****Preprint. Final citation is:**
Hutson, J. P., Magliano, J. P., Smith, T. J., & Loschky, L. C. (2021). "This Ticking Noise in My Head": How Sound Design, Dialogue, Event Structure, and Viewer Working Memory Interact in the Comprehension of Touch of Evil (1958). *Projections*, *15*(1), 1-27.

"This Ticking Noise in My Head" How Sound Design, Dialogue, Event Structure, and Viewer Working Memory Interact in the Comprehension of Touch of Evil (1958)

John P. Hutson, Joseph P. Magliano, Tim J. Smith, Lester C. Loschky

Abstract:

This study tested the role of the audio soundtrack in the opening scene of Orson Welles' *Touch of Evil* (Welles and Zugsmith 1958) in supporting a predictive inference that a time-bomb will explode, as the filmmakers intended. We designed two experiments and interpreted their results using The Scene Perception and Event Comprehension Theory (SPECT). Across both experiments, viewers watched the scene, we manipulated their knowledge of the bomb, and they made a predictive inference just before the bomb would explode. Experiment 1 found that the likelihood of predicting the explosion decreased when the soundtrack was absent. Experiment 2 showed that individual differences in working memory accounted for variability in generating the prediction when the soundtrack was absent. We explore the implications for filmmaking in general.

Keywords: Film Comprehension, Predictive Inferencing, Soundtrack, Working Memory, Event Structure

> *We are now having a very innocent little chat. Let's suppose that there is a bomb underneath this table between us. Nothing happens, and then all of a sudden, "Boom!" There is an explosion. The public is surprised, but prior to this surprise, it has seen an absolutely ordinary scene, of no special consequence. Now, let us take a suspense situation. The bomb is underneath the table and the public knows it, probably because they have seen the anarchist place it there. The public is aware the bomb is going to explode at one o'clock and there is a clock in the decor. The public can see that it is a quarter to one. In these conditions, the same innocuous conversation becomes fascinating because the public is participating in the scene. The audience is longing to warn the characters on the screen: "You shouldn't be talking about such trivial matters. There is a bomb beneath you and it is about to explode!"* (Truffaut, Hitchcock, and Scott 1984, 73)

Hitchcock's famous quote provides a window into the art of storytelling in film and why film is so popular. The quote posits that the success of some films is due to the filmmaker's creating a common affective experience (Smith 1995), *suspense* in this case. Affective reactions, such as surprise, suspense, and anticipation rely on how viewers comprehend the film (Lichtenstein and Brewer 1980, Brewer and Ohtsuka 1988). In the scenario above, the experience of suspense would require viewers to maintain a representation of the bomb in working memory (i.e., the maintenance and processing of information over a short time period), as the hypothetical scene unfolds. Maintaining the activation of information in working memory is fundamental to comprehension, and it is well understood that over time activation of narrative elements wanes (Just and Carpenter 1992). To aid comprehension, filmmakers may provide retrieval cues to help viewers maintain narratively relevant information (e.g., a characters says "This news is going to be explosive.") (Myers and O'Brien 1998). As such, suspense in this scene would require that viewers maintain activation of the bomb and also make a prediction about it exploding. Importantly, conventions of filmmaking and storytelling, such as those described by Hitchcock, constrain predictive inference (Magliano, Dijkstra, and Zwaan 1996).

The opening scene of Orson Welles' *Touch of Evil* (Welles and Zugsmith 1958) illustrates the importance of providing retrieval cues to support a prediction that is essential to experiencing suspense, and, indeed, the scene is regarded as an important example of using cinematic practices to create suspense (Comito 1986, D'Angelo 2012). The scene is illustrated by nine frames shown in Figure 1. It unfolds during a technically audacious single tracking shot that lasts 3 minutes and 12 seconds, opens on a close-up of someone setting a time bomb. The time bomb is then placed into the trunk of a car, after which a couple unknowingly gets into the car. When they start the car, the radio starts blaring loud music, and they drive off as the camera follows them. They drive down busy streets crowded with pedestrians. About halfway through the scene, the camera begins to follow a walking couple. As the car goes in and out of the shot, its radio gets louder and softer based on how close it is to the center of the screen. Of critical importance for our discussion, after the bomb is put into the car, it is never seen again for the remainder of the long shot. While viewing the scene, this creates a very suspenseful experience for the viewer, as they wait for the time bomb to explode while the events of the scene unfold.

INSERT FIGURE 1 HERE

Figure 1. Frames illustrating important shots in the 3 minute 12 second clip from the film *Touch of Evil* (Welles and Zugsmith 1958).

Welles argued that the soundtrack would be critical for the scene to work as intended (Tully 1999). After seeing that the studio added a non-diegetic audio track and opening credits for their release, Welles was concerned viewers would not experience the suspense of the bomb potentially exploding. His intuitions reflect an understanding that the soundtrack provided a

critical retrieval cue for viewers regarding the presence of the bomb. Walter Murch undertook a re-edit of the entire film based on Welles' memo (Welles 1998), with notable changes to the opening scene that fulfilled Welles' vision. One specific sound mix change Murch made was including the diegetic music from the radio of the car with the bomb. Concerning this decision, Murch said, "I invented the idea of putting music on the radio of the car that's about to explode. So as this car goes in and out of frame, the music anticipates the car. It's like a marker, or perfume, that says: 'This is the car.' So you have both the visual and the sound that identifies the car, to make it easier to understand what's going on" (Tully 1999). By adding the car radio to the sound mix, Murch was providing viewers with an additional token for the bomb, which is likely to both help viewers maintain the bomb in working memory, and/or reactivate it from long term memory.

Another aspect of the soundtrack that may have been important in maintaining or reinstating the bomb in working memory if it was forgotten, is the dialog. In particular, there is a line of dialog shortly before the end of the shot in which the car's female passenger repeatedly tries to get the attention of a border patrol officer, saying "Hey! Hey, I've got…I've got this ticking noise!… No really, I've got this ticking noise in my head!" This dialog towards the end of the long shot, just before next shot of the bomb exploding, was likely considered an important point to reactivate knowledge of the bomb for viewers who had forgotten about it.

This study tested the filmmakers' intuitions of the importance of the soundtrack in supporting the prediction that the bomb will explode (Experiment 1), and then, from a cognitive perspective, that without the soundtrack, there is a heavy burden placed on working memory (Experiment 2). As such, viewers with limited working memory capacities should be less likely

to generate the intended prediction regarding the bomb, foiling Welles' intention to create suspense.

Scene Perception and Event Comprehension Theory (SPECT)

To frame our conceptualization of the interaction between viewer cognition and the film stimulus in the creation of suspense, we utilize the Scene Perception and Event Comprehension Theory (SPECT; Loschky, Larson, Smith, & Magliano, 2020). SPECT is a theoretical framework designed to make predictions about the perception and comprehension of visual narratives (Loschky et al. 2020). SPECT explains how viewers create a cognitive representation of a filmed narrative, which we refer to as the *event model*. SPECT integrates established theories from scene perception (Henderson and Hollingworth 1999), event perception (Radvansky and Zacks 2014), and narrative comprehension (Gernsbacher 1990). A key distinction in SPECT is between *front-end* processes that involve visual perception and attention and *back-end* processes in working and long-term memory that support constructing event models. Front-end visual processes include information extraction (i.e., recognizing agents, actions, and objects, etc.) and attentional selection (i.e., determining where one looks, and which information one processes). Back-end processes include event segmentation, which allows one to parse the smaller events that make up a larger event (Kurby and Zacks 2008), inference generation, and event model updating (i.e., integrating the understanding of the current event with prior events in the narrative, which are in long-term memory (Loschky et al. 2020).

