AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 6, 1449–1463
https://doi.org/10.1037/xge0001577

Evidence for an Event-Integration Window: A Cognitive Temporal Window Supports Flexible Integration of Multimodal Events

Madison Lee and Daniel T. Levin
Department of Psychology and Human Development, Vanderbilt University

Just as the perception of simple events such as clapping hands requires a linkage of sound with movements that produce the sound, the integration of more complex events such as describing how to give an injection requires a linkage between the instructor's utterances and their actions. However, the mechanism for integrating these complex multimodal events is unclear. For example, it is possible that predictive temporal relationships are important for multimodal event understanding, but it is also possible that this form of understanding arises more from meaningful causal between-event links that are temporally unspecified. This latter approach might be supported by a cognitive temporal window within which multimodal event information integrates flexibly with few default commitments about specific temporal relationships. To test this hypothesis, we assessed the consequences of disrupting temporal relationships between instructors' actions and their speech in both narrated screen-capture instructional videos (Experiment 1) and live-action instructional videos (Experiment 2) by displacing the audio channel forward or backward relative to the video by 0, 1, 3, or 7 s. We assessed learning, event segmentation, disruption awareness, segmentation uncertainty, and perceived workload. Across two experiments, 7-s temporal disruptions consistently increased uncertainty and workload and decreased learning in Experiment 2. None of these effects appeared for 3-s disruptions, which were barely detectable. One-second disruptions produced no effects and were undetectable, even though much intraevent information falls within this range. Our results suggest the presence of an event-integration window that supports the integration of events independent of constraining temporal relationships between subevents.

Public Significance Statement

To perceive complex events, we must be able to integrate visual information such as people's actions or gestures with corresponding auditory information such as speech. Although these two forms of information are mutually supportive, it is not clear whether the precise temporal relationship between these streams is perceptually and cognitively important. These experiments demonstrate that, within a several-second window, the temporal relationship between these modalities can be disrupted without interfering with effective event perception and understanding. We demonstrate that cognitive integration of multimodal events is temporally flexible, and this may support forms of event understanding that are robust over small variations in event synchronization and temporal attention.

Keywords: event perception, learning, multimodal integration, psychological present

Effective perception and understanding of real-world events often require 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

This article was published Online First April 4, 2024.

Madison Lee https://orcid.org/0000-0001-6395-0976

Results from Experiment 1 were presented as a poster at the Psychonomic Society's Annual Conference in 2021. The authors would like to thank Eric Hall at Vanderbilt's School of Nursing for contributing to the creation of our live-action instructional videos. These studies were not preregistered. The data and materials are publicly available (https://osf.io/x2cm3/).

Madison Lee served as lead for data curation, formal analysis, project administration, software, visualization, and writing–original draft, contributed equally to investigation, and served in a supporting role for resources. Daniel T. Levin served as lead for resources and supervision and served in a supporting role for formal analysis and writing–original draft. Madison Lee and Daniel T. Levin contributed equally to conceptualization, writing–review and editing, and methodology.

Correspondence concerning this article should be addressed to Madison Lee, Department of Psychology and Human Development, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203-5721, United States. Email: madison.j.lee@vanderbilt.edu

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 et al., 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, on the one hand, some theories of event perception at least imply that specific interevent 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 interevent 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 (Zacks et al., 2007) imply the necessity of temporal expectations for effective event perception. According to event segmentation theory, 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 (Flores et al., 2017; Kurby & Zacks, 2008). Specifically, participants who provide relatively normative segmentation patterns tended to remember events better (Kurby & Zacks, 2011), and segmenting (vs. not segmenting) has been shown to improve event memory (Flores et al., 2017). Importantly, predictions are described in event segmentation theory as inherently sequential and forward-looking because they rely on comparisons between represented events that have already happened and the event that is currently being perceived. For many events, predictions are developed within one modality and confirmed via another modality, creating distinct cognitive demands for meaningful intermodal predictions and comparisons. These intermodal predictions are associated with meaningful temporal information that at least has the potential to be incorporated into the processing stream. For example, students viewing a computer instructional video might hear an instructor say, "Next, select cells A5 to A20" and then see a visual selection action. In this example, several temporal parameters are inherent in any prediction that the participant might make. First, the visual selection should occur in a specific location and after the verbalization because the word "next" has explicitly foreshadowed the selection action. Second, the selection should occur relatively quickly after the verbalization because it is simple. In contrast, an action that requires more thought to implement might reasonably take longer to initiate, and this delay might constitute valuable temporal information signaling the instructor's mental states, which could in turn aid in understanding the lesson. This would be an analog of the role speaker disfluencies play in helping listeners understand speech content by highlighting the relative complexity of different moments in their partner's speech stream (e.g., Fraundorf & Watson, 2011).

Several lines of behavioral and neurophysiological evidence further reinforce the hypothesis that temporal information is central to event perception and downstream cognitive processes (Reynolds et al., 2007; Zacks et al., 2011). For example, Eisenberg et al. (2018) propose that movie viewers who look to objects that actors are about to interact with 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 continue 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). Furthermore, 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). Other research demonstrates that very precise temporal information can impact visual processing. For example, Graf et al. (2007) found that participants projected the motion of point-light walkers forward during a timed occlusion, as indicated by priming for targets that precisely matched the forward-projected configuration.

While most would agree that temporal expectations are at least in some cases included in 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 above 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. Each movie included an average of 11 shots, and for some movies, one of the actions was presented in reverse order (for example, the shot depicting the use of the screwdriver was shown before the shot depicting grabbing it). The mean duration of the reversed shot was between 300 and 1,000 ms, so the reversal lasted up to 2 s. 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. Also, participants never detected the reversals when they were asked to attend closely to the movies but were not specifically instructed to look for them.

In the context of multimodal learning, research shows that memory improves with the presentation of congruent multimodal information compared to unimodal information (Meitz et al., 2020; Meyerhoff & Huff, 2016). Furthermore, the benefits of multimodal information seem to be widespread and flexible as event memory is surprisingly robust against violations of audiovisual synchrony. Meyerhoff and Huff (2016) presented participants with short film clips (3–4 s) that were either unmanipulated or the playback of the

visual track was reversed. Participants demonstrated no significant differences in recognition memory across conditions despite the latter condition's obvious temporal mismatch. Additionally, participants' memory was unaffected when multimodal information was presented sequentially compared to simultaneously (e.g., the audio track was presented immediately after the visual track, or vice versa; Meyerhoff & Huff, 2016). Broadly, these results suggest that semantic congruency rather than temporal synchrony contributes to multimodal superiority over unimodal information with regard to event memory. However, in a final experiment, Meyerhoff and Huff (2016) did find significant decreases in memory performance when audio and visual tracks were completely randomized across videos, demonstrating that this robustness against audiovisual asynchrony does have limitations.

Findings such as these suggest that audiovisual temporal mismatches within events, and even the temporal sequence between short adjacent events, may not have consistent cognitive impacts. Broadly, these findings may be consistent with the longstanding idea of a "psychological present" consisting of a several-second window within which time can be "immediately perceived" (James, 1892). More recently, Pöppel (2009) summarized empirical work from movement control, spontaneous speech, and auditory and visual processing that all provide evidence consistent with the hypothesis that conscious activity is presemantically integrated within 2–3-s windows. Fairhall et al. (2014) extended these ideas to naturalistic event perception by presenting participants with 13-s silent movie clips. The clips were divided into chunks, varying in duration from less than 1 s to the full 13 s representing the entire clip. Then, each chunk was further divided into 1/4 s intervals, and the intervals were scrambled within the chunks. Participants rated how difficult the scrambled clips were to follow. Video clips that were scrambled within 2-s chunks were reportedly easy to follow, but as chunk duration increased, the participant's difficulty rating increased considerably. Fairhall et al. (2014) concluded that the increased difficulty derived from the fact that scrambling within the larger chunks displaced action information beyond the psychological present.

