences on event perception and cognitive processes. Because we displaced audio channels forward and backward, it is important to note that we are in some cases disrupting specific temporal delays (for example Figure 2B, displacing audio 1.5 seconds before the paired action instead of .5 seconds before) and in other cases disrupting temporal sequences (for example Figure 2C, displacing the verbalization to occur .5 seconds after the paired action instead of .5 seconds before).

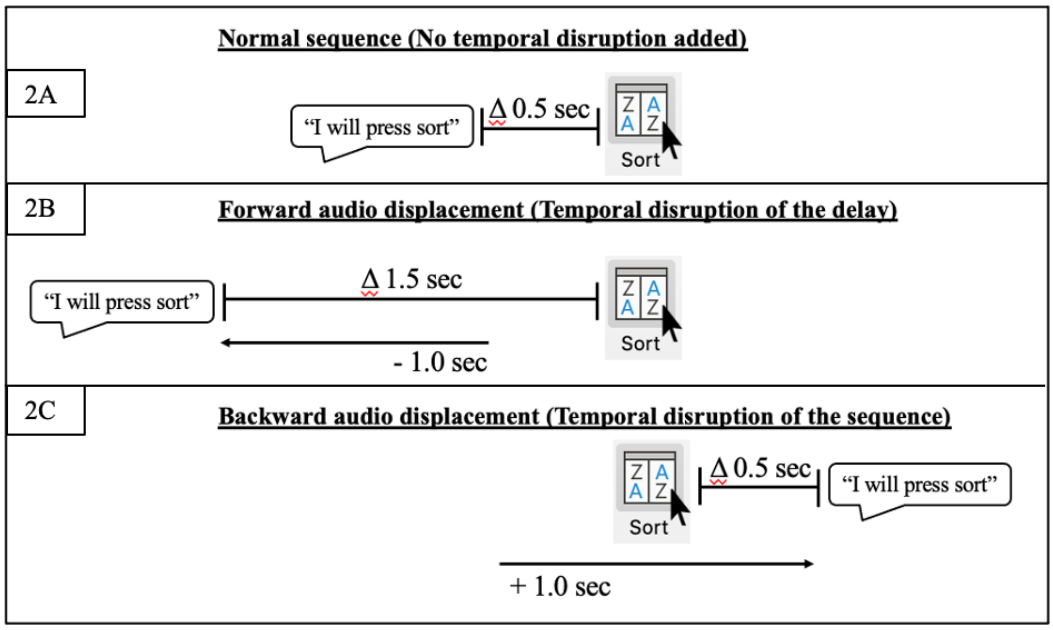

Figure 2. Applying 500ms displacement to an entire instructional video uniquely disrupts each inter-event relationship.

Grounded by the research described above, we hypothesize that only disruptions beyond the psychological present will significantly impact our cognitive and perceptual measures. Only as temporal disruption increases to 7 seconds, should we observe an increase in prediction error and event model updating which should be represented by an increase in event segmentation and a decrease in participants' segmentation agreement scores. We also predict a decrease in learning for 7-second disruptions as ineffective event segmentation has been shown to negatively affect learning and memory (Flores et al., 2017). Additionally, just as Fairhall (2014) saw an increase in difficulty measures for temporal disruptions lasting beyond the psychological present, we hypothesize a 7-second disruption will cause a spike in segmentation uncertainty and perceived workload. Finally, because most intra-event temporal relationships vary within a range of ~1 second (Figure 1), we hypothesize participants will remain unaware of the briefest displacements.

CHAPTER II

Experiment 1

Method

Participants

60 participants from Vanderbilt University's undergraduate participant pool completed Experiment 1 on-line. 11 participants failed the instruction check and were excluded from analyses, leaving 49 participants in analyses. Participant's average age was 19.5 years old. 22 reported as female, 26 as male, and 1 preferred not to answer. The sample size was determined primarily based on the amount of data that could be collected in a given timespan, though a post-hoc power analysis reveals all significant F-tests in Experiment 1 achieved above .99 power.

Videos

Participants watched 8 screen-captured instructional videos where a narrator explained and demonstrated various topics in Microsoft Excel and Microsoft Paint. Lessons in Excel were about how to transpose data, how to freeze panes, how to use a formula for changing letter cases, and how to calculate averages and medians. Lessons in Paint were about how to use the right click erase feature, how to create textured lines, how to use transparent selection features, and how to create clean outlines when drawing. The average video duration was 70 seconds, ranging from 50 to 98 seconds in length.