Previous tests of SPECT have explored the extent to which back-end processes influence front-end attentional selection in film and picture stories. In film, back-end processes appear to have little effect on attentional selection (i.e., where viewers fixate) as a scene unfolds (Hutson et al. 2017, Loschky et al. 2015). In contrast, the need to generate an inference has an impact on

attentional selection in picture stories (Hutson, Magliano, and Loschky 2018). In the opening scene of *Touch of Evil*, differences between viewers in generating a predictive inference that the bomb in the car was going to explode did not influence attentional selection (Hutson et al. 2017). On the other hand, whether characters were treated as protagonists (versus not) did briefly influence attentional selection.

While SPECT is concerned with explaining visual scene perception and event comprehension, it could potentially be extended to the auditory modality. Specifically, SPECT's front-end focuses on the processes of visual attentional selection and information extraction. Nevertheless, it seems reasonable to posit that auditory processing of diegetic sound in a film's soundtrack would initially involve the same two key front-end processes of attentional selection and information extraction, since these have both been studied in audition (Bregman 1994). However, how attentional selection and information extraction interact in audition compared to vision is likely to differ due to fundamental differences between the two sensory systems (for an example of this dissociation in film see Smith 2014). Furthermore, auditory processing of dialog would involve specific linguistic processes (word recognition, semantic role assignment, syntactic processing) (Dick et al. 2016), and their output would be passed along to the back-end. On the other hand, regardless of the information modality, the output of all of these front-end processes could be assumed to be amodal information about the event (i.e., "event indices": people, objects, actions, causal relationships, and characters' goals; Zwann & Radvansky, 1997; Magliano et al. 2012) which is sent to the back-end. SPECT's back-end processes are agnostic to modality. Thus, the effects of visual information, diegetic sound, and spoken dialog on the event model would be equivalent in SPECT. Consistent with the above ideas, both film scholars (Bordwell 1985) and researchers (Batten and Smith 2018), have argued that diegetic audio

should have a large impact on the viewer's comprehension of film. In support of this argument, is evidence that diegetic information in the audio track supports predictive inferences (Magliano, Dijkstra, and Zwaan 1996) and more broadly comprehension (Manfredi et al. 2018). As noted, Welles and Murch argued that the soundtrack is important to supporting predictions and feelings of suspense regarding the bomb. Cognitively speaking, the visual presence of the car, the sound of the car radio, and the dialogue containing the "ticking comment" at the border check point, after being attended to and extracted in the front-end, all would serve as retrieval cues for the bomb, and thus contribute to generating an inference of an up-coming explosion in the back-end. Activating relevant narrative content is necessary for generating inferences based on that content (Graesser, Singer, and Trabasso 1994, McKoon and Ratcliff 1998, Albrecht and Myers 1998). Removing the soundtrack should reduce the retrieval cues for the bomb, and decrease the likelihood that the predictive inference of the bomb exploding will be regenerated.

Nevertheless, there is recent evidence that diegetic sound and visual information in film are not always combined in the back-end event model to guide a viewer's attention (Batten and Smith 2018). Thus, it is possible that the addition of the diegetic sound and dialog to the visual stream in the opening scene of Touch of Evil might not affect the viewers' event models.

Working memory and knowledge activation in the event model

Scenes such as the opening one in *Touch of Evil*, place a burden on the cognitive resource of working memory, which is important for comprehension and inference generation. Working memory constitutes a complex set of cognitive systems that control the extent to which people maintain the activation of relevant information over time, and perform mental computations on it (Baddeley 1992, Just and Carpenter 1992). Working memory resources are critical for understanding narrative texts, particularly because understanding narratives often requires

making connections between information over long distances in the text, or in this case over long stretches of time in a film (Just and Carpenter 1992, King and Just 1991). Moreover, it is well established that working memory resources are measurable and vary across people (Just and Carpenter 1992). Numerous studies (Whitney, Ritchie, and Clark 1991, Calvo 2001, Rai, Loschky, and Harris 2014, Rai et al. 2011, St George, Mannes, and Hoffman 1997) have shown that readers with a high working memory capacity are more likely to generate inferences that require the activation of prior information in a story than readers with low working memory capacity.

Working memory demands in processing texts operate at the word, sentence, and discourse levels (Just and Carpenter 1992). Most germane to the present study is at the discourse level, which involves establishing semantic relationships between discourse constituents (i.e., clause and sentences). The more distantly related discourse constituents are (i.e., number of sentences between them), the greater the demands placed on working memory to establish those relationships (Just and Carpenter 1992). Narrative films similarly require the viewer to make connections between story content that is displaced in time, and as such there should be greater demands on working memory the more screen time between story elements that are necessary for an inference. The opening scene of *Touch of Evil* is a case in point. As the time between the introduction of the bomb increases, the demands on working memory to maintain that content increase, and in particular as new situational content is introduced (e.g., the couple is introduced). Thus, we predicted that working memory capacity should be correlated with the likelihood of generating the prediction that the bomb will explode near the end of the scene. The presence of the soundtrack likely reduces the burdens on working memory by providing retrieval cues to supports the maintenance of activation of knowledge of the bomb in working memory.

To our knowledge few studies have explored the extent to which working memory is related to inference generation in narrative film.

The Current Study

SPECT offers a theoretical framework within which to recast the intuitions of filmmakers about how this scene works, and to empirically test them. Experiment 1 was conducted to test the intuition of Welles and Murch that both the "ticking" comment in the dialog, and the car radio sound editing, would remind viewers who had forgotten about the car and the bomb in it, of their existence, thus engendering a predictive inference about the car exploding. In Experiment 2, we assessed the extent to which individual differences in working memory capacity accounted for variability in generating the prediction by removing the soundtrack and relating the rate of predicting the explosion to individual viewers' working memory capacity.

In both experiments, viewers watched the opening scene, which ended *just before the bomb exploded* in the second shot, and were asked to make a prediction of what would happen next. Both experiments manipulated the viewing context to provide a control condition in which the prediction should not be made. Specifically, in the No-context condition, viewers were not shown the segment of the scene in which the bomb was placed in the car (see Figure 1). This is equivalent to the experience of an audience member who arrived 18 seconds late for the movie, thus we call this the "jumped-in-the-middle" paradigm. Experiment 1 employed a 2 (Bomb knowledge: Context, No-context) X 2 (Soundtrack: present, absent) between-participants design, while Experiment 2 only manipulated knowledge of the bomb (Context, No-context).

Experiment 1: Context and Audio Influences on Predictive Inference Generation

Method

Participants. There were 94[1] participants included in the data analyses (54 females; mean age = 18.6 years; *SD* = 1.3). The number of participants for this and the following experiment were based on previous work that used the same context manipulation and similar materials (Loschky et al. 2015). The experiment was run online using Qualtrics, and participants used their personal computers. Participants were pseudo-randomly assigned to one of four viewing conditions of the opening scene of *Touch of Evil*. The participants included in the data analysis had a fairly equal representation in each condition (Context + Audio, $n = 21$; Context + No-Audio, $n = 29$, No-context + Audio, $n = 24$, No-context + No-Audio, $n = 20$)[2]. Participants were Kansas State University undergraduate students participating for course research credit. The University Institutional Review Board determined that this and the following experiment posed minimal risk to the participants, and thus determined informed consent was unnecessary.

Stimuli. Two versions of the opening scene of Orson Welles' *Touch of Evil* were used (Welles and Zugsmith 1958). The Context version showed the bomb placed in the car trunk (Figure 1) at the beginning, and ran for 3:12. The No-context version omitted the first 18 seconds when the bomb was placed in the car, and ran for 2:54. Both clips ended on the close-up of the walking couple kissing. The Context and No-context conditions were presented in both Audio and No-audio conditions, with the soundtrack taken from the Walter Murch audio mix (Welles 1998).