The inherent temporal leads and lags between gestures and their accompanying speech may serve as a multimodal reflection of the psychological present. This hypothesis is supported by evidence indicating that the precise timing of gesture–speech pairs may not be cognitively impactful. Anikin et al. (2015) presented participants with multimodal instructional videos that were either unmanipulated or the audio track was displaced from the video track by ± 1.5 s. These videos portrayed an instructor's hand gestures as they explained how to form a geometric shape with presented materials. After watching the videos, participants were tasked with reconstructing the shape described in the video. The ± 1.5-s temporal disruption did not impact participants' performance or their perceived task difficulty. Relatedly, Kirchhof (2014) found 60% of participants perceived gesture–speech pairs that had delays between −600 and +600 ms as natural and synchronous. Additionally, when asked to choose the optimal synchronization position for auditory and visual tracks, participants' selection varied from −1.8 to +1.2 s. Together, these results support the possibility of a larger event-integration window.

Combining gesture–speech lag findings, the temporal nondiscrimination findings, and the idea of presemantic integration implies that events within the psychological 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 event segmentation theory. 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 (e.g., Graf et al., 2007). It is therefore possible that evidence for an event-integration window characterized by nondiscrimination of temporal information is generated only from limited situations. One important limit of previous research is that it often relies on measures of visual recognition memory, but a true cognitive temporal window should constrain a broader spectrum of cognitive and perceptual processes. Moreover, the "learning" contexts presented in these studies often lack clear real-world applicability and ecological validity. 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. In addition, most of the above studies only tested a limited range of displacements, and no study has parametrically manipulated displacements to demonstrate a transition between small displacements that do not impact processing and larger ones that do.

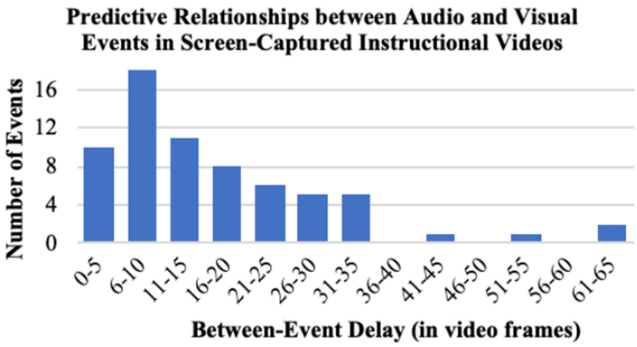
Current Study and Hypotheses

To test for evidence of a multimodal event-integration window in situations involving relatively deep processing, we disrupted the temporal relationship between a model'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 s relative to the video. So, a 1-s forward displacement would entail moving the instructor's speech 1 s earlier relative to the actions they are producing, and a 1-s backward displacement would move the speech to be 1-s 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 (such as hand gestures and facial expressions), it is an extremely common learning setting that viewers find comfortable, as evidenced by the literally billions of views these videos receive (Jaeger et al., 2021).

To better understand temporal information in multimodal events, we coded 12 min of screen-captured instructional video for the frequency and timing of predictive relationships between audio and video channels (Figure 1). For example, in one of our videos, an instructor says, "Now let's create a new Excel sheet" (prediction-generating action). Twenty-two frames after "sheet" is 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 four frames after "sort" is verbalized, the instructor presses the "Sort" button. The frequency distribution shown in Figure 1 suggests that substantial variation in intraevent information falls within the range of 1 s (∼30 video frames).

Figure 1

Distribution of Between-Event Delay Durations (in Video Frames) Between the Audio and Video Channel of an Event



Note. Six different screen-captured instructional videos were coded. See the online article for the color version of this figure.

In the current experiments, participants watched eight screen-captured instructional videos, each varying in their level of disruption. 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 that temporally disrupting multimodal relationships would have negative consequences on event perception and cognitive processes. It is important to note that the instructional videos in the current experiments are naturalistic and were shot without any attempt at constraining whether audio information predicted video information and to what degree. Consequently, the impact of a temporal disruption depends not only on the direction of the disruption (forward vs. backward) but also on the magnitude of the disruption (0, 1, 3, and 7 s), the natural multimodal order (audio naturally leading or lagging video for an event), and the natural delay duration.

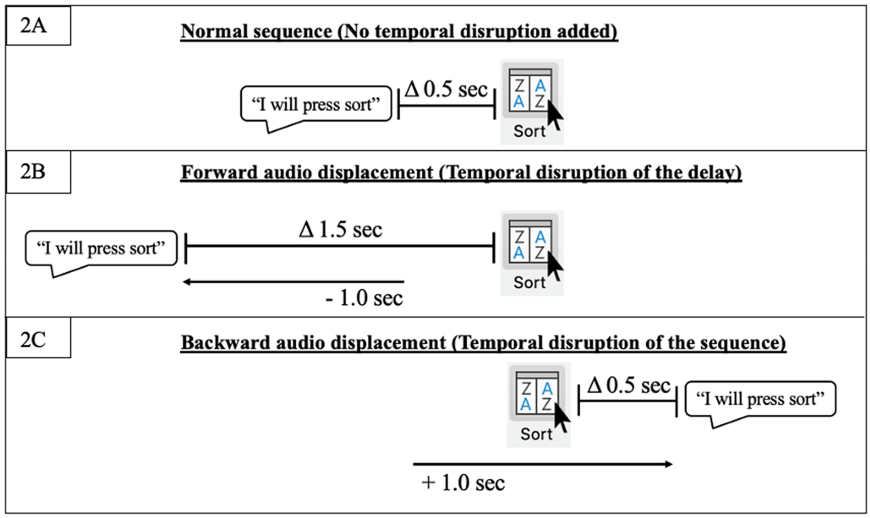
Figure 2 illustrates the possible impacts of a temporal disruption, but these illustrations are not exhaustive. Figure 2B and 2C illustrate an example of auditory information predicting visual information. Figure 2B presents a disruption of the natural temporal delay where the displacement causes the event's audio information to occur 1.5 s before the paired visual information instead of 0.5 s before. Figure 2C presents a disruption of the natural temporal sequence where the displacement causes the event's audio information to occur 0.5 s after the paired visual information instead of 0.5 s before.

In contrast, Figure 3 demonstrates how the magnitude of the displacement, the natural multimodal sequence, and the natural delay duration impact the unique effect a forward and backward displacement have on each event. While in Figure 2, the forward displacement increases the natural temporal delay, Figure 3 presents an instance of the forward displacement disrupting the sequence of the delay. These instances are provided to demonstrate the unique impact a displacement has on every event and to justify why we would not expect any significant difference between forward and backward displacements of equal magnitude for the videos used in the present experiment.