Audio channels for each video were manipulated using Final Cut Pro. In each video, the audio channel was displaced (i.e., moved independently from the associated video track) to lead or lag the video channel by 0, 1, 3, or 7 seconds. The video channel was held at a freeze frame at

the beginning or end of the video to match the duration the audio channel was displaced. In forward displaced videos, the audio channel was manipulated to start ahead of the video channel. In backward displaced videos, the audio channel was manipulated to lag the video channel. In this within-subjects design, each participant watched all 8 videos: two unmanipulated videos (0-second displacement), two 1-second displaced videos (one forward displaced, one backward displaced), two 3-second displaced videos, and two 7-second displaced videos. Participants were randomly assigned one of eight possible block orders which counterbalanced the degree to which each video was temporally displaced. Differences between forward and backward displacements will not be analyzed or reported. Considering this extensive variability in how a temporal disruption could alter a given multimodal event, and because it is possible that actions predict utterances and utterances predict actions, we had no specific hypotheses for significant differences between forward and backward conditions. We therefore collapsed across forward and backward displacements for all analyses.

Procedure

Participants began the experiment by completing basic demographic questions and reading instructions. Instructions explained that participants would be watching 8 videos, two times each. They were instructed to learn as much as they could during their first viewing, then to segment events during their second viewing. Prior to reading the instructions, participants were warned they would be asked a question about the instructions immediately after reading them. Participants were asked, “Which of the following sentences was NOT in the instructions?”. This instruction check included six possible answers, five were key points taken directly from the instructions and one was a sentence that was not in the instructions (i.e., “The videos you will

be watching are clips about geography"). Those who incorrectly responded to this question were excluded from analyses.

Eight multiple-choice content questions were created for each video. Participants answered all 64 questions at the start of the experiment to measure their baseline knowledge about Excel and Paint. We explicitly aimed to create questions that tested participants understanding of the sequence in which events occurred and questions that tested participants on their comprehension of the tasks demonstrated. For example, a sequence question about the transparent selection feature video was "In the previous video, what steps were taken before the author began using the spray paint tool?". A comprehension question about that video was "Based on this video, when should transparent selection be turned on?". As demonstrated in these example questions, some questions were specific to the video while some questions were more general and could have potentially been answered correctly if a participant frequently used Excel and Paint.

After the pre-test, participants practiced segmenting events on a novel, unmanipulated video. Participants were told to, "Press the "N" key when you believe one meaningful event ends and another event begins. There is no right or wrong answer; we are simply interested in how you do this task." Participants had one minute to practice segmenting events.

After segmentation practice, participants watched their first video twice. For the first viewing, participants were told, "Your primary task is to learn as much as you can. Do not segment events yet, your responses will not be recorded.". For the second viewing, participants were told to find event boundaries and to "press the "N" key when you believe one meaningful event ends and another event begins".

After the second viewing, participants completed a series of questionnaires. The first questionnaire assessed participants' awareness of the temporal mismatches. To avoid cueing participants to the manipulation, participants reported whether each of four different possible "abnormalities" occurred in the video they just watched. Participants who reported "Yes" to "The audio was out of synchronization with the video" abnormality were classified as being aware of the disruption. Other abnormality options that did not take in any video were "Important pieces of audio were cut out", "Some key events in the video were not discussed in the audio", "The video randomly froze either momentarily or for a long time".

Next, participants responded to a five-point Likert scale asking, "How uncertain were you while segmenting events in this video (i.e., detecting the end of one event and beginning of another event)?" (1 – Not at all uncertain, 5 – Very uncertain). Participants then responded to the NASA task load index (TLX), a questionnaire that measures an individual's perceived workload. The questionnaire has 6 scales (mental demand, physical demand, temporal demand, performance, effort, frustration. We used all but the physical demand scale. Finally, participants answered the same 8 multiple-choice content questions they had seen in the pre-test. This procedure repeated for all 8 videos.

Results

Learning

We assessed learning by subtracting each participants' pre-test score from their post-test score. Although there was a downward trend of learning as disruption increased, there was no significant effect of temporal disruption on learning scores, $F(3, 144) = 1.371$, $p = 0.254$, $\eta^2 =$

0.028 (Figure 3A). Across all conditions, on average participant's post-test scores were 39% higher than their pre-test scores, suggesting a high level of learning, $t(195) = 23.774$, $p = < .001$.