Data accessibility. The data and R scripts can be accessed online at the Open science framework [3].

Procedure. In all four conditions, the experimental procedure was the same. Viewers were randomly assigned to one of four Context x Audio conditions. After viewing the video clip, viewers were presented with a series of written questions. To check whether viewers in the Context condition maintained the bomb in their event model, the first question was, "What do you think will happen next?" and they were prompted to type their response in a text box. The following questions were to ensure that viewers had not seen the clip before. No viewers reported having seen it before.

Results and Discussion

To identify whether viewers' predictive inferences at the end of the clip were influenced by having seen the bomb put in the car trunk, and therefore having the bomb in their event model, two research assistants, who were blind to the conditions, coded each predictive inference. Responses having no mention of the bomb or an explosion were coded as '1'; unclear responses were coded as '2'; responses making clear reference to the bomb or an explosion were coded as '3'. The coders had a relatively high level of inter-rater reliability (Cohen's *Kappa* = .89). Discrepancies between the two coders were resolved through discussion. Unclear responses (coded as '2') were resolved as relating to the bomb or not, resulting in a final dichotomous coding indicating whether participant predictions did mention the bomb '1', or did not '0'.

The results of these analyses are shown in Figure 2. We ran Chi-square tests to test for differences in the frequencies of making a predictive inference about the bomb between the four conditions. There were large differences based on viewer Context (X^2 (1, N = 94) = 28.517, p < .001; *Eta* = .551). Unsurprisingly, viewers who saw the bomb at the beginning were far more likely (56%) to make an inference about it than those who had not (4.5%) (Figure 2).

Importantly, however, within the Context condition, viewers in the Audio condition were much more likely to draw bomb-related inferences (76%) than those in the No-audio condition (41%)(X^2 (1, N = 50) = 5.99, p = .014; Eta = .346). There was no significant difference between audio conditions when collapsed across Context conditions (X2 (1, N = 94) = 2.597, p = .107; Eta = .166), indicating that drawing the inference was primarily dependent on having prior knowledge of the bomb. For this reason, it was very surprising to find that two viewers in the *No-context* +Audio condition *did* make a predictive inference about the bomb exploding, without having seen the bomb put in the car. Importantly, these viewers did *not* indicate having seen the film before. A plausible explanation is that they could make the inference because the female character in the car mentioned a "ticking noise." This is also one explanation for why, in the Context condition, the Audio group outperformed the No-audio group. Specifically, perhaps some viewers in the Context + Audio condition forgot about the bomb, but were reminded of it near the end of the film clip when they heard the "ticking" comment. Conversely, if there were an equal proportion of viewers in the Context + No-audio condition who forgot about the bomb, they would not have been reminded of the bomb, because they could not hear the "ticking" comment. If so, then the difference in performance between the Context + Audio versus the Context + No-audio groups (76% - 41% = 35%) would represent the proportion of Context condition viewers who forgot about the bomb. We tested this hypothesis in Experiment 2.

INSTERT FIGURE 2 HERE

Figure 2. Probability of making an inference about the bomb. Context condition is on the left, and No-context is on the right. Audio condition is in Blue, and No-audio is in Red.

Note that the strong effect of Context (i.e., knowledge of the bomb) on making bomb-

relevant inferences is unsurprising. However, it is important for the study, because it establishes that the context manipulation produced clear differences in the information viewers had in their event models as they viewed the film clip. Additionally, the much higher probability of viewers making the explosion inference when presented with Context+Audio supports the prediction by the filmmakers and film theorists that the sound mix increases the suspense felt by viewers by helping them maintain the bomb (in the car) in their event model in working memory.

Experiment 2: **Event Segmentation and Working Memory**

We hypothesized that for viewers of the opening shot of *Touch of Evil*, when the soundtrack was absent, there was a heavier burden placed on working memory. Experiment 2 tested this possibility by assessing the extent to which, when the soundtrack was absent, individual viewers' differences in working memory capacity could account for variability in their likelihood of generating the bomb prediction. However, working memory is not the only contributing factor that supports inference generation. Theories of inference generation assume inferences are constrained by contents of an event model for a narrative (Graesser, Singer, and Trabasso 1994, Van den Broek 1990). Presumably, differences in event models should lead to differences in the inferences that are generated as one processes a narrative (Anderson and Pichert 1978). Indeed, that is the basis for the explanation for why viewers were less likely to generate the prediction in the No-context condition relative to the Context condition (Hutson et al. 2017). Specifically, differences in viewers' knowledge of the bomb change the nature of their event model (Loschky et al. 2015).

One approach to assessing viewers' event models while watching narrative films is the *event segmentation* task. In this task, while viewers watch a narrative, they are asked to identify

any time that they think events have changed (Magliano, Miller, and Zwaan 2001, Zacks, Speer, and Reynolds 2009). The likelihood of perceiving an event change is correlated with shifts in situational continuities in characters, their actions, space, time, causality, and characters' goals (Zacks et al. 2009). If viewers differ in their understanding of the scene as a function of the context manipulation, we would expect to find differences in how they *segment the film* across the Context and No-context conditions. We tested this hypothesis in Experiment 2. We hypothesized that viewers in the Context condition would be more regular in their segmentation when the car containing the bomb appeared on the screen.

However, most importantly, we tested the hypothesis that viewers' working memory capacity would predict their likelihood of generating the predictive inference of the explosion at the end of the clip. Specifically, it is possible that some viewers may forget about the bomb between the time they see it put in the car at the beginning of the shot, and the end of the shot roughly 3 minutes later. During that 3 minutes, there are a plethora of people, places, things, and events to look at, which could fill-up viewers' working memory capacity. This could lead to forgetting about the bomb being in the car, especially for viewers with lower working memory capacity, reducing the likelihood of generating a predictive inference involving the bomb, which we call the *working memory and predictive inference hypothesis*.

We also assessed whether individual differences in agreement between viewers in the segmentation task in the Context condition accounted for differences in viewers' likelihood of generating the explosion prediction over-and-above their working memory capacity. It has been shown that variability in segmentation agreement is predictive of memory for filmed events (Sargent et al. 2013), which formed the basis for a further hypothesis that viewers' level of

segmentation agreement would be positively correlated with their likelihood of generating the explosion prediction in the Context condition.

In Experiment 2, viewers viewed the scene without audio. This removed the reminders of the bomb in the dialog and the sound editing, placing a greater importance on viewers' working memory to maintain the bomb in their event model for the scene. While watching the scene, viewers engaged in the event segmentation task, and then answered the prediction question at the end of the clip. We measured viewers' working memory capacity and event segmentation, and used logistic regression to assess the extent to which those variables predicted viewers' likelihood of generating a bomb-related predictive inference.

Method

Participants. A total of 81 students enrolled in an introductory psychology course at Kansas State University participated for course research credit, and were pseudo-randomly assigned to either the Context (n = 41) or the No-context (n = 40) condition. The participant number was based on the results of Experiment 1, as well as other eye tracking experiments run with the same manipulations and materials (Hutson et al. 2017). All participants had 20/30 or better corrected or uncorrected vision (FrACT; Bach 2006). Only participants who had previously participated in a working memory capacity study were given the possibility of signing up to participate in this experiment. Despite this, working memory scores were only available for 31 of 41 participants in the Context condition, due to participants either not completing the working memory study or the inability to match participants between the studies[4].

Stimuli. The same opening scene from *Touch of Evil* was used in Experiment 2, but it was presented without the soundtrack. There was one additional change to the materials. In order to create a condition in which the bomb, the car, and it's passengers are rendered

completely irrelevant to the viewer's event model for the scene, our No-context clip started 1 minute and 49 seconds into the opening scene, when the walking couple was shown alone on the screen, with the car off-screen (Figure 1).