Grounded by the research described above, we hypothesize that only disruptions beyond the psychological present will significantly impact participants' cognitive and perceptual measures. Only as temporal disruption increases to 7 s should we observe an increase in prediction error and event model updating, which should be

Figure 2

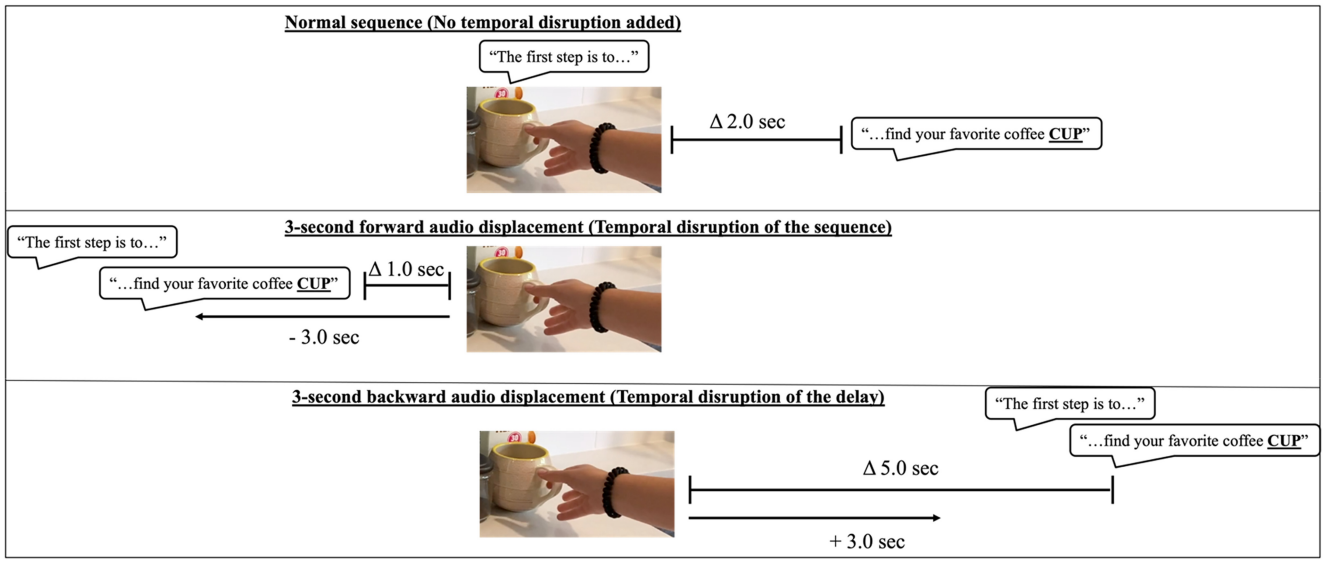
How a 1-s Displacement Uniquely Disrupts Each Inter-Event Relationship in an Instructional Video



Note. (2A) In this example, the natural delay between an event's auditory information and its corresponding visual information is 0.5 s with the auditory information preceding the visual information. (2B) Applying a 1 s forward displacement to the auditory information leads the auditory information to now take place 1.5 s prior to the visual information. (2C) Applying a 1 s backward displacement to the auditory information leads the auditory information to now take place 0.5 s after the visual information. It is evident that applying a 1 s displacement to an entire instructional video uniquely disrupts each inter-event relationship. See the online article for the color version of this figure.

Figure 3

How a 3-s Displacement Uniquely Disrupts Each Inter-Event Relationship in an Instructional Video



Note. In this example, the natural delay between an event's auditory information and its corresponding visual information is 2 s with the visual information preceding the auditory information. Applying a 3 s forward displacement to the auditory information leads the auditory information to now take place 1 s prior to the visual information. Applying a 3 s backward displacement to the auditory information leads the auditory information to now take place 5 s after the visual information. See the online article for the color version of this figure.

represented by an increase in event segmentation and a decrease in participants' segmentation agreement scores. We also predict a decrease in learning for 7-s disruptions as ineffective event segmentation has been shown to negatively affect learning and memory (Flores et al., 2017). Additionally, just as Fairhall et al. (2014) saw an increase in difficulty measures for temporal disruptions lasting beyond the psychological present, we hypothesize a 7-s disruption will cause a spike in segmentation uncertainty and perceived workload. Finally, even though most intraevent temporal relationships vary within a range of less than 1 s (Figure 1), if they are not normally monitored, we hypothesize participants will remain unaware of the briefest displacements.

Experiment 1

In this experiment, participants watched eight screen-capture instructional videos about various topics in Microsoft Excel and Microsoft Paint. Participants watched each video twice and were tasked with learning as much as they could during the first viewing and segmenting the video into meaningful events during the second viewing. The audio and video channels of each video were offset at varying magnitudes (0, 1, 3, or 7 s) to disrupt temporal information existing between modalities. To investigate the possibility of a multimodal event-integration window, cognitive and perceptual consequences of this disruption were assessed via measures of learning, event segmentation, disruption awareness, and task load.

Method

Transparency and Openness

In line with Transparency and Openness guidelines, this article specifies how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Experiments 1 and 2 data and materials are publicly available on Open Science Framework (https://osf.io/x2cm3/). Data were analyzed using JASP, Version 0.16.2. Study designs and analysis plans were not preregistered for Experiment 1 or 2.

Participants

Sixty participants from Vanderbilt University's undergraduate participant pool completed Experiment 1 online in 2021. Eleven participants failed the instruction check and were excluded from analyses, leaving 49 participants in analyses. The participant's average age was 19.5 years old. Twenty-two reported as female, 26 as male, and one preferred not to answer. Participants reported their gender identities using a drop-down menu with the following possible options: "Male," "Female," "Transgender male-to-female," "Transgender female-to-male," "Gender variant/non-conforming," "Prefer to self-describe:," or "Prefer not to answer." No a priori power analysis was performed; the sample size was determined based on the amount of data that could be collected in the available timespan. This experiment was not preregistered. Data and materials are publicly available (https://osf.io/x2cm3/).

Videos

Participants watched eight screen-captured instructional videos where an off-screen narrator explained and demonstrated various topics in Microsoft Excel and Microsoft Paint. Lessons in Excel explained 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 explained 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 s, ranging from 50 to 98 s 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 s. 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 eight videos: two unmanipulated videos (0-s displacement), two 1-s displaced videos (one forward displaced, one backward displaced), two 3-s displaced videos, and two 7-s displaced videos. Participants were randomly assigned one of eight possible block orders, which counterbalanced the degree to which each video was temporally displaced.

Considering this 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 eight videos, two times each. They were instructed to learn as much as they could during their first viewing and 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. This question 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 not only participants' comprehension of the tasks demonstrated but also their memory for the sequence in which events occurred. For example, a comprehension question about the transparent selection video was "Based on this video, when should transparent selection be turned on?." 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?." 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 pretest, participants practiced segmenting events on a novel, unmanipulated video. Participants were told, "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 1 min 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 5-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)?" ($1 = not\ at\ all\ uncertain$ to $5 = very\ uncertain$). Participants then responded to the National Aeronautics and Space Administration's Task Load Index (NASA-TLX; Hart & Staveland, 1988), a questionnaire that measures an individual's perceived workload. The questionnaire has six scales (mental demand, physical demand, temporal demand, performance, effort, and frustration). We used all but the physical demand scale. Finally, participants answered the exact same eight multiple-choice content questions they had seen in the pretest. This procedure was repeated for all eight videos.

Results

Learning

We assessed learning by subtracting each participant's pretest score from their posttest score. Although there was a downward trend of learning as disruption increased, the effect of temporal disruption on learning scores was nonsignificant, $F(3, 144) = 1.37$, $p = .25$, $\eta^2 = .028$ (Figure 4A). Across all conditions, on average participant's posttest scores were 39% higher than their pretest scores, suggesting a high level of learning, $t(48) = 15.73$, $p = < .001$, $d = 2.25$.