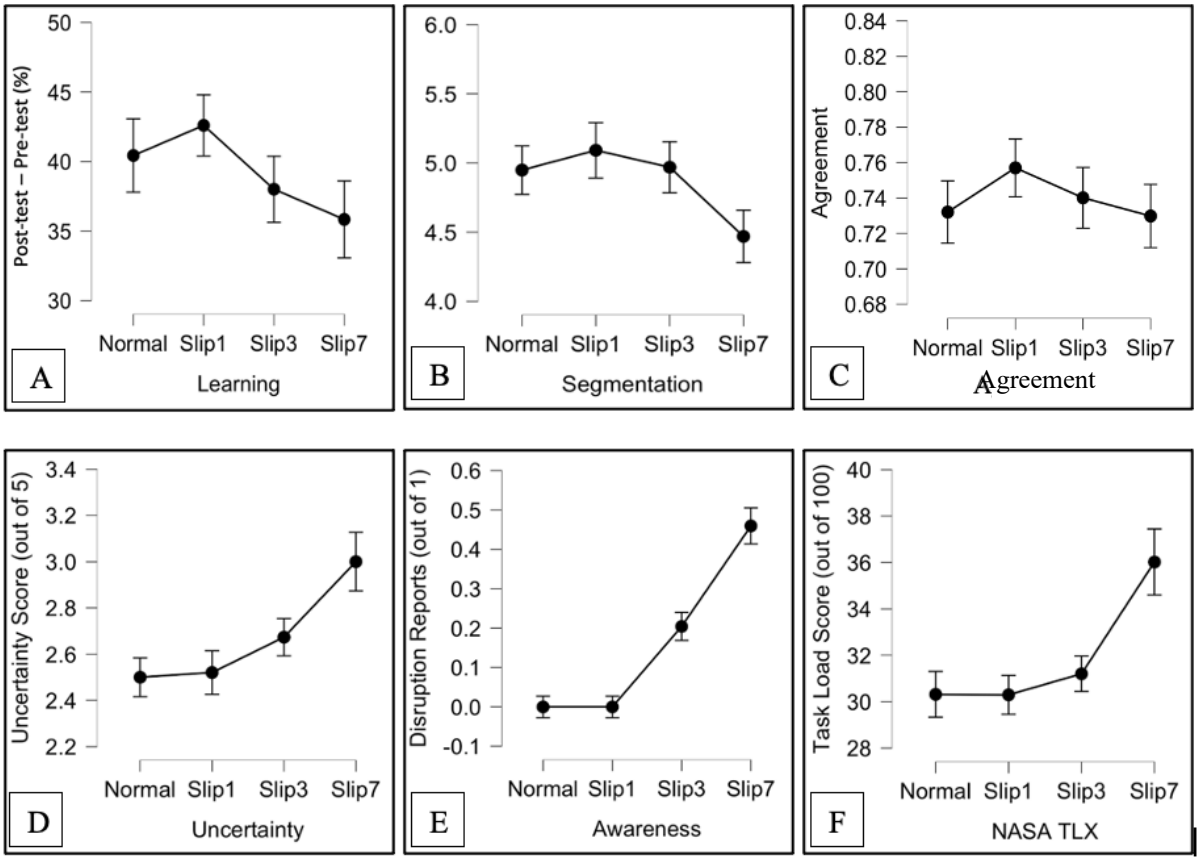

Figure 3. Results from all six measures of cognitive and perceptual processing: (A) Average learning scores calculated from post-test minus pre-test scores presented by condition. (B) Average number of time participants pressed "N" to represent a segmentation presented by condition. (C) Participant's average level of agreement in segmentation patterns by condition. (D) Average level of uncertainty while segmenting events (E) Awareness of temporal disruption by condition. (F) Average percieved workload by condition.

Event Segmentation

Segmentation Count. The number of times each participant pressed "N" to report an event boundary was recorded. Contrary to the prediction that disruption would increase event segmentation, there was a downward trend in the number of segmentations as disruption increased, but this effect was not significant, $F(3, 144) = 2.150$, $p = 0.097$, $\eta^2 = 0.043$ (Figure

3B). The difference between the 7-second disrupted videos and the undisrupted videos was nonsignificant correcting for multiple comparisons (t = -1.813, p = 0.072). Across all videos and all conditions (average duration of 70 seconds), participants segmented 4.87 times on average, approximately 1 segment every 14 seconds.

Segmentation Agreement. Segmentation agreement was calculated using a point biserial correlation across 1 second bins comparing a participant's segmentations with the segmentation pattern of the rest of the group who viewed the exact same video (Zacks et al., 2006). Each participant received an agreement score for each video they watched. Averaging over all videos, there was no difference in segmentation agreement scores between conditions, $F(3, 144) = 0.512$, $p = 0.674$, $\eta^2 = 0.004$ (Figure 3C). Across all videos and all conditions, participant's average segmentation agreement score was 0.740 (on a 0 to 1 scale).

Segmentation Uncertainty

Displacement significantly affected segmentation uncertainty, $F(3, 144) = 5.547$, $p = 0.005$, $\eta^2 = 0.104$ (Figure 3D). Bonferroni corrected post-hoc comparisons reveal watching a video that was temporally displaced 7 seconds caused significantly more uncertainty than 0 second temporally displaced videos (t = -3.604, p = 0.003), and 1 second temporally displaced conditions (t = -3.457, p = 0.004).

Disruption Awareness

Temporal displacement significantly affected participants' awareness of disruption, $F(3, 144) = 39.072$, $p < 0.001$, $\eta^2 = 0.449$ (Figure 3E). Not once did a participant report awareness of disruption in the 1-second displacement condition. Approximately 20% of participants noticed the disruption in the 3 second displacement and 45% in the 7 second displacement. Post-hoc proportion tests reveal 3 second displacements led to a significantly greater proportion of

awareness compared to 0 second displacements (z = -3.300, p < 0.001) and 1 second displacements (z = -3.300, p < 0.001). 7 seconds displacements led to a significantly greater proportion of awareness compared to all other conditions (0 second displacement: (z = -5.334, p < 0.001), 1 second displacement: (z = -5.334, p < 0.001), 3 second displacement: (z = -2.641, p < 0.004).