Working memory measures. Working memory capacity was measured in a separate experiment in which viewers completed the Operation Span (OPSAN, Turner and Engle 1989), reading span (RSPAN, Daneman and Carpenter 1980), and counting span (CSPAN, Case, Kurland, and Goldberg 1982) tasks. Our working memory predictions were not specific to a given measure, so for the analyses below, we used a composite measure of working memory that combined viewers' scores on all three measures. Specifically, for each participant, we calculated z-scores for each measure, and took their average across all measures, so each participant had a composite span score.

Procedure. Participants viewed the clip in the laboratory in groups of up to four. Chin rests were used to maintain a constant viewing distance. The monitors were set to refresh at 60Hz, to ensure no dropped frames for the video playing at 30 frames per second. To learn the event segmentation process, we gave viewers practice with the task using an example video of a person folding laundry. They were instructed to press a button whenever they perceived "new" events that were "natural" and "meaningful." After the practice video, the instructions were repeated, and viewers began the experimental video clip. As in Experiment 1, at the end of the video we asked viewers "What will happen next?" We removed the data from two participants because they reported having seen the film before.

Results and discussion

We report 3 main results sections below, and an exploratory analysis. First, we report the proportion of viewers who generated the predictive inference about the bomb. Second, we report

the event segmentation analyses, which tested both the frequency of segmentation and the level of agreement between viewers as to when event boundaries occurred. Here we report an exploratory analysis with an accompanying hypothesis to further explore unexpected results. Finally, we report whether segmentation agreement and/or working memory were related to the likelihood of generating the inference about the bomb.

Predictive inference. We followed the same procedure for coding the explosion inference as Experiment 1. Two coders independently categorized the predictive inferences viewers made as either relating to the bomb, or not. Their inter-rater reliability was high (*Kappa* = .96). The coders resolved coding discrepancies through discussion.

A chi-square analysis showed that, as before, viewers in the Context condition were far more likely to make the predictive inference about the bomb (39% of viewers) than viewers in the No-context condition (0% of viewers, X^2 (1, N = 82) = 18.932, $p < .001$, *Eta* = .480). Thus, the *jumped-in-the-middle* context manipulation again strongly affected participant's event model of the film clip. Importantly, adding the event segmentation task did not affect the likelihood of viewers generating the explosion inference in the Context condition compared to Experiment 1 (Exp 2: 39% vs. Exp 1: 41% respectively).

Event segmentation. To test for a general difference in event segmentation throughout the portion of the clip that both conditions saw, we calculated the proportion of events viewers identified in the Context and No-context conditions. For this analysis, we first divided the clip into one-second bins. Next, for each participant, we calculated whether they pressed their button to identify an event in each of the bins. This produced an overall distribution of when events were identified by each participant throughout the film clip. Then we aggregated across viewers to produce an overall group distribution for events in the clip. The No-context viewers

perceived a significantly higher proportion of events for any given second of the scene (M = .079, SD = .061) than the Context viewers (M = .055, SD = .036)(X^2 (1, N = 44) = 169.00, p < .001; Eta = 1.0) (Figure 3). These data are convergent with the prediction data and support the interpretation that the context manipulation created differences in viewers' event models for the film clip. The finding that the Context condition identified fewer events replicates the same pattern found using the jumped-in-the-middle paradigm with a different film clip (Loschky et al. 2015). It also replicates a difference between experts and novices when segmenting dance performances, in which novices segmented more frequently than experts (Bläsing 2015). One explanation for this is that a better understanding should allow better prediction of upcoming events, and event segmentation typically occurs when there is a prediction error in an event (Reynolds, Zacks, and Braver 2007) or a break in coherence (Loschky et al. 2015). Alternatively, the No-context group may have identified more events because they were tracking more events occurring outside of the primary narrative about the bomb (e.g., the introduction of new characters and events that are irrelevant to the bomb narrative), which is a general indicator of poor comprehension (Gernsbacher, Varner, and Faust 1990).

INSERT FIGURE 3 HERE

Figure 3. Proportion of new events indicated within 3-second bins by Context condition. Three-second bins were chosen for this figure, as opposed to the 1-second bins used in that analyses. The use of 3-second bins gives a clearer depiction of the segmentation frequency and agreement effects than the figure with 1-second bins.

The event segmentation data was also analyzed in term of segmentation agreement which provide an assessment of the extent that a group of participants identify event boundaries at similar locations (i.e., similarly across the 1 second time bins). The greater the segmentation agreement within a condition, the more similar the event models for participants within that condition. Moreover, differences in segmentation agreement across conditions would reflect that there are differences in the event models constructed by participants across conditions. For this analysis, we used Zacks' (Zacks 1999) segmentation agreement scoring method. A group segmentation baseline was calculated (i.e., the proportion of participants that identified event boundaries for each time bin) and correlations were computed between the baseline and each participants segmentation judgments. This created an agreement score for each participant. A value of zero indicates no agreement between an individual and the group, and a value of 1 indicates perfect agreement. The correlation is scaled to account for variability in the number of events viewers identified, so the score is not biased towards those who indicate either very few or very many events (i.e., No-context viewers' agreement scores would not be biased by their higher segmentation frequency).

Mean segmentation agreement scores were computed for each condition. Surprisingly, a Welch's *t*-test showed that viewers in the No-context condition had higher agreement scores (M = .51, SD = .16) than the Context viewers (M = .42, SD = .17; $t(74.38) = -2.36$, p = .021, *Cohen's d* = .54). This shows that No-context viewers' segmentation behavior was more systematic. It is surprising that the No-context group showed higher agreement because 1) we had initially predicted they would have lower agreement due to having worse comprehension, and 2) they also had a higher overall segmentation frequency, which we have already attributed to having worse comprehension.

Exploratory analysis of prediction protocols. Our initial hypothesis was that the presence of the bomb in the Context condition would lead those viewers to have more similar event models. However, upon consideration of the segmentation agreement results, an alternative hypothesis arose. Specifically, only 41% of the viewers in the Context condition generated a prediction about the bomb. This suggests that there was actually considerable variability in how viewers within the Context condition understood the scene, which could have reduced segmentation agreement compared to the No-context condition. Viewers who did not include the bomb in their event model may have more universally understood the scene as simply showing a couple having a leisurely walk.

To test this, we recoded viewers' prediction protocols into one of four mutually exclusive categories. Viewers could either make a predictive inference about 1) the bomb, 2) continued social interaction between the walking couple (e.g., the couple will go to dinner), 3) continued social interaction between couple in the car (e.g., they will drive off to another town), or 4) an interaction between the two couples (e.g., the two couples will meet). There were a few protocols that could not be classified as falling into any of these categories, and so we added a category of "Other." Two raters coded each prediction protocol independently, and reliability was high (*Fliess' Kappa* = .955, $z = 12.8$, $p < .001$). The few discrepancies were discussed and adjudicated.

The results of this analysis were consistent with the hypothesis that there was greater variability in interpreting the scene in the Context than the No-context condition. That is, viewers in the No-context condition were much more consistent in their predictive inferences, with 87.5% making an inference about the walking couple (Table 1). Context condition viewers showed more variability in the categories of their predictive inferences, with the bomb and

walking couple inferences almost tied for frequency (38% and 40% respectively). The differences in frequencies across conditions and the types of predictions were significant ($X^2(3) = 29.44$, $p < .001$, *Eta* = .599). Thus, this analysis explains the higher segmentation agreement in the No-context condition than in the Context condition.