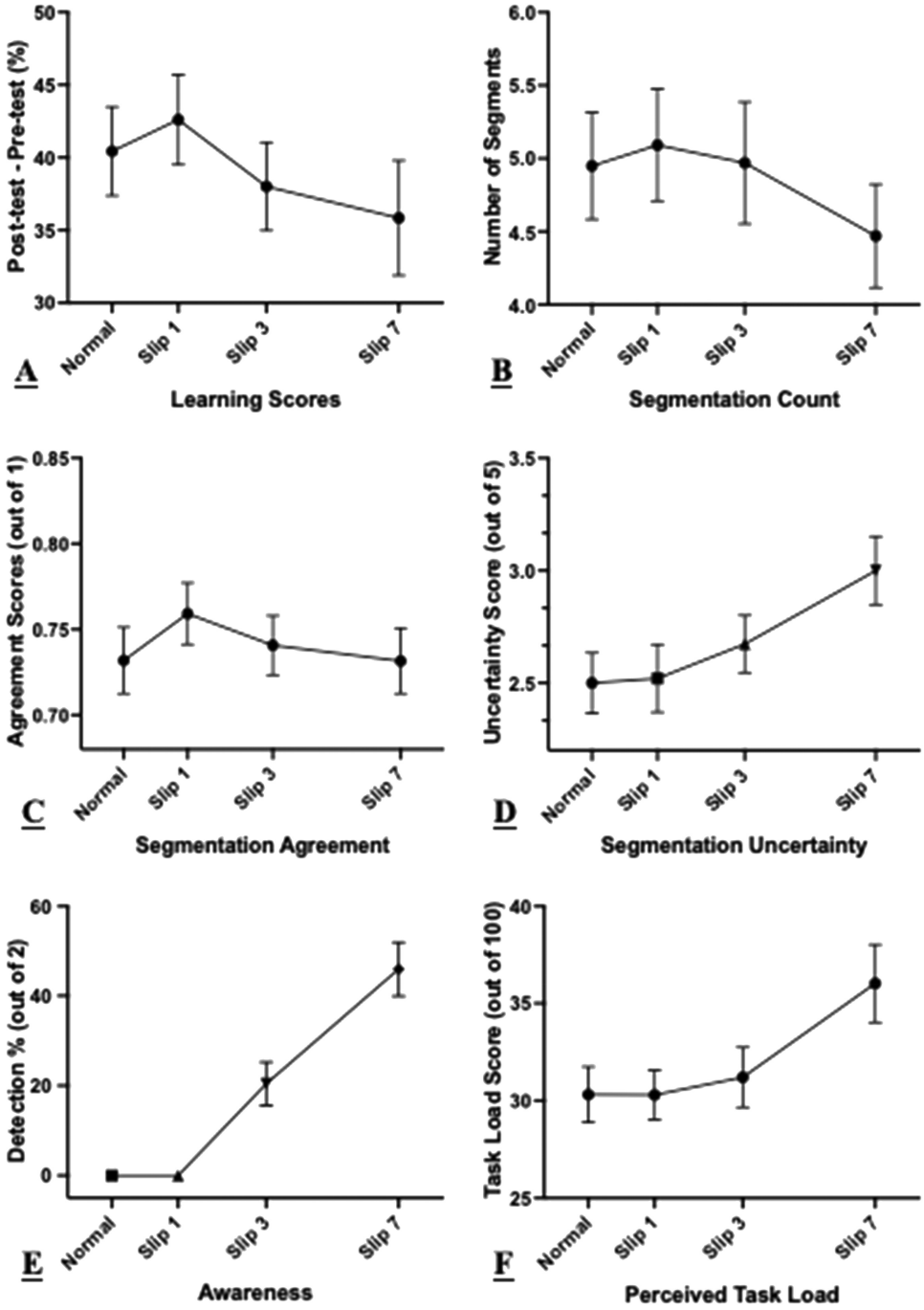
Table 1 presents Bayes factors in favor of the null over alternative hypotheses when comparing displacement conditions for each measure. The Bayes factor analysis was run with a Cauchy prior of 0.707, the default prior in JASP (Morey & Rouder, 2015). This table shows that for each comparison between normal, Slip 1, and Slip 3 conditions (other than disruption awareness), BF_{01} is between 2.3 and 6.4. These values signify moderate evidence for the null hypothesis; these conditions are unlikely to differentially affect perception and cognition.

Event Segmentation

Segmentation Count. 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, $F(3, 144) = 2.15$, $p = .10$, $\eta^2 = .04$ (Figure 4B). The difference between the 7-s disrupted videos and the undisrupted videos was nonsignificant correcting for multiple comparisons with Holm–Bonferroni, $t(48) = -1.81$, $p = .30$, $d = 0.18$. Across all videos and all conditions (average duration of 70 s), participants segmented 4.87 times on average, approximately one segment every 14 s.

Figure 4

Results From All Six Measures of Cognitive and Perceptual Processing



Note. (A) Average learning scores calculated from posttest minus pretest 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 perceived workload by condition. Error bars show the *SEM*.

Segmentation Agreement. Segmentation agreement was calculated using a point biserial correlation across 1-s 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). Comparison groups varied in size from four to eight individuals per video ($M = 6.15$). Each participant received an agreement score

Table 1
Bayesian Paired Samples t Tests

Outcome measure	Condition 1	Condition 2	BF_{01}	Error (%)
Learning score	Normal	Slip 1	5.26	0.06
	Normal	Slip 3	5.19	0.06
	Normal	Slip 7	3.43	0.05
	Slip 1	Slip 3	2.31	0.04
	Slip 1	Slip 7	1.25	0.03
	Slip 3	Slip 7	5.44	0.06
Segmentation uncertainty	Normal	Slip 1	6.26	0.07
	Normal	Slip 3	2.39	0.04
	Normal	Slip 7	0.14	2.88×10^{-7}
	Slip 1	Slip 3	3.38	0.05
	Slip 1	Slip 7	0.25	5.86×10^{-7}
	Slip 3	Slip 7	0.40	0.01
Perceived load	Normal	Slip 1	6.44	0.07
	Normal	Slip 3	5.14	0.06
	Normal	Slip 7	0.16	3.32×10^{-7}
	Slip 1	Slip 3	4.77	0.06
	Slip 1	Slip 7	0.10	1.84×10^{-7}
	Slip 3	Slip 7	0.04	5.64×10^{-8}
Segment count	Normal	Slip 1	5.72	0.06
	Normal	Slip 3	6.42	0.07
	Normal	Slip 7	0.79	0.02
	Slip 1	Slip 3	5.74	0.06
	Slip 1	Slip 7	0.75	0.02
	Slip 3	Slip 7	1.53	0.03
Disruption awareness	Slip 3	Slip 7	7.26×10^{-4}	2.05×10^{-7}

Note. BF_{01} values greater than 3.0 indicate moderate evidence for the null hypothesis where the observed data are 3 times more likely to fall under the null. BF_{01} values less than 0.33 indicate moderate evidence for the alternative hypothesis (van Doorn et al., 2021). BF_{01} values could not be calculated for comparisons involving normal or Slip 1 disruption awareness because their variances equal zero. BF = Bayes factor.

for each video they watched. Averaging over all videos, there was no difference in segmentation agreement scores between conditions, $F(3, 144) = 0.51$, $p = .67$, $\eta^2 = .004$ (Figure 4C). Across all videos and all conditions, the participant's average segmentation agreement score was 0.74 (on a 0–1 scale).

Segmentation Uncertainty

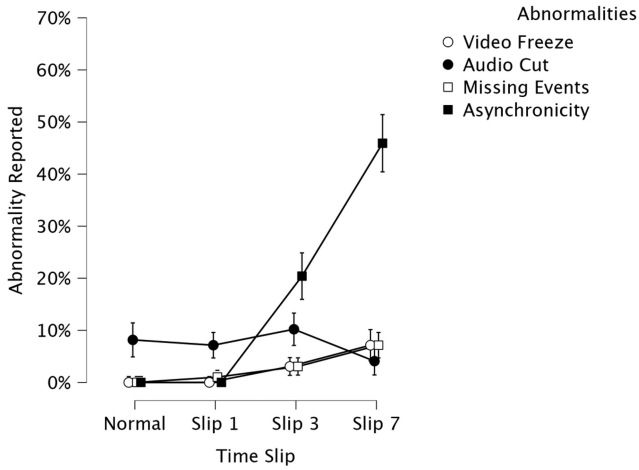
Displacement significantly affected segmentation uncertainty, $F(3, 144) = 5.55$, $p = .005$, $\eta^2 = .10$ (Figure 4D). Holm–Bonferroni corrected post hoc comparisons reveal watching a video that was temporally displaced 7 s caused significantly more uncertainty than 0 s temporally displaced videos, $t(48) = -3.60$, $p = .003$, $d = -0.51$, and 1 s temporally displaced conditions, $t(48) = -3.46$, $p = .004$, $d = -0.48$.