Participants rarely false alarmed by reporting the other three abnormalities (Figure 4).

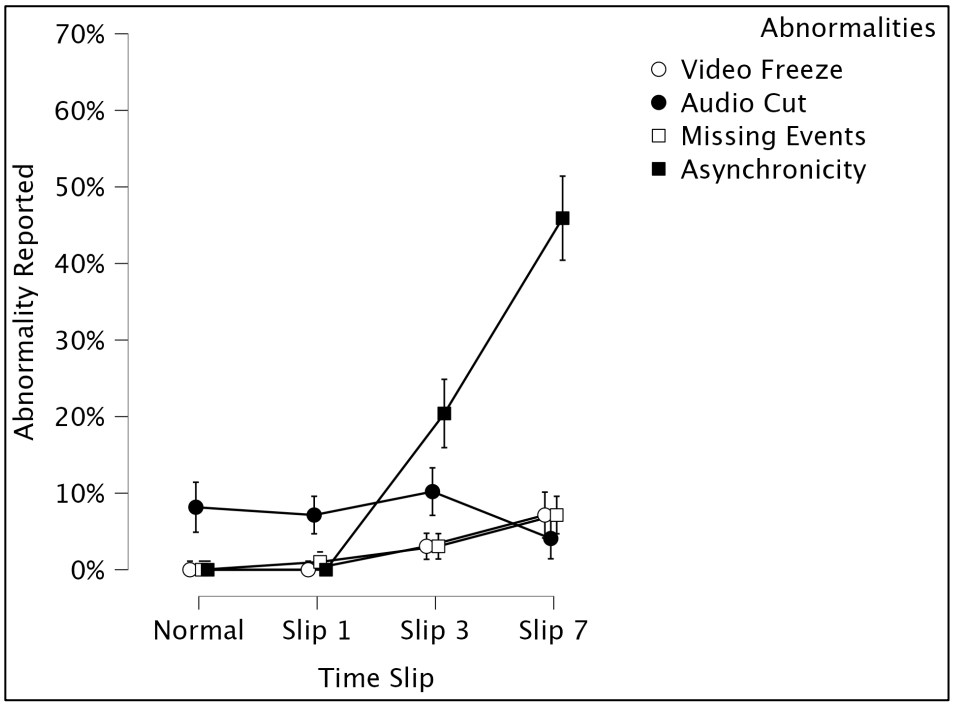

Figure 4. Number of reports for each abnormality are presented. "Video Freeze", "Audio Cut", and "Missing Events" are all abnormalities that never took place.

Perceived Workload

The degree of temporal displacement significantly affected participants' perceived workload, as measured by NASA TLX, $F(3, 144) = 7.035$, $p = 0.002$, $\eta^2 = 0.128$ (Figure 3F). Post-hoc tests reveal watching a video that was temporally displaced 7 seconds created significantly greater perceived workload than all other conditions (0 second displacement: (t = -

3.904, $p < 0.001$), 1 second displacement: ($t = -3.918$, $p < 0.001$), 3 second displacement: ($t = -3.296$, $p = 0.005$).

Experiment 1 Discussion

Temporal displacement increased segmentation uncertainty, disruption awareness, and perceived workload, but in the case of segmentation uncertainty and perceived workload only large 7-second disruptions had any impact. When asked, participants were able to detect 3-second disruptions, but only in about 20% of cases. The 7-second disruptions were detectable on about half of trials. While nonsignificant, a general downward trend appeared in learning scores as disruption increased. Contrary to our predictions, event segmentation was not significantly affected by temporal displacements, but there was a nonsignificant trend for 7-second displacements to result in fewer segmentations.

Most notably, 3-second temporal displacements only significantly impacted disruption awareness and 1-second temporal displacements had no impact on any measure and was never detected even though most intra-event temporal relationships vary within this range (Figure 1). This suggests there might exist an event-integration window, a temporal window of perceptual flexibility within which specific temporal expectations are not necessarily tracked.

CHAPTER III

Experiment 2

Experiment 1 suggests that temporal event integration may be more flexible than current theories predict. However, it is possible that screen-capture instructional videos uniquely encourage this flexibility. For example, it is possible that the visual events associated with language in these videos are unnatural and therefore do not benefit much from possibly more deeply engrained processes that may associate hand gestures with language. To combat this possible limit in the generalizability of Experiment 1's results, Experiment 2 followed the same procedure but instead used live-action instructional videos as stimuli. These videos depicted an instructor explaining how to perform some task, so they included audio of the instructor's voice and video of their hands engaging in the action and gesturing. Crucially, the videos did not show the instructor's face so the videos could test event integration between actions and language in the absence of multimodal perceptual information specifying the relationship between speech and lip movements.