INSERT TABLE 1 HERE

Table 1. Percentage of inference type by condition

Individual differences and prediction. This final planned analysis tested the hypotheses that 1) viewers who had higher working memory capacity would be more likely to maintain knowledge of the bomb in working memory, thus they would be more likely to make a predictive inference about it, and 2) that viewers who made an inference about the bomb would have greater segmentation agreement. We had working memory scores for 31 Context condition participants, 12 of which made the inference about the bomb. Given this relatively small number of participants with working memory scores, these result are tentative.

We entered segmentation agreement (a correlation value between 0 and 1) and working memory (standardized z-score of working memory) scores for the Context condition viewers into a logistic regression to predict their likelihood of generating the inference about the bomb. Only the main effects were entered in the regression, because 1) we did not have specific predictions about the interaction of segmentation agreement and working memory, and 2), with 31 viewers, the model including the interaction would be underpowered.

Consistent with our *working memory and predictive inference hypothesis*, viewers with higher working memory scores were more likely to make a bomb related inference. Those who made the predictive inference about the bomb got a higher proportion of working memory span items correct ($M = .86$, $SD = .07$) than viewers who did not make the inference ($M = .68$, $SD =$

.22). This effect was significant (b = 1.78, $z(29)$ = .90, p = .047, *Odds Ratio* = 5.93). Specifically, the model predicted that participants having a perfect working memory span score would make a bomb-related inference about 85% of the time, while participants answering only half the working memory span questions correctly would make a bomb-related inference only about 10% of the time. However, contrary to our event segmentation and predictive inference hypothesis, a participant's segmentation agreement did not significantly predict participant likelihood of generating a bomb-related inference (Bomb Inference Agreement M = .51, SD = .15; No Bomb Inference Agreement M = .54, SD = .15; b = -1.23, $z(29)$ = -.45, p = .653). The overall model predicted approximately 20% of the variance in whether viewers in the Context condition made a predictive inference about the bomb (R^2 = .20). This suggests that viewers' working memory capacity plays an important role in whether they will maintain the bomb in their working memory throughout the *Touch of Evil* opening shot, and thus experience Welles' intended continual suspense throughout the shot[5].

General Discussion

People around the world enjoy films, because filmmakers can use the medium to create intended enjoyable cognitive and affective states. The opening scene of *Touch of Evil* is a famous example of this, and is unique, among many other reasons, because of the written documents in which the filmmakers described how the scene should work (Tully 1999). In this paper, we used the SPECT narrative cognition framework (Loschky et al. 2020) and what is known about narrative comprehension to test explanations about how the scene "works". We hypothesized that the soundtrack, including the dialog and diegetic sounds, serve as retrieval cues, so knowledge of the bomb, and the car it is in, remain available in working memory, which should support generating inferences about the narrative (e.g., Myers and O'Brien 1998).

Specifically, the soundtrack likely reduces the burden of working memory demands to maintain the activation of this information. Therefore, when the soundtrack is absent, there should be a reduction in the likelihood of viewers generating the prediction that the bomb will explode. Furthermore, when the soundtrack is absent, individual differences in working memory capacity should account for variability in the likelihood of viewers generating the explosion prediction.

Our results were consistent with the hypotheses about the role of the soundtrack and viewers' working memory capacity in supporting their ability to generate Welles' intended prediction. Specifically, less than half of the viewers generated the explosion prediction when the soundtrack was absent (about 40%) (Experiments 1 and 2), whereas the majority generated the prediction (76%) when it was present (Experiment 1). Moreover, in Experiment 2, we found that individual differences in viewers' working memory capacity accounted for significant variability in their generating Welles' intended prediction. Our exploratory analysis of the prediction protocol for Experiment 2 indicated that there was, in fact, greater variability in the types of predictive inferences that viewers made in the Context condition relative to the No-context condition. These data are consistent with the idea that some of the viewers were able to maintain the activation of the bomb in their event model in working memory during the scene, while others were not. However, we did not find evidence consistent with the hypothesis that segmentation agreement was related to the likelihood of generating the prediction. This suggests that individual differences in working memory accounted for the variability in generating the prediction about the bomb in the No-audio condition, as opposed to differences in understanding the scene's event structure.

However, this last conclusion is tentative. While segmentation agreement has been shown to be correlated with comprehension performance in other contexts and working memory

capacity (Sargent et al. 2013), it may be the case that variability in segmentation agreement is not directly related to predicting a specific narrative event (i.e., the bomb exploding). The exploratory analyses of the prediction protocols suggest the possibility that differences in event models across viewers in the No-audio condition may also have contributed to variability in generating the prediction. Specifically, contrary to what was predicted, there was actually greater segmentation agreement in the No-context condition than the Context condition in Experiment 2. The exploratory analysis of the prediction protocol showed that there was greater variability in the Context condition than the No-context condition. Thus, the segmentation agreement and prediction protocols provided converging evidence of greater variability in the event models in the Context condition than the No-context condition when the soundtrack was absent.

One caveat to this interpretation of the converging segmentation agreement and prediction data is that segmentation frequency was higher in the No-context condition than the context condition. Narrative events that lead viewers to make more prediction errors have been associated with relatively higher segmentation frequency (Reynolds, Zacks, and Braver 2007). Concerning the high rate of prediction error, an explanation may be related to what features viewers in each condition were basing their segmentation responses on. The predictions that drive segmentation are typically at a perceptual level (e.g., predictions that perceptual motion continuity will persist), but can also be modulated by higher-level goal-directed activity (Zacks 2004). Based on this, viewers in the No-context condition may have been following events around the walking couple, and relying on low-level perceptual features of the scene (e.g., walking motion) to guide their segmentation behavior. Conversely, viewers in the Context condition could have followed events around different characters and the bomb, and could have

thus segmented more on predicted causal-relations in the narrative (e.g., the bomb exploding, and its effects on the walking couple, other pedestrians, buildings, etc.).

It is also important to acknowledge that while others have used event segmentation to measure differences in viewers' event models across a context manipulation in narrative film (Loschky et al. 2015), it is always best practice to use converging methodologies to study inference processing, particularly for naturalistic materials (Magliano and Graesser 1991). In fact, the present study employed converging methodologies by using the question answering task, which is also sensitive to the nature and quality of event models (Graesser and Franklin 1990). However, a key challenge when conducting research on narrative film is developing unobtrusive measures of event model processing that are sensitive to the moment-to-moment processing of film (i.e., changes in the event model as a film is viewed). Because reading text is self-paced, researchers can use sentence reading times, which are sensitive to inference processing (e.g., Clark 1977). However, this method is more difficult to apply to film perception and comprehension studies. However, event segmentation is an unobtrusive activity that does not appear to change the processing of events (Zacks, Speer, and Reynolds 2009). Moreover, event segmentation in narrative film is sensitive to factors associated with event model updating. For example, the likelihood of perceiving event boundaries increases as situational changes increase (e.g., jumps in narrative time, shifts to new locations, new characters introduced, causally anomalous event depicted, etc.) (Huff, Meitz, and Papenmeier 2014, Magliano, Miller, and Zwaan 2001, Zacks, Speer, and Reynolds 2009). Nonetheless, the success of future research on event model processing in narrative film rests on developing and using multiple paradigms, as each has its own limitations. Using converging methodologies is one approach to overcome these limitations.

The present study illustrates the virtue of empirically investigating insights from film makers. Both Orson Welles and Walter Murch (Tully 1999) have documented statements on how this opening scene of *Touch of Evil* (Welles and Zugsmith 1958) should work. Specifically, they focus on how the use of diegetic sound should increase the suspense felt by viewers. Murch specifically stated that the radio of the car with the bomb would work as a retrieval cue for the car, and thus also the bomb contained in it (Tully 1999). In other words, without the diegetic sound, some viewers may forget about the car, and also the bomb, which would remove the emotional impact of the scene. Similarly, we hypothesize that Welles included the dialog with the "ticking" comment specifically for the purpose of reminding viewers who forgot about the bomb of its existence, shortly before it exploded. Based on our results, Welles and theorists' perspectives on how the scene works are correct.