Disruption Awareness

Temporal displacement significantly affected participants' awareness of disruption, $F(3, 144) = 39.07$, $p < .001$, $\eta^2 = .45$ (Figure 4E). Not once did a participant report awareness of disruption in the 0-s or 1-s displacement condition. Participants noticed approximately 20% of the 3-s displacements and 45% of the 7-s displacements. One-sample t tests revealed that mean awareness for 7-s and 3-s displacement conditions was significantly greater than zero, respectively, $t(48) = 7.77$, $p < .001$, $d = 1.10$; $t(48) = 4.24$, $p < .001$, $d = 0.61$. Seven-second displacements caused a significantly greater proportion of awareness compared to 3-s displacements, $t(48) = -5.17$, $p < .001$, $d = -0.95$.

To provide evidence about the degree to which detection of asynchrony was more prevalent than false detection of other abnormalities. And the degree to which detection of asynchronicity might, in fact, be reflected in misattributions to other abnormalities, Figure 5 presents positive responses for all of the possible abnormalities surveyed. As the figure demonstrates, participants in most cases did not false alarm by reporting the other three abnormalities, although false alarms for audio cutouts occurred in approximately 10% of trials. However,

Figure 5
Prevalence of Reports for Each Abnormality



Note. "Video freeze," "audio cut," and "missing events" are all abnormalities that never took place.

the prevalence of audio cutout false alarms clearly did not increase for increasing amounts of asynchrony.

Perceived Workload

The degree of temporal displacement significantly affected participants' perceived workload, as measured by NASA-TLX, $F(3, 144) = 7.04$, $p = .002$, $\eta^2 = .13$ (Figure 4F). Post hoc tests using Holm–Bonferroni correction reveal watching a video that was temporally displaced 7 s created significantly greater perceived workload than all other conditions: 0-s displacement, $t(48) = -3.90$, $p < .001$, $d = -0.51$; 1-s displacement, $t(48) = -3.92$, $p < .001$, $d = -0.52$; and 3-s displacement, $t(48) = -3.30$, $p = .005$, $d = -0.43$.

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-s disruptions had any impact. When asked, participants were able to detect 3-s disruptions, but only in about 20% of cases. Otherwise, 3-s displacements produced no significant impacts. In no case did 1-s disruptions produce any effects. 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-s displacements to result in fewer segmentations.

Experiment 2

Experiment 1 suggests that temporal event integration may be flexible within a 1–3-s window. However, it is possible that screen-capture instructional videos uniquely encourage this flexibility. For example, the visual events associated with language in these videos may be unnatural and therefore would not benefit much from 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 task 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

Eighty-three participants from Vanderbilt University's undergraduate participant pool completed Experiment 2 online in 2021. Ten participants failed the instruction check and were excluded from analyses, leaving 73 participants in analyses. The participant's average age was 18.9 years old. Forty-four reported as female, 28 as male, and one reported as nonbinary. No a priori power analysis was performed; the sample size was determined based on the amount of data that could be collected in the available timespan. This experiment was not preregistered. Data and materials are publicly available (https://osf.io/x2cm3/).

Videos

Participants watched eight live-action instructional videos in which a narrator demonstrated various tasks. Lessons included administering a vaccine, inserting an intravenous tube, making a latte, setting up a sewing machine, setting up a board game, crafting a cocktail, potting a plant, and playing a card game (Figure 6). The average video duration was 3 min and 14 s, ranging from 63 s to 4 min. The process of editing videos to create each temporally displaced condition was identical to Experiment 1. These were essentially single-shot videos. The only additional editing occurred when there were long moments of no relevant audio or video (e.g., 20 s of milk frothing). In these moments, we speeded the video 8× and removed the audio. While filming these videos, we did not include the instructor's face to avoid participants noticing disruptions only because instructor's 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

Learning

The degree of temporal displacement significantly affected learning scores, $F(3, 216) = 3.65$, $p = .013$, $\eta^2 = .048$ (Figure 7A). Post hoc tests using Holm–Bonferroni correction revealed that watching a 7-s displaced video caused significantly less learning than watching a 0-s displaced video, $t(72) = 2.97$, $p = .020$, $d = 0.34$.

Table 2 presents the Bayes factors in favor of the null over alternative hypotheses when comparing displacement conditions for each measure. This Bayes factor analysis was run with a Cauchy prior of 0.707, the default prior in JASP (Morey & Rouder, 2015). This table shows that for each comparison between normal, Slip 1, and Slip 3 conditions (other than comparisons of disruption awareness), the BF_{01} is between 2.5 and 7.7. These values signify moderate evidence for the null hypothesis; these conditions are unlikely to differ in their effects on our measures of perception and cognition.

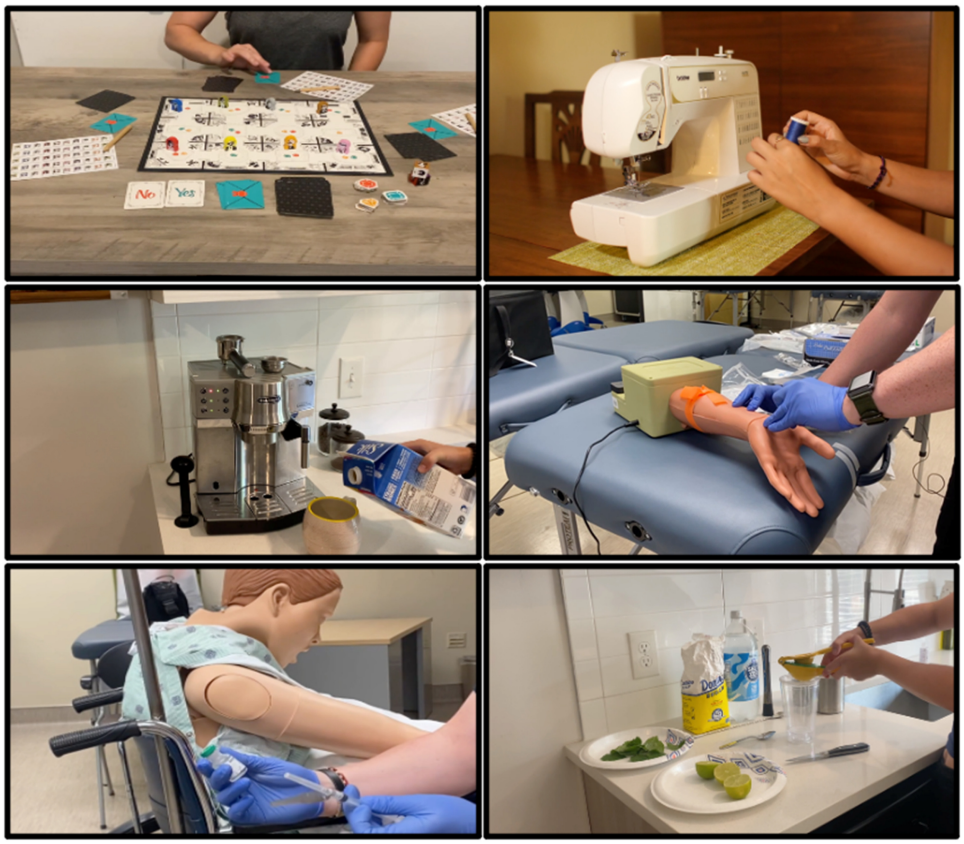
Event Segmentation

Segmentation Count. The effect of temporal displacement on segmentation count was not significant, $F(3, 216) = 1.77$, $p = .15$, $\eta^2 = .024$ (Figure 7B). Across all videos and all conditions (average duration of 3 min and 14 s), participants segmented 9.13 times on average, approximately one segment every 21 s.