Method

Participants

83 participants from Vanderbilt University's undergraduate participant pool completed Experiment 2 on-line. 10 participants failed the instruction check and were excluded from analyses, leaving 73 participants in analyses. Participant's average age was 18.9 years old. 44 reported as female, 28 as male, and 1 reporting as nonbinary. The sample size was determined

primarily based on the amount of data that could be collected in a given timespan, though a post-hoc power analysis reveals all significant F-tests in Experiment 2 achieved above .99 power.

Videos

Participants watched 8 live-action instructional videos in which a narrator demonstrated various tasks. Lessons included administering a vaccine, making a latte, setting up a sewing machine, setting up a board game, crafting a cocktail, potting a plant, playing a card game, and inserting an intravenous tube. The average video duration was 3 minutes and 14 seconds, ranging from 63 seconds to 4 minutes. The process of editing videos to create each temporally displaced condition was identical to Experiment 1. These were single-shot videos, so the only additional editing that took place was when there were long moments of no relevant audio or video (e.g., 20 seconds of milk frothing). In these moments, we speeded the video 8x and removed the audio. While filming these videos, we did not include the instructors face to avoid participants noticing disruptions only because instructors' lips were out of sync.

Procedure

Experiment 2 followed a procedure identical to Experiment 1. The only change was that we created 64 new multiple-choice questions for the live-action videos.

Results

Experiment 2 results followed the same pattern as Experiment 1 across all measures. The only difference was, unlike in Experiment 1, there was a significant effect of displacement on learning scores in Experiment 2.

Learning

The degree of temporal displacement significantly affected learning scores, $F(3, 216) = 3.652$, $p = 0.013$, $\eta^2 = 0.048$ (Figure 5A). Post-hoc tests revealed that watching a 7-second displaced video caused significantly less learning than watching a 0-second displaced video ($t = 2.970$, $p = 0.020$).

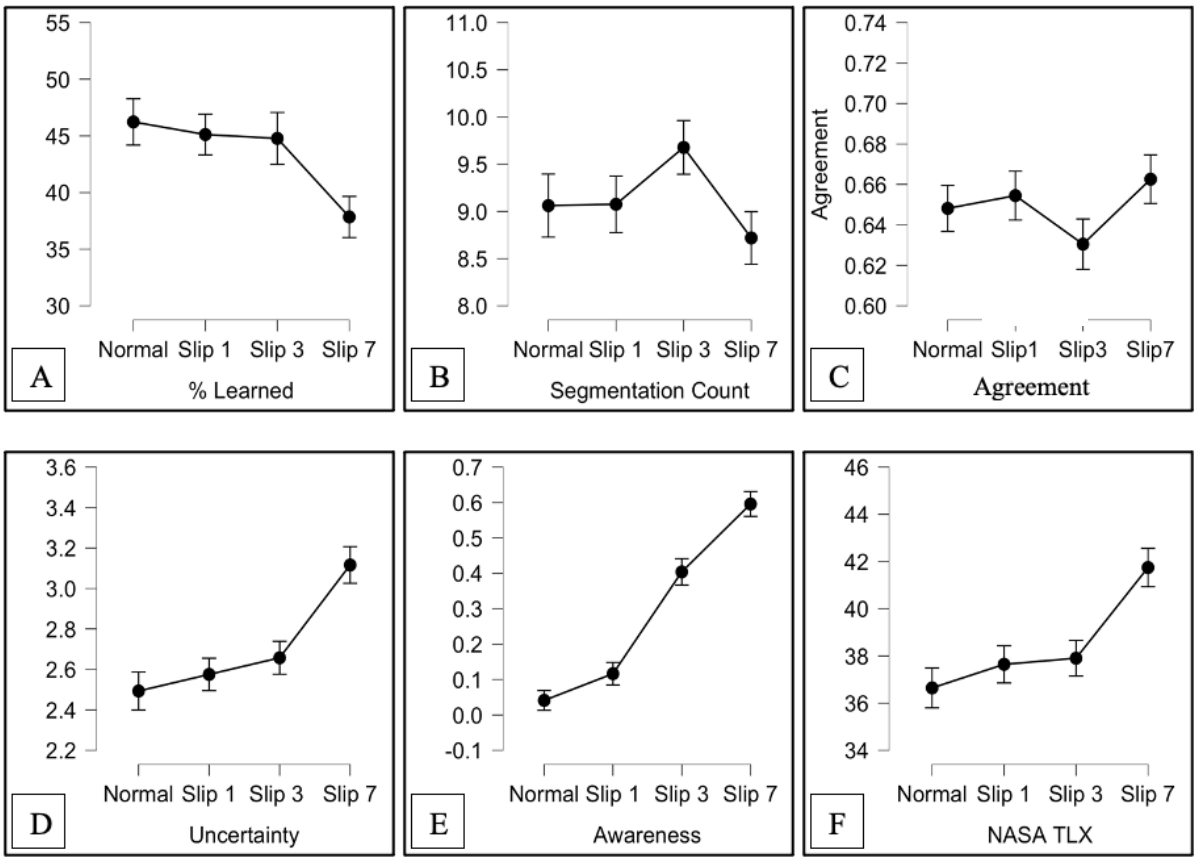

Figure 5. Results from all six measures of cognitive and perceptual processing: (A) Average learning scores calculated from post-test minus pre-test scores presented by condition. (B) Average number of time participants pressed "N" to represent a segmentation presented by condition. (C) Participant's average level of agreement in segmentation patterns by condition. (D) Average level of uncertainty while segmenting events (E) Awareness of temporal disruption by condition. (F) Average percieved workload by condition.