While one could argue that the conclusions we can draw here are limited to this scene, in the spirit of case study designs, we argue that it is possible to generalize beyond it. Specifically, the use of shot scale, dialog, and sound editing to constrain inference processes are not unique to this scene (Magliano, Dijkstra, and Zwaan 1996). We see this study as illustrating the promise of collaborations between filmmakers, film theorists, and cognitive scientists (Clinton et al. 2017).

Implications for SPECT

With respect to the contributions from cognitive science, SPECT (Loschky et al. 2020) was used in this study as a framework to make predictions about how film features and the manipulation of context would influence viewers' comprehension, and it also informed the tasks used in the study. SPECT assumes that when viewers watch or read visual narratives, their event models are generated through the interaction of front-end processes occurring during single eye

fixations and back-end processes in working memory and long-term memory. The perceptual processing of the visual stream and the soundtrack constitute front-end processes that constrain the back-end processes of inference generation and event segmentation. As such, SPECT provided a cognitive framework that shows Welles' and Murch's insights had psychological foundations we were able to use to explain the cognitive processes that make the scene work.

SPECT also provided a framework for determining the tasks used in the study (i.e., the prediction and event segmentation tasks). SPECT assumes that predictive inference generation and event segmentation are closely related to the "shifting" processes in the event model. Specifically, based on the viewer's current event model, the viewer generates predictions, and if these predictions are not met, the viewer perceives an event boundary, as measured by the event segmentation task, and *shift* to creating a new event model. The convergence between the segmentation agreement analysis and the exploratory analysis of the prediction protocols is consistent with this idea (see also Radvansky 2012). However, in this study, segmentation agreement was not predictive of the likelihood of generating the prediction. More research is warranted on assessing this relationship in the context of film comprehension. Moreover, this study also illustrated a possible future direction of the SPECT framework acknowledged by Loschky et al. (2020). Specifically, SPECT is currently constrained to describing the front-end processes that support film processing in the visual system, and is agnostic about those involved in linguistic and non-linguistic auditory perception. Film is obviously a multimodal experience, and a comprehensive framework should eventually account for a range of front- and back-end processes associated with processing scenes; including linguistic information, diegetic sounds, and non-diegetic sounds (Magliano et al. 2013, Magliano, Higgs, and Clinton 2019). Specifically, there is relatively recent work exploring the roles of audio in guiding attention in

film (Batten and Smith 2018), decreasing awareness of cuts (Smith and Martin-Portugues Santacreu 2017), affecting event processing in film (Meitz, Meyerhoff, and Huff 2019), and computational work attempting to model a film's multimodal influence over viewing behavior (Coutrot and Guyader 2016). This recent work, along with the long history of theorizing about the multimodal aspects of film (e.g., Chion 1994) point to the need for future versions of SPECT to accommodate multimodal factors.

An additional future direction for the development of SPECT is the integration of how individual cognitive differences between viewers, such as their working memory capacity, influence their moment-to-moment processing of scenes and events, in both the front-end and the back-end. Most theories of text comprehension that describe back-end processing are agnostic about the impact of individual differences in general (McNamara and Magliano 2009). While many assume working memory constrains back-end processes (e.g., Kintsch 1988), few contain specific assumptions and parameters regarding working memory (see as an exception, Just and Carpenter 1992). SPECT similarly acknowledges that working memory constrains backend processes, but makes no assumptions regarding how this is the case. The SPECT framework is amenable to the integration of individual differences in mechanisms such as working memory. For example, SPECT does assume viewers lack the capacity in working memory, attentional resources, or executive resources to encode many surface details in films, which results in phenomena of viewers not noticing when edits occur in films (Smith and Henderson 2008, Smith and Martin-Portugues Santacreu 2017), or even a change in actor across cuts (Levin and Simons 1997). Further work needs to be done to test the role of individual differences in the development of event models during scene and event perception.

Limitations and future directions

There are a number of aspects of this study that warrant further consideration. First, the working memory and inference generation effect could have been influenced by the task of event segmentation. Specifically, it is possible that the segmentation task may have required participants to engage in task switching between constructing their event model by watching and comprehending the scene and remembering to press the segmentation button. Thus, participants higher in working memory may have been less impacted by task switching (Liefooghe et al. 2008). Event segmentation has been shown to be a naturally occurring process (Zacks et al. 2001), but as an explicit task, it can have an impact on participants' perceptual processes and goals (Zacks and Swallow 2007). Similarly, correlation research has shown that working memory and segmentation are related (Sargent et al. 2013), but it is not clear if the relationship is directly related to task performance or the development of the event model. As such, future work should address this in working memory experiments that do not include a secondary task. Nevertheless, it is worth reiterating that working memory capacity has repeatedly been shown to both be correlated with inference generation during reading, and to experimentally influence it (Calvo, 2001; Rai, Loschky, & Harris, 2014; Rai, Loschky, Harris, Peck, & Cook, 2011; St George, Mannes, & Hoffman, 1997; Whitney, Ritchie, & Clark, 1991), consistent with our *working memory and predictive inference hypothesis* and results.

Second, this study used a single well-known scene from *Touch of Evil*. This experimental case-study design mirrored previous seminal studies that implemented a single, or very limited number, of complex naturalistic stimuli to demonstrate unique cognitive phenomena (Anderson and Pichert 1978, Yarbus 1967, Magliano, Dijkstra, and Zwaan 1996). Nevertheless, relying on limited cases reduces the generalizability beyond those materials. As such, those studies were followed-up using more diverse stimuli to test the generalizability of the

psychological phenomena. In this study, the method allowed for very specific predictions about the back-end processes for this scene. To generalize these results beyond this clip, future work will need to extend these findings using more diverse film clips. Also, given the complexities of controlling for a variety of factors that can affect processing, but not of experimental interest, future research may require the development of materials generated by experimenters.

This study did not take into account other factors that affect segmentation. For example, there is considerable research that shows that event boundaries in narrative film are perceived when the film depicts situational changes (e.g., jumps in narrative time, changes to new locations, the introduction of new characters (Zacks et al. 2009, Huff, Meitz, and Papenmeier 2014, Magliano, Miller, and Zwaan 2001). It's possible that one would need to account for these factors to adequately assess the relationship between segmentation and prediction in narrative film. This was not done in the present study because the scene did not contain the situational factors that are specified by theory to be important for narrative comprehension (e.g., Zwaan and Radvansky 1998). For example, the scene takes place in one location, is temporally contiguous, and the characters maintain the same goal. As such, there was no theoretical-based approach for identifying important situational changes. However, exploring the relationship between segmentation and prediction in narrative film using longer film sequences that contain situational changes would afford partialling out variance accounted for by monitoring changes in situational continuity, thus making it possible to account for any uniquely shared variance between segmentation and the prediction of narrative events.

Conclusion

For a little over 100 years, filmmakers have been developing a practical skillset to create engaging cinema, and the popularity of film shows its effectiveness. This study drew on one of

the most famous examples of effective filmmaking to test if it worked the way the filmmakers predicted, using a cognitive framework of scene perception and event comprehension (SPECT). The results showed that, at least in our tests, the opening scene of *Touch of Evil* does work as the filmmakers intended. Specifically, the audio track used likely helped viewers maintain a key piece of narrative information, the car and the time-bomb it contained, in working memory, and the dialog containing the "ticking comment" likely helped reinstate the bomb in viewers' event models. Theoretically, this is required for the scene to have its intended effect—engendering continual suspense over the duration of the long shot. Importantly, maintaining knowledge of the bomb in viewers' event models did not influence their online processing of the scene, as measured through event segmentation, but the perceived agents of the narrative did influence event model construction. In conclusion, we have empirically demonstrated what Alfred Hitchcock called the "public…. Participation in the [suspense] scene" (Truffaut, Hitchcock, and Scott 1984) by showing that cinematic suspense can be created in viewers' working memory, given the right audiovisual memory cues together with individuals' working memory capacities.