Segmentation Agreement. There was no difference in segmentation agreement scores between conditions, $F(3, 600) = 1.30$, $p = .27$, $\eta^2 = .006$ (Figure 7C). Across all videos and all conditions, the participant's average segmentation agreement score was 0.65 (on a 0–1 scale).

Figure 6

Video Frames From Six of Our Live-Action Instructional Videos



Note. See the online article for the color version of this figure.

Segmentation Uncertainty

Temporal displacement significantly increased segmentation uncertainty, $F(3, 216) = 10.35$, $p < .001$, $\eta^2 = .13$ (Figure 7D). Post hoc tests using Holm–Bonferroni correction reveal watching a video that was temporally displaced 7 s caused significantly more uncertainty than all other conditions: 0-s displacement, $t(72) = -5.09$, $p < .001$, $d = -0.67$; 1-s displacement, $t(72) = -4.42$, $p < .001$, $d = -0.58$; and 3-s displacement, $t(72) = -3.74$, $p < .001$, $d = -0.49$.

Disruption Awareness

The degree of temporal displacement significantly affected participants' awareness of disruption, $F(3, 216) = 59.89$, $p < .001$, $\eta^2 = .45$ (Figure 7E). Participants only rarely reported the 1-s disruptions (17 out of 146 trials), and these reports were not significantly more frequent than the false alarms in the no-displacement condition, six out of 146 trials, $t(72) = 1.60$, $p = .11$. Participants noticed about 40% of the 3-s displacements and 60% of the 7-s displacements. Post hoc tests using Holm–Bonferroni correction reveal 3-s displacements caused a significantly greater proportion of awareness compared to 0-s displacements, $t(72) = -7.72$, $p = < .001$, $d = -1.13$, and 1-s displacements, $t(72) = -6.12$, $p < .001$, $d = -0.89$. Seven-second displacements caused a significantly greater proportion of awareness compared to all other conditions: 0-s displacement, $t(72) = -11.80$, $p < .001$, $d = -1.73$; 1-s displacement, $t(72) = -10.20$, $p < .001$, $d = -1.49$; and 3-s displacement, $t(72) = -4.08$, $p < .001$, $d = -0.60$. In addition, asynchronicity reports were more frequent than reports of other abnormalities for both the 3-s condition and the 7-s condition, 3-s condition: asynchrony versus video freeze, $t(72) = 3.01$, $p = .004$, $d = 0.35$, versus audio cut, $t(72) = 4.46$, $p < .001$, $d = 0.52$, versus missing event, $t(72) = 5.29$, $p < .001$, $d = 0.62$; 7-s condition: asynchrony versus video freeze, $t(72) = 5.69$, $p < .001$, $d = 0.67$, versus audio cut, $t(72) = 6.66$, $p < .001$, $d = 0.78$, versus missing event, $t(72) = 5.32$, $p < .001$, $d = 0.62$.

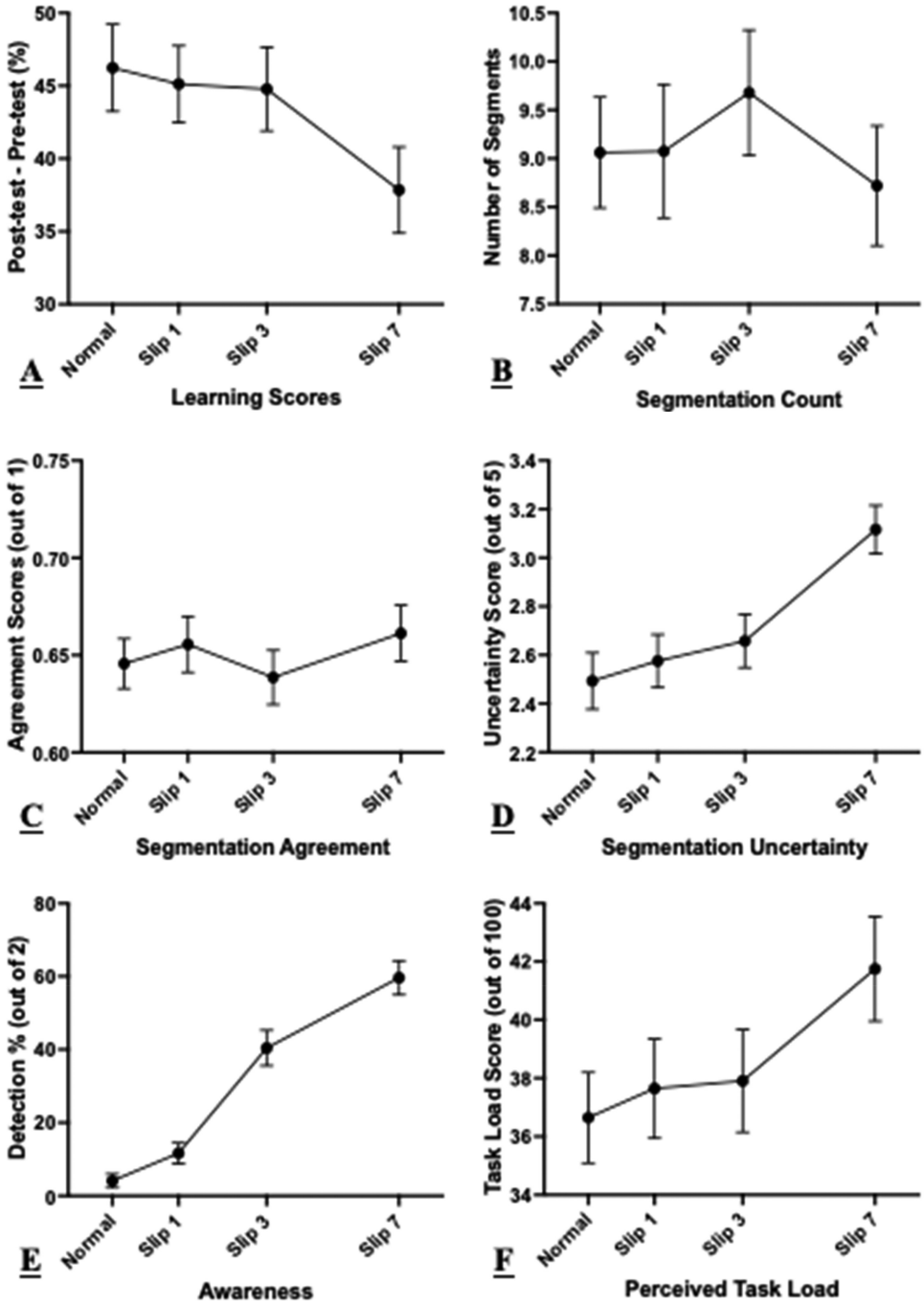
It is interesting to note that participants false alarmed by reporting nonpresent abnormalities more frequently than in Experiment 1. As evident in Figure 8, the overall level of false alarms for the nonmanipulated disruptions was very similar to the number of reports of the manipulated temporal disruptions for the normal and 1-s videos, and false alarms did rise for 3- and 7-s videos and reached substantially higher levels (20%–30%) in Experiment 2 than in Experiment 1 (less than 10%).

Perceived Workload

Displacement significantly increased participants' perceived workload as measured by NASA-TLX, $F(3, 216) = 7.84$, $p < .001$,

Figure 7

Results From All Six Measures of Cognitive and Perceptual Processing



Note. (A) Average learning scores calculated from posttest minus pretest 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 perceived workload by condition. Error bars show the *SEM*.

$\eta^2 = .098$ (Figure 7F). Post hoc tests using Holm–Bonferroni correction reveal watching a video that was temporally displaced 7 s created significantly greater perceived workload than all other conditions: 0-s displacement, $t(72) = -4.51$, $p < .001$, $d = -0.35$; 1-s displacement, $t(72) = -3.62$, $p = .002$, $d = -0.28$; and 3-s displacement, $t(72) = -3.39$, $p = .003$, $d = -0.26$.