Event Segmentation

Segmentation Count. The effect of temporal displacement on segmentation count was not significant, $F(3, 216) = 1.768$, $p = 0.154$, $\eta^2 = 0.024$ (Figure 5B). Across all videos and all conditions (average duration of 3 minutes and 14 seconds), participants segmented 9.13 times on average, approximately 1 segment every 21 seconds.

Segmentation Agreement. There was no difference in segmentation agreement scores between conditions, $F(3, 600) = 1.30$, $p = 0.273$, $\eta^2 = 0.006$ (Figure 5C). Across all videos and all conditions, participant's average segmentation agreement score was 0.649 (on a 0 to 1 scale).

Segmentation Uncertainty

Temporal displacement significantly affected segmentation uncertainty, $F(3, 216) = 10.346$, $p < 0.001$, $\eta^2 = 0.126$ (Figure 5D). Post-hoc tests reveal watching a video that was temporally displaced 7 seconds caused significantly more uncertainty than all other conditions (0 second displacement: ($t = -5.086$, $p < 0.001$), 1 second displacement: ($t = -4.415$, $p < 0.001$), 3 second displacement: ($t = -3.744$, $p < 0.001$).

Disruption Awareness

The degree of temporal displacement had a significant effect on participants' awareness of disruption, $F(3, 216) = 59.89$, $p < 0.001$, $\eta^2 = 0.454$ (Figure 5E). Only a hand full of times did a participant become aware of disruption in the 1 second displacement condition (17 out of 146 trials). Similarly, a few participants had false alarms, reporting disruption in the normal condition (6 out of 146 trials). Approximately 40% of individuals noticed the disruption in the 3 second displacement and 60% in the 7 second displacement. Post-hoc proportion tests reveal 3 seconds displacements led to a significantly greater proportion of awareness compared to 0 second displacements ($z = -4.320$, $p = < 0.001$) and 1 second displacements ($z = -3.250$, $p < 0.001$). 7

seconds displacements led to a significantly greater proportion of awareness compared to all other conditions (0 second displacement: (z = -5.896, p < 0.001), 1 second displacement: (z = -4.962, p < 0.001), 3 second displacement: (z = -1.901, p = 0.0287). The proportion of awareness for 1 second displacements was not significantly greater than that of 0 second displacements (z = 1.38, p = 0.08). Participants false alarmed by reporting other abnormalities on occasion. As evident in Figure 6, the overall level of false alarms for the nonmanipulated disruptions was very similar to the number of reports of the manipulated temporal disruptions for the one-second videos, and that 3-second disruptions did begin to increase relative to false alarms.

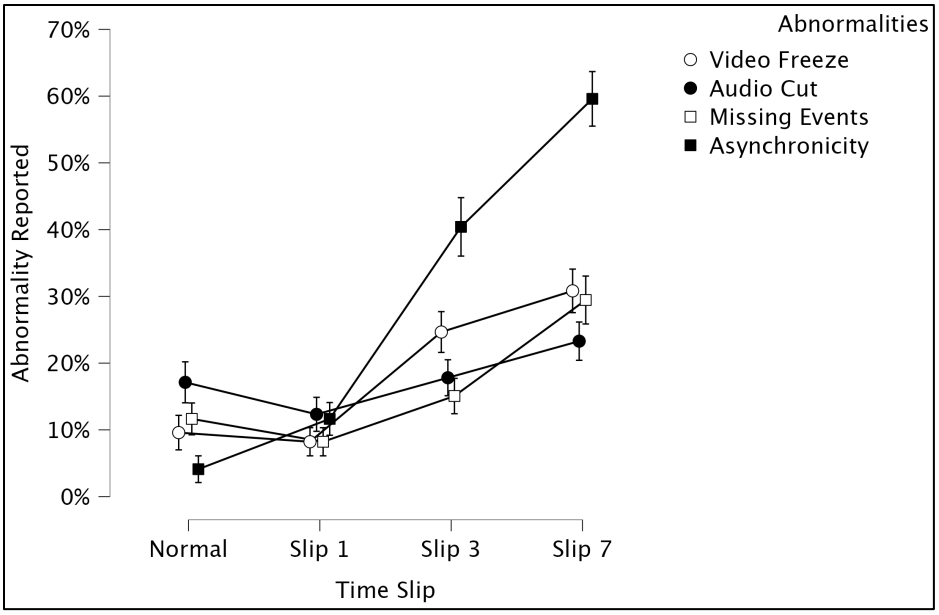

Figure 6. Amount of reports for each abnormality are presented. “Video Freeze”, “Audio Cut”, and “Missing Events” are all abnormalities that never took place.