References

Albrecht, J. E., and J. L. Myers. 1998. "Accessing distant text information during reading: Effects of contextual cues." *Discourse processes* 26 (2-3):87-107.

Anderson, R. C., and J. W. Pichert. 1978. "Recall of previously unrecallable information following a shift in perspective." *Journal of Verbal Learning and Verbal Behavior* 17 (1):1-12. doi: 10.1016/s0022-5371(78)90485-1.

Bach, M. 2006. "The Freiburg Visual Acuity Test-Variability unchanged by post-hoc re-analysis." *Graefe's Archive for Clinical and Experimental Ophthalmology* 245 (7):965-971. doi: 10.1007/s00417-006-0474-4.

Baddeley, A. 1992. "Working memory." *Science* 255 (5044):556-559. doi: 10.1126/science.1736359.

Batten, J. P., and T. J. Smith. 2018. "Looking at sound: sound design and the audiovisual influences on gaze." In *Seeing into Screens: Eye Tracking the Moving Image*, edited by T. Dwyer, C. Perkins, S. Redmond and J. Sita, 85-102. London, UK: Bloomsbury Publishing.

Bläsing, Bettina E. 2015. "Segmentation of dance movement: effects of expertise, visual familiarity, motor experience and music." *Frontiers in psychology* 5:1500.

Bordwell, D. 1985. *Narration in the fiction film*. Madison, WI: University of Wisconsin Press.

Bregman, A. S. 1994. *Auditory scene analysis: The perceptual organization of sound*: MIT press.

Brewer, W. F., and K. Ohtsuka. 1988. "Story structure, characterization, just world organization, and reader affect in American and Hungarian short stories." *Poetics* 17 (4-5):395-415.

Calvo, M. G. 2001. "Working memory and inferences: Evidence from eye fixations during reading." *Memory* 9 (4-6):365-381.

Case, Robbie, D. Midian Kurland, and Jill Goldberg. 1982. "Operational efficiency and the growth of short-term memory span." *Journal of Experimental Child Psychology* 33 (3):386-404. doi: http://dx.doi.org/10.1016/0022-0965(82)90054-6.

Chion, M. 1994. *Audio-vision: Sound on Screen, trans*. Translated by Claudia Gorbman, . New York: Columbia UP.

Clark, H. H. 1977. "Inferences in comprehension." In *Basic processes in reading: Perception and comprehension*, edited by D. LaBerge and S. J. Samuels, 243-263. Hillsdale, NJ: Earlbaum.

Clinton, J. A., S. W. Briner, A. M. Sherrill, T. Ackerman, and J. P. Magliano. 2017. "The role of cinematic techniques in understanding character affect." *Scientific Study of Literature* 7 (2):177-202.

Comito, T. 1986. "Welles Labyrinths: Introduction to a Touch of Evil." *Avant Scene Cinema* (346-47):6-34.

Coutrot, A., and N. Guyader. 2016. "Multimodal saliency models for videos." In *From Human Attention to Computational Attention*, 291-304. Springer.

D'Angelo, M. 2012. "Touch of Evil." Onion Inc., accessed April 20th, 2017. http://www.avclub.com/article/touch-of-evil-71058.

Daneman, M., and P. A. Carpenter. 1980. "Individual differences in working memory and reading." *Journal of Verbal Learning and Verbal Behavior* 19 (4):450-466.

Dick, F., S. Krishnan, R. Leech, and A. P. Saygin. 2016. "Environmental sounds." In *Neurobiology of Language*, 1121-1138. Elsevier.

Gernsbacher, M. A. 1990. *Language comprehension as structure building*. Vol. xi. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Gernsbacher, M. A., K. R. Varner, and M. E. Faust. 1990. "Investigating differences in general comprehension skill." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16 (3):430-445.

Graesser, A. C., and S. P. Franklin. 1990. "QUEST: A cognitive model of question answering." *Discourse processes* 13 (3):279-303.

Graesser, A. C., M. Singer, and T. Trabasso. 1994. "Constructing inferences during narrative text comprehension." *Psychological Review* 101 (3):371-395.

Henderson, J. M., and A. Hollingworth. 1999. "High-level scene perception." *Annual Review of Psychology* 50:243-271. doi: 10.1146/annurev.psych.50.1.243.

Huff, M., Tino GK Meitz, and F. Papenmeier. 2014. "Changes in situation models modulate processes of event perception in audiovisual narratives." *Journal of Experimental Psychology: Learning Memory and Cognition* 40 (5):1377–1388.

Hutson, J. P., J. P. Magliano, and L. C. Loschky. 2018. "Understanding Moment-to-moment Processing of Visual Narratives." *Cognitive Science* 42:2999-3033. doi: 10.1111/cogs.12699.

Hutson, J. P., T. J. Smith, J. P. Magliano, and L. C. Loschky. 2017. "What is the role of the film viewer? The effects of narrative comprehension and viewing task on gaze control in film." *Cognitive Research: Principles and Implications* 2 (1):46. doi: 10.1186/s41235-017-0080-5.

Just, M. A., and Patricia A. Carpenter. 1992. "A capacity theory of comprehension: Individual differences in working memory." *Psychological Review* 99 (1):122-149.

King, J., and M. A. Just. 1991. "Individual differences in syntactic processing: The role of working memory." *Journal of memory and language* 30 (5):580-602.

Kintsch, W. 1988. "The role of knowledge in discourse comprehension: A construction-integration model." *Psychological Review* 95 (2):163-182.

Kurby, C. A., and J. M. Zacks. 2008. "Segmentation in the perception and memory of events." *Trends in Cognitive Sciences* 12 (2):72.

Levin, D. T., and D. J. Simons. 1997. "Failure to detect changes to attended objects in motion pictures." *Psychonomic Bulletin & Review* 4 (4):501-506.

Lichtenstein, Edward H., and W. F. Brewer. 1980. "Memory for goal-directed events." *Cognitive Psychology* 12 (3):412-445.

Liefooghe, B., P. Barrouillet, A. Vandierendonck, and V. Camos. 2008. "Working memory costs of task switching." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34 (3):478.

Loschky, L. C., A. M. Larson, J. P. Magliano, and T. J. Smith. 2015. "What Would Jaws Do? The Tyranny of Film and the Relationship between Gaze and Higher-Level Narrative Film Comprehension." *PLoS ONE* 10 (11):1-23. doi: 10.1371/journal.pone.0142474.

Loschky, L. C., A. Larson, T. J. Smith, and J. P. Magliano. 2020. "The scene perception & event comprehension theory (SPECT) applied to visual narratives." *Topics in Cognitive Science*:1-41.

Magliano, J. P., K. Dijkstra, and R. A. Zwaan. 1996. "Generating predictive inferences while viewing a movie." *Discourse Processes* 22 (3):199 - 224.

Magliano, J. P., and A. C. Graesser. 1991. "A three-pronged method for studying inference generation in literary text." *Poetics* 20 (3):193-232. doi: http://dx.doi.org/10.1016/0304-422X(91)90007-C.

Magliano, J. P., K. Higgs, and J. Clinton. 2019. "6 Sources of Complexity in Narrative Comprehension across Media." *Narrative Complexity: Cognition, Embodiment, Evolution*:149.