Table 2
Bayesian Paired Sample t Tests

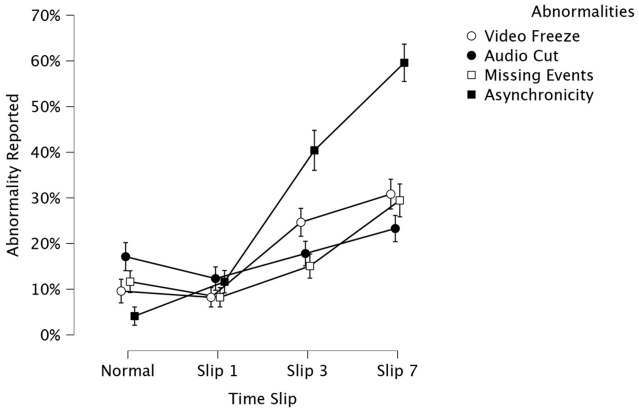
Outcome measure	Condition 1	Condition 2	BF_{01}	Error (%)
Learning score	Normal	Slip 1	7.13	0.099
	Normal	Slip 3	7.02	0.098
	Normal	Slip 7	0.11	2.39×10^{-7}
	Slip 1	Slip 3	7.72	0.11
	Slip 1	Slip 7	0.11	2.46×10^{-7}
	Slip 3	Slip 7	0.59	0.015
Segmentation uncertainty	Normal	Slip 1	7.77	0.12
	Normal	Slip 3	2.93	0.052
	Normal	Slip 7	5.72	0.085
	Slip 1	Slip 3	2.58	0.047
	Slip 1	Slip 7	5.15	0.079
	Slip 3	Slip 7	0.63	0.016
Perceived load	Normal	Slip 1	5.26	0.080
	Normal	Slip 3	4.45	0.071
	Normal	Slip 7	0.004	3.61×10^{-9}
	Slip 1	Slip 3	7.56	0.10
	Slip 1	Slip 7	0.032	5.70×10^{-8}
	Slip 3	Slip 7	0.018	2.81×10^{-8}
Segment count	Normal	Slip 1	6.32	0.091
	Normal	Slip 3	3.42	0.058
	Normal	Slip 7	6.60×10^{-4}	2.87×10^{-10}
	Slip 1	Slip 3	5.86	0.086
	Slip 1	Slip 7	0.001	8.80×10^{-10}
	Slip 3	Slip 7	0.017	2.65×10^{-8}
Disruption awareness	Normal	Slip 1	0.33	0.009
	Normal	Slip 3	3.52×10^{-8}	8.33×10^{-10}
	Normal	Slip 7	1.11×10^{-15}	4.77×10^{-20}
	Slip 1	Slip 3	7.84×10^{-5}	3.01×10^{-11}
	Slip 1	Slip 7	5.49×10^{-12}	4.19×10^{-17}
	Slip 3	Slip 7	0.003	3.30×10^{-9}

Note. BF_{01} values greater than 3.0 indicate moderate evidence for the null hypothesis where the observed data are 3 times more likely to fall under the null. BF_{01} values less than 0.33 indicate moderate evidence for the alternative hypothesis (van Doorn et al., 2021). BF = Bayes factor.

Experiment 2: Discussion

The consequences of disrupting the relationship between auditory and visual streams were 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-s disruptions decreased learning. Event segmentation and segmentation agreement remained unaffected. Similar to Experiment 1, 7-s displacement was the only condition that significantly affected all of these cognitive and perceptual processes. The 3-s displacement impacted only awareness of disruptions. Abnormality reports were higher overall for Experiment 2 compared to Experiment 1, likely because the small speed up edits and irrelevant audio cut edits were more noticeable than intended in the live-action stimuli and led participants to increase their abnormality reporting.

Figure 8
Prevalence of Reports for Each Abnormality



Note. "Video freeze," "audio cut," and "missing events" are all abnormalities that never took place.

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 1-s temporal disruptions and only an effect of awareness for 3-s disruptions. We would like to point out that these noneffects occurred in a setting that allowed participants ample opportunity to encode the videos. Not only did participants know that they would be questioned extensively about the videos, but they also had

the opportunity to view each video twice. This produced healthy learning gains of 40%–45% in correct responding from the pre- to posttests.

In contrast to the 1- and 3-s disruptions, the 7-s disruptions produced clear effects, especially for disruption awareness, perceived workload, segmentation uncertainty, and, in Experiment 2, 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 event-integration window that is longer than the multisensory temporal binding window, potentially extending to 3 s (Wallace & Stevenson, 2014).

While the 3-s temporal displacement reliably produced disruption awareness, we propose that the event-integration window extends to 3 s 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." Furthermore, our conclusion of a 3-s 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 intraevent 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 et al. (2014) demonstrates that participants can detect small one- to two-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 an 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, e.g., impossible to use a screwdriver before grabbing it). This is consistent with the previous research by Hymel et al. (2016), demonstrating that reversals in action sequences such as these are difficult to detect. The present findings converge with this previous finding by demonstrating that the link between actions and verbal narrations is similarly flexible. However, this flexibility stems from a different source. Rather than an external action constraint obviating the need for encoding, we suspect that the action–language linkages are not encoded temporally because they usually reinforce each other in a highly undetermined sequence over short intervals. This can be seen in the word choice often used by instructors. In most cases, they refer to actions they are performing in the absence of directional temporal markers such as "next" and instead refer to doing things "now." This may occur because using short-term predictive sequence markers induces cognitive load associated with generating and verbalizing precise short-term predictions. Although there is little evidence for this predictive load in language production, research on narrative comprehension does suggest that generating narrative predictions can be difficult for readers (Magliano et al., 1999). Thus, it may be efficient for speakers to meaningfully reinforce their actions by verbalizations that support flexible, iterative, action–language interactions that allow speakers both to use their statements to remind themselves of the actions they are about to produce and to use the actions they are producing as cues for necessary explanations. Thus, it may be the timing of this interactive loop, along with the rarity of necessary-to-encode short-term action-to-action relationships, that establishes a useful event-integration window.

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 complement 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.

Although we propose the possibility of a 3-s event-integration window, we recognize that this window may not have a fixed duration as it could be influenced by factors such as event structure and prior knowledge. In terms of event structure, it could be argued that the duration of this window would vary between a quick 10-s clip and a feature film. Although Experiment 1 and 2's stimuli did not include videos at the long end of this range, the instructional videos in Experiment 2 were nearly triple the duration of Experiment 1's and the content found in screen-capture versus live-action videos is substantially different. Considering the similar results across these heterogeneous materials, we believe that the event-integration window reflects a broad constraint on processing. Additionally, the event-integration window might reasonably be expected to depend in part on the availability of cognitive representations that can organize incoming information about actions. In cases where organizing knowledge is weak, it may be possible to observe the increased impact of small displacements. Just as event segmentation theory posits that prior memory representations influence event boundaries, we would expect the ability to integrate multimodal information into a single event model to depend in part on top-down knowledge guiding event construction (Zacks, 2020). Regarding gesture–speech pairs, Cavicchio and Busà (2023) found evidence that second language speakers, not native speakers, respond significantly faster when gestures and corresponding speech are synchronous versus asynchronous. In other words, native speakers were able to flexibly integrate gesture and speech pairs despite temporal disruptions, but second-language speakers were not. These results suggest prior knowledge can influence the duration of an event-integration window and there are limitations to this window that have consequential cognitive impact. A worthy future direction would be to investigate the flexibility of this event-integration window's duration by experimentally manipulating event structure and prior knowledge.

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 presemantic processes involved in predictions and the possibility of an event-integration window. However, such similarity in segmentation agreement scores across conditions could have occurred because agreement scores were calculated by comparing individuals' segmentation patterns only with participants who watched the exact same video. Each of the eight videos in Experiments 1 and 2 was associated with seven different temporal manipulation conditions, so comparison groups for segmentation agreement varied in sizes between four and eight individuals. Although differences between random and nonrandom segmentation patterns can still be detected in such small group sizes

(Sasmita & Swallow, 2023), it remains possible that our between-condition tests were underpowered if one assumes that these would be associated with a smaller effect size that tested in the previous study.