Perceived Workload

The degree of displacement had a significant effect on participants’ perceived workload as measured by NASA TLX, $F(3, 216) = 7.835$, $p < 0.001$, $\eta^2 = 0.098$ (Figure 4F). Post-hoc

tests reveal watching a video that was temporally displaced 7 seconds created significantly greater perceived workload than all other conditions (0 second displacement: (t = -4.506, p < 0.001), 1 second displacement: (t = -3.623, p = 0.002), 3 second displacement: (t = -3.393, p = 0.003).

Experiment 2 Discussion

The consequences of disrupting the relationship between auditory and visual streams was very similar for the live action videos in Experiment 2 and the screen-captured videos in Experiment 1. As predicted, disruptions increased segmentation uncertainty, disruption awareness, and perceived workload. Also 7-second disruptions decreased learning. Event segmentation and segmentation agreement remained unaffected. Similar to Experiment 1, 7 second displacement was the only condition that significantly affected all of these cognitive and perceptual processes. The 3-second displacement impacted only awareness of disruptions.

CHAPTER IV

General Discussion

We observed very similar results using very different materials across two experiments that manipulated the temporal relationships between visual events and utterances. In both screen-captured videos, where verbalizations were associated with actions such as mouse movements and menu selections, and in live-action videos where verbalizations were associated with hand movements and gestures, there was no impact of one-second temporal disruptions, and only an effect of awareness for 3-second disruptions. There were more clear effects of 7-second disruptions, especially for disruption awareness, perceived workload, segmentation uncertainty, and in Experiment 2 for learning. A lack of impact from smaller disruptions suggests our cognitive and perceptual systems maintain a flexibility of multimodal integration that needs to be better accounted for in current theories of event perception. In other words, our results reveal an approximately 3-second event-integration window within which specific temporal expectations are not necessarily tracked.

While the 3-second temporal displacement reliably produced disruption awareness, we propose that the event-integration window extends to 3 seconds because the relatively low levels of post-hoc awareness could appear from just one detection of asynchronicity, and participants were likely to some degree on the lookout for “abnormalities”. Further, our conclusion of a 3-second event-integration window is consistent with work on the psychological present described previously (Pöppel, 2009). That said, it is likely that participants will be able to discriminate intra-event temporal relationships at shorter timeframes via focused attention, even for the

natural events we have explored here, with sufficient support and possibly repetition (Zmigrod & Hommel, 2011). For example, research by Shimamura and colleagues (2014) demonstrates that participants can detect small 1-2 frame mismatches in action overlap and ellipsis across movie edits with repeated scrutiny even though these differences appear undetectable with less repetition and scrutiny, as reviewed in the introduction.

The idea that a 3-second event-integration window will under most circumstances allow temporal variability may reflect a natural rhythm to the events that people must understand. One source for this cognitive rhythm may be that many brief temporal and sequential relationships are highly constrained by the basic mechanics of action (it is, for example, impossible to use a screwdriver before grabbing it). Accordingly, variations in these subsequences can normally go unencoded, unless the less-frequent need for scrutiny arises. This might be a highly efficient approach to event perception, especially in the context of phenomena such as attentional blink which has been explained as a temporally constrained cost whereby awareness in one moment prevents awareness briefly thereafter (e.g. Chun & Potter, 1995). Recently, evidence has suggested attentional blink is due to constraints in the encoding stages of visual short-term memory (Petersen & Vangkilde, 2022). The proposed event-integration window here may compliment this evidence in that the need for a window is due to the limitations of our perceptual processes. In other words, the temporal structure of a fine-grained event can be disrupted and not interfere with cognitive and perceptual processes, as long as the information stays within the psychological present.

Contrary to our hypotheses, event segmentation count and segmentation agreement measures produced nonsignificant downward trends as disruption increased in Experiment 1 and were largely unaffected by disruption in Experiment 2. These results may suggest that event

perception theories need to better account for pre-semantic processes involved in predictions and the possibility of an event-integration window. However, one might argue that the event-integration window we found occurs only because of the timescale of prediction error signal in participants. For example, less perceptual flexibility might be seen in videos in smaller timescales (i.e., >10 seconds) rather than our videos that were up to 4 minutes long. As EST states, integrating the prediction error signal at different time constants leads to segmentation at different timescales (Zacks, 2020). As a future direction, we aim to rerun this experiment while participants' eye movements and EEG are recorded. While our research question of interest focuses on temporal precision, most measures were collected from participants only after each video was viewed. Using temporally precise measures like eye movement and EEG could either validate our event-integration window or they could reveal neural signatures of temporal disruption impact within a 3-second window that support current theories of event perception. For instance, Simon and Wallace (2018) investigated multisensory integration utilizing stimuli that visually and auditorily presented an individual verbalizing the 'BA' syllable. They temporally displaced the audio and visual channels of the stimuli somewhere between 0ms up to 450ms. The authors found evidence suggesting increased theta power and alpha suppression may be markers of incongruity processing in multisensory temporal perception. This finding and others similar (Venskus & Hughes, 2021; London et al., 2022; Bhat et al., 2018) guide our motivation to investigate our current paradigm under EEG. Aside from EEG, eye tracking would allow us to infer participants' predictions via gaze patterns such as look-ahead fixations, and differences in these gaze patterns across conditions might reflect impacts of disruption. Look-ahead fixations may be a low-cost method of prediction which, if violated only produces another eye movement but no particular cognitive or perceptual processing alarm, leading to the lack of