Magliano, J. P., L. C. Loschky, J. A. Clinton, and A. M. Larson. 2013. "Is reading the same as viewing? An exploration of the similarities and differences between processing text- and visually based narratives." In *Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension*, edited by B. Miller, L. Cutting and P. McCardle, 78-90. Baltimore, MD: Brookes Publishing Co.

Magliano, J. P., J. Miller, and R. A. Zwaan. 2001. "Indexing space and time in film understanding." *Applied Cognitive Psychology* 15 (5):533-545.

Manfredi, Mirella, N. Cohn, Mariana De Araújo Andreoli, and Paulo Sergio Boggio. 2018. "Listening beyond seeing: Event-related potentials to audiovisual processing in visual narrative." *Brain and Language* 185:1-8. doi: https://doi.org/10.1016/j.bandl.2018.06.008.

McKoon, G., and R. Ratcliff. 1998. "Memory-based language processing: Psycholinguistic research in the 1990s." *Annual Review of Psychology* 49 (1):25-42.

McNamara, D. S., and J. P. Magliano. 2009. "Toward a comprehensive model of comprehension." In *Psychology of Learning and Motivation*, edited by B. H. Ross, 297-384. New York, NY: Elsevier Science.

Meitz, T. G. K., H. S. Meyerhoff, and M. Huff. 2019. "Event related message processing: Perceiving and remembering changes in films with and without soundtrack." *Media Psychology*:1-31.

Myers, J. L., and E. J. O'Brien. 1998. "Accessing the discourse representation during reading." *Discourse Processes* 26 (2-3):131-157.

Radvansky, G. A. 2012. "Across the event horizon." *Current Directions in Psychological Science* 21 (4):269-272.

Radvansky, G. A., and J. M. Zacks. 2014. *Event Cognition*: Oxford University Press.

Rai, M. K., L. C. Loschky, and R. J. Harris. 2014. "The effects of stress on reading: A comparison of first language versus intermediate second-language reading comprehension." *Journal of Educational Psychology*. doi: http://dx.doi.org/10.1037/a0037591.

Rai, M. K., L. C. Loschky, R. J. Harris, N. R. Peck, and L. G. Cook. 2011. "Effects of stress and working memory capacity on foreign language readers' inferential processing during comprehension." *Language learning* 61 (1):187-218. doi: 10.1111/j.1467-9922.2010.00592.x.

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

Sargent, Jesse Q., Jeffrey M. Zacks, David Z. Hambrick, Rose T. Zacks, Christopher A. Kurby, Heather R. Bailey, Michelle L. Eisenberg, and Taylor M. Beck. 2013. "Event segmentation ability uniquely predicts event memory." *Cognition* 129 (2):241-255. doi: http://dx.doi.org/10.1016/j.cognition.2013.07.002.

Smith, Murray. 1995. "Film spectatorship and the institution of fiction." *Journal of Aesthetics and Art Criticism* 53 (2):113-127.

Smith, T. J. 2014. "Audiovisual correspondences in Sergei Eisenstein's Alexander Nevsky: a case study in viewer attention." In *Cognitive Media Theory (AFI Film Reader)*, edited by P. Taberham and T. Nannicelli. Abingdon, UK: Taylor & Francis.

Smith, T. J., and J. M. Henderson. 2008. "Edit Blindness: The relationship between attention and global change blindness in dynamic scenes." *Journal of Eye Movement Research* 2 (2:6):1-17.

Smith, T. J., and J. Y. Martin-Portugues Santacreu. 2017. "Match-Action: The Role of Motion and Audio in Creating Global Change Blindness in Film." *Media Psychology* 20 (2):317-348. doi: 10.1080/15213269.2016.1160789.

St George, M., S. Mannes, and J. E. Hoffman. 1997. "Individual differences in inference generation: An ERP analysis." *Journal of Cognitive Neuroscience* 9 (6):776-787.

Truffaut, F., A. Hitchcock, and G. S. Scott. 1984. *Hitchcock*. Edited by Revised Edition. New York City: Simon & Schuster.

Tully, T. 1999. "The Sounds of Evil." accessed February 27th, 2020.

Turner, Marilyn L., and Randall W. Engle. 1989. "Is working memory capacity task dependent?" *Journal of Memory and Language* 28 (2):127-154. doi: http://dx.doi.org/10.1016/0749-596X(89)90040-5.

Van den Broek, P. 1990. "The causal inference maker: Towards a process model of inference generation in text comprehension." *Comprehension processes in reading*:423-445.

Welles, O. 1998. Touch of Evil. Universal Studios.

Welles, O., and A. Zugsmith. 1958. Touch of Evil [Film]. Universal Pictures.

Whitney, P., B. G. Ritchie, and M. B. Clark. 1991. "Working-memory capacity and the use of elaborative inferences in text comprehension." *Discourse processes* 14 (2):133-145.

Yarbus, Alfred L. 1967. *Eye movements and vision*. Translated by Basil Haigh. New York, NY, US: Plenum Press.

Zacks, J. M. 1999. *Event Structure Perception Studies in Perceiving, Remembering, and Communicating*: Stanford University.

Zacks, J. M. 2004. "Using movement and intentions to understand simple events." *Cognitive Science* 28 (6):979-1008.

Zacks, J. M., T. Braver, M. Sheridan, D. Donaldson, A. Snyder, J. Ollinger, R. Buckner, and M. Raichle. 2001. "Human brain activity time-locked to perceptual event boundaries." *Nature Neuroscience* 4 (6):651-655. doi: 10.1038/88486.

Zacks, J. M., S. Kumar, R. Abrams, and R. Mehta. 2009. "Using movement and intentions to understand human activity." *Cognition* 112 (2):201-216. doi: 10.1016/j.cognition.2009.03.007.

Zacks, J. M., N. Speer, and J. Reynolds. 2009. "Segmentation in reading and film comprehension." *Journal of Experimental Psychology: General* 138 (2):307-27. doi: 2009-05547-010 [pii] 10.1037/a0015305.

Zacks, J. M., and K. Swallow. 2007. "Event segmentation." *Current Directions in Psychological Science* 16 (2):80-84. doi: 10.1111/j.1467-8721.2007.00480.x.

Zwaan, R. A., and G. A. Radvansky. 1998. "Situation models in language comprehension and memory." *Psychological Bulletin* 123 (2):162-185.

[1] Twenty-two participants were removed for not reporting any predictive inference at the end of the experiment. It is unclear why there was a relatively high dropout rate for this online experiment. We did not have a method for identifying where participants dropped out of the experiment in the platform we used, so we cannot speculate as to why some participants did not complete the experiment. That said, the dropout rate did not appear to affect the results, because the inference results from this experiment were replicated in the following in-person experiments.

[2] The unequal number of participants in each condition occurred because not all participants completed the online study, and those participants who dropped out were not included in any analyses.
[3] https://osf.io/69hjm/?view_only=a707f60f1b7c4e05a70efa81c185b66a
[4] To match participants between studies, we asked them to provide a student ID number. However, the correct ID number was not always given in both the prior working memory experiment and the current experiment.
[5] A limitation of this analysis is that it only used the agreement for participants with a working memory score. As a final test of whether participants in the Context condition who did versus did not make a predictive inference about the bomb had differences in their segmentation agreement, we ran agreement analyses with the three groups of participants (Context+Inference, Context+No-inference, and No-context). The analysis was run three times, with reference baselines for each of the groups in the analysis (e.g., one analysis used the Context+Inference group as the agreement reference group, another analysis used Context+No-inference as the reference group, and the final used No-context). A leave-one-out procedure avoided participants being compared to themselves. These results were are consistent with the above results. The only significant effect was that No-context participants had higher agreement when compared to their reference group. Context+Inference and +No-inference did not significantly differ from each other for any of the baselines used, and never had significantly greater agreement than the No-context condition.