Beyond this methodological limitation, one might argue that the event-integration window we found occurs only because of the timescale of the prediction error signal. For example, less perceptual flexibility might be seen in videos in smaller timescales (i.e., >10 s) rather than our videos, which were up to 4 min long. As event segmentation theory 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 electroencephalogram 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 such as eye movements and electroencephalogram could either validate our event-integration window or they could reveal neural signatures of temporal disruption impact within a 3-s window. For instance, Simon and Wallace (2018) investigated multisensory integration utilizing stimuli that visually and auditorily presented an individual verbalizing the "BA" syllable. Simon and Wallace temporally displaced the audio and visual channels by up to 450 ms. The authors observed increased theta power and alpha suppression as these temporal displacements increased (see also London et al., 2022; Venskus & Hughes, 2021). Eye tracking may also allow us to infer participants' predictions via gaze patterns such as look-ahead fixations (for example, Sullivan et al., 2021), which may be impacted by temporal disruptions. Look-ahead fixations occur when participants look to objects that an actor in a video is about to use. If temporal predictions are violated, this may produce an added eye movement. It would be particularly interesting if this occurred but was associated with no particular cognitive or perceptual processing alarm, leading to the lack of disruption impact seen in our behavioral measures (Venskus & Hughes, 2021). 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 rote events that require little added processing.

Constraints on Generality

Across Experiments 1 and 2, we found evidence to suggest the cognitive and perceptual impact of temporally disrupting the relationship between audio and video channels in instructional videos generalizes across videos that vary in content and duration. However, our sample relied on incentivized volunteer participants from Vanderbilt University's undergraduate subject pool, which is ethnically diverse but highly selected for academic achievement. Because our target population is a wide range of adults who might view cognitively complex informational videos, it is important to consider how this selectivity may affect the generalizability of our findings. On the one hand, this selectivity may overestimate the size and impact of a cognitive temporal window. For example, it is possible that our population may possess high working memory capacity that would support forms of integrative flexibility (see Baum & Stevenson, 2017) that might increase the duration of an event-integration window. On the other hand, it is possible that our academically motivated population would overscrutinize the details in our videos, which might lessen the impact of an event-integration window. Thus, additional research might usefully assess variations in event-integration windows across different populations both to ensure generality and to explore mechanisms that produce these processing constraints.

References

Anikin, A., Nirme, J., Alomari, S., Bonnevier, J., & Haake, M. (2015). *Compensation for a large gesture-speech asynchrony in instructional videos. Gesture and speech in interaction—4th edition (GESPIN 4)* (pp. 19–23). Lund University.

Baum, S. H., & Stevenson, R. A. (2017). Shifts in audiovisual processing in healthy aging. *Current Behavioral Neuroscience Reports*, *4*(3), 198–208. https://doi.org/10.1007/s40473-017-0124-7

Cavicchio, F., & Busà, M. G. (2023). The role of representational gestures and speech synchronicity in auditory input by L2 and L1 speakers. *Journal of Psycholinguistic Research*, *52*(5), 1721–1735. https://doi.org/10.1007/s10936-023-09947-2

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), 109–127. https://doi.org/10.1037//0096-1523.21.1.109

Claus, B., & Kelter, S. (2006). Comprehending narratives containing flashbacks: Evidence for temporally organized representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1031–1044. https://doi.org/10.1037/0278-7393.32.5.1031

Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive Research: Principles and Implications*, *3*(1), Article 53. https://doi.org/10.1186/s41235-018-0146-z

Fairhall, S. L., Albi, A., & Melcher, D. (2014). Temporal integration windows for naturalistic visual sequences. *PLOS ONE*, *9*(7), Article e102248. https://doi.org/10.1371/journal.pone.0102248

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), 1183–1202. https://doi.org/10.1037/xlm0000367

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175. https://doi.org/10.1016/j.jml.2011.03.004

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., & Prinz, W. (2007). Predicting point-light actions in real-time. *Neuroimage*, *36*(Suppl 2), T22–T32. https://doi.org/10.1016/j.neuroimage.2007.03.017

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

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. https://doi.org/10.1037/xhp0000082

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.

James, W. (1892). *Psychology: A briefer course*. Henry Holt and Company.

Kirchhof, C. (2014). *Desynchronized speech-gesture signals still get the message across*. The 7th International Conference on Multimodality (7ICOM), Hongkong.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79. https://doi.org/10.1016/j.tics.2007.11.004

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition*, *39*(1), 75–91. https://doi.org/10.3758/s13421-010-0027-2

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. https://doi.org/10.1037/aca0000462

London, R. E., Benwell, C. S., Cecere, R., Quak, M., Thut, G., & Talsma, D. (2022). EEG alpha power predicts the temporal sensitivity of multisensory perception. *European Journal of Neuroscience*, *55*(11-12), 3241–3255. https://doi.org/10.1111/ejn.15719

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, *91*(4), 615–629. https://doi.org/10.1037/0022-0663.91.4.615

Meitz, T. G., Meyerhoff, H. S., & Huff, M. (2020). Event related message processing: Perceiving and remembering changes in films with and without soundtrack. *Media Psychology*, *23*(5), 733–763. https://doi.org/10.1080/15213269.2019.1636660

Meyerhoff, H. S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition*, *44*(3), 390–402. https://doi.org/10.3758/s13421-015-0575-6

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9. 11-1. Comprehensive R Archive Network*, 82(87), 126.

Petersen, A., & Vangkilde, S. (2022). Decomposing the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(8), 812–823. https://doi.org/10.1037/xhp0001018

Pöppel, E. (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *364*(1525), 1887–1896. https://doi.org/10.1098/rstb.2009.0015

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. https://doi.org/10.1080/15326900701399913

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. https://doi.org/10.1523/JNEUROSCI.0804-12.2012

Sasmita, K., & Swallow, K. M. (2023). Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods*, *55*(1), 428–447. https://doi.org/10.3758/s13428-022-01832-5

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), 77–80. https://doi.org/10.1037/a0034595

Simon, D. M., & Wallace, M. T. (2018). Integration and temporal processing of asynchronous audiovisual speech. *Journal of Cognitive Neuroscience*, *30*(3), 319–337. https://doi.org/10.1162/jocn_a_01205

Sullivan, B., Ludwig, C. J., Damen, D., Mayol-Cuevas, W., & Gilchrist, I. D. (2021). Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. *Journal of Vision*, *21*(3), Article 13. https://doi.org/10.1167/jov.21.3.13

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., … Wagenmakers, E. J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Venskus, A., & Hughes, G. (2021). Individual differences in alpha frequency are associated with the time window of multisensory integration, but not time perception. *Neuropsychologia*, *159*, Article 107919. https://doi.org/10.1016/j.neuropsychologia.2021.107919

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. https://doi.org/10.1016/j.neuropsychologia.2014.08.005

Wiener, M., & Kanai, R. (2016). Frequency tuning for temporal perception and prediction. *Current Opinion in Behavioral Sciences*, *8*, 1–6. https://doi.org/10.1016/j.cobeha.2016.01.001

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*(1), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101

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. https://doi.org/10.1162/jocn_a_00078

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*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

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), 466–482. https://doi.org/10.1037/0882-7974.21.3.466

Zhou, H. Y., Cheung, E. F., & Chan, R. C. (2020). Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions. *Neuropsychologia*, *140*, Article 107396. https://doi.org/10.1016/j.neuropsychologia.2020.107396

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. https://doi.org/10.1016/j.concog.2011.01.011

Received March 17, 2023
Revision received February 5, 2024
Accepted February 17, 2024 ■