disruption impact seen in our behavioral measures. Additional future directions include investigating whether perceptual flexibility in the event-integration window is constant or if it changes based on the current event in the psychological present. For example, participants may have more or less flexibility for events requiring deeper levels of information processing versus common sense or irrelevant events.

REFERENCES

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology: Human perception and performance*, *21*(1), 109.

Claus, B., & Kelter, S. (2006). Comprehending narratives containing flashbacks: Evidence for temporally organized representations. Journal of Experimental Psychology. Learning, Memory, and Cognition, 32, 1031–1044.

Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive research: principles and implications*, *3*(1), 1-12.

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? Psychological Science, 22(2), 243–252.

Fairhall, S. L., Albi, A., & Melcher, D. (2014). Temporal integration windows for naturalistic visual sequences. *PloS one*, *9*(7), e102248.

Flores, S., Bailey, H. R., Eisenberg, M. L., & Zacks, J. M. (2017). Event segmentation improves event memory up to one month later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1183.

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., & Prinz, W. (2007). Predicting point light actions in real-time. *Neuroimage, 36,* T22-T32.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., ... & Hasson, U. (2012). Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, *76*(2), 423-434.

Hommel B, Muesseler J, Aschersleben G, Prinz W. 2001. The Theory of Event Coding (TEC): A framework for perception and action planning. Behavioral & Brain Sciences 24(5):849–937 Describes TEC and reviews a wide range of empirical support.

Hymel, A., Levin, D. T., & Baker, L. J. (2016). Default processing of event sequences. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 235-246.

Jaeger, C.B., Little, J.W., & Levin, D.T. (2021). The prevalence and utility of formal features in YouTube screen-capture instructional videos. *Technical Communication*, 68(1), 56-72.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72-79.

Kurby, C. A., & Zacks, J. M. (2021). Priming of movie content is modulated by event boundaries. Journal of experimental psychology. Learning, memory, and cognition, 10.1037/xlm0001085.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906-2915.

Levin, D. T., Baker, L. J., Wright, A. M., Little, J. W., & Jaeger, C. B. (2022). Perceiving versus scrutinizing: Viewers do not default to awareness of small spatiotemporal inconsistencies in movie edits. *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication.

Petersen, A., & Vangkilde, S. (2022). Decomposing the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(8), 812.

Pöppel, E. (2009). Pre-semantically defined temporal windows for cognitive processing. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 364, 1887–1896.

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current opinion in behavioral sciences*, *17*, 133-140.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive science*, *31*(4), 613-643.

Rohenkohl, G., Cravo, A. M., Wyart, V., & Nobre, A. C. (2012). Temporal expectation improves the quality of sensory information. *Journal of Neuroscience*, *32*(24), 8424-8428.

Shimamura, A. P., Cohn-Sheehy, B. I., & Shimamura, T. A. (2014). Perceiving movement across film edits: A psychocinematic analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(1), 77.

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., Camarata, S., & Wallace, M. T. (2016). Keeping time in the brain: Autism spectrum disorder and audiovisual temporal processing. *Autism Research*, *9*(7), 720-738.

Tanenhaus, Michael K; Spivey-Knowlton, Michael J; Eberhard, Kathleen M; Sedivy, Julie C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.

Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, *64*, 105-123.

White, P.A. (2017). The three-second “Subjective Present”: A critical review and a new proposal. *Psychological Bulletin*, 143(7), 735-756.

Wiener, M., & Kanai, R. (2016). Frequency tuning for temporal perception and prediction. *Current Opinion in Behavioral Sciences*, *8*, 1-6.

Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the Alzheimer type. Psychology and aging, 21(3), 466.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind– brain perspective. *Psychological Bulletin*, 133, 273–293.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, *23*(12), 4057-4066.

Zacks J. M. (2020). Event Perception and Memory. *Annual review of psychology*, *71*, 165–191.

Zhou, H. Y., Cheung, E. F., & Chan, R. C. (2020). Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions. *Neuropsychologia*, *140*, 107396.

Zmigrod, S., & Hommel, B. (2011). The relationship between feature binding and consciousness: Evidence from asynchronous multi-modal stimuli. Consciousness and cognition, 20(3), 586-593.