Viewing Static Visual Narratives Through the Lens of the Scene Perception and Event Comprehension Theory (SPECT)

Lester C. Loschky, John P. Hutson, Maverick E. Smith

Kansas State University, USA

Tim J. Smith

Birkbeck, University of London, UK

Joseph P. Magliano

Northern Illinois University, USA

To Appear in "Empirical Comics Research: Digital, Multimodal, and Cognitive Methods"

Edited by Jochen Laubrock, Janina Wildfeuer, and Alexander Dunst

Author Note

Lester C. Loschky, Department of Psychological Sciences, Kansas State University, USA

We would like to thank Adam M. Larson, Morton Gernsbacher, and G. A. Radvansky for helpful discussions of SPECT.

Correspondence concerning this article should be addressed to Lester C. Loschky, 471 Bluemont Hall, 1114 Mid-Campus Drive North, Manhattan, KS 66506-5302, USA

Contact: loschky@ksu.edu

10 Key words:

1. event segmentation -- the process by which people parse the continuous activities into meaningful events (e.g., “washing dishes” includes scrubbing dishes, rinsing dishes, drying dishes)
2. bridging inference -- the process of inferring a link between two events that explains their relationship (e.g., “John dropped the vase. He went to the kitchen to get a dust pan and broom.” We can explain the second sentence by inferring that the vase broke into many pieces.)
3. event model -- a person’s understanding of what is happening right now (e.g., “Looking out my front door, I see a woman walking her dog.”)
4. scene perception -- the application of the last 150 years of perception research to understand how people perceive the real world
5. event perception -- how people perceive and make sense of events in the real world or in narratives
6. scene gist -- the rapid holistic understanding of a scene gained within a single eye fixation (e.g., “a beach,” “a street,” or “a woman and her dog”)
7. (eye) fixation -- the brief time period when the eyes are still while looking at something (usually about 225-330 milliseconds, while reading, or looking at scenes)
8. Saccade -- the brief time period when the eyes are moving from one fixation to the next (usually about 50 milliseconds)
9. working memory -- short term memory used to both briefly store and processes a limited amount of information

10. long-term memory -- the unlimited store of memories that are not in short-term memory or working memory (i.e., anything you remember more than 30 seconds to a minute ago)

Abstract

This paper briefly sketches out the Scene Perception & Event Comprehension Theory (SPECT) and reports on tests of the theory in a series of studies using the "Boy, Dog, Frog" (BDF) wordless picture stories (e.g., Mayer 1969). SPECT is an integrative framework synthesizing a number of theories from the areas of scene perception, event perception, and narrative comprehension. SPECT distinguishes between front-end mechanisms that involve *information extraction* and *attentional selection* during single eye fixations, and back-end mechanisms that involve creating the event models (one's current understanding) across multiple fixations in working memory and storing them in long-term memory. The chief back-end mechanisms are laying the foundation for the event model, mapping incoming information to it, and shifting to create a new event model. In the BDF studies reported, we show evidence from event segmentation and bridging inference generation data for the generalizability of these back-end mechanisms, originally proposed for text comprehension, to visual narratives. We also show some apparent differences from text processing due to the visual narrative modality. We then report tests of novel hypotheses from SPECT about the bidirectional interactions of the front-end and back-end processes. These include the changes to eye movements due to 1) laying the foundation for the event model, and 2) generating bridging inferences while mapping incoming information to the event model.

Viewing Visual Narratives Through the Lens of the Scene Perception and Event Comprehension Theory (SPECT)

1. Introduction In everyday life in the information age, we are continuously presented with visual information, much of it in the form of stories, or narratives. The visual narratives we consume every day come in many forms, from picture stories read by children, to the comics read by adolescents and adults, to the advertisements, TV shows, and movies watched by people of all ages. But, while we have learned a great deal about reading and comprehending textual narratives over the last 100+ years of research (Rayner 1998, Rayner et al. 2001, McNamara and Magliano 2009, Kintsch 1998), we know far less about how we perceive and comprehend visual narratives (see Cohn 2013). Thus, in this paper, we ask, how does moment-to-moment processing of visual narratives progress, both in terms of perceptual and comprehension mechanisms? And, importantly, how are these perceptual and comprehension mechanisms coordinated? In doing so, we frame our discussion in terms of the Scene Perception and Event Comprehension Theory (SPECT)(Loschky et al. 2014, June, Loschky, Magliano, and Smith 2016, July, Loschky et al. in preparation), which we use as a lens to describe research that we have done using visual narratives. The studies we describe have used a particular set of visual narratives by Mercer Mayer, collectively known by many researchers as the "Boy, Dog, Frog" stories (Mayer 1967, 1969, Mayer and Mayer 1971, Mayer 1973, 1974, 1975). We use those studies to test hypotheses generated by SPECT. Nevertheless, a full account of the theory is beyond the scope of the current paper, and can be found elsewhere (Loschky et al. in preparation). As a small caveat for this volume on Research on Comics, we note that while the Boy, Dog, Frog (BDF) picture stories are visual narratives, which are very similar to comics,

they do not have multiple panels on a page and do not have the layout of those panels. Nevertheless, we feel that readers interested in comics perception and comprehension will have a similar interest in our studies. And likewise, we believe that our studies using picture stories may also be relevant to readers interested in other forms of visual narratives, including film, which share similar demands on visual scene perception and sequential visual narrative comprehension even if they differ radically from static visual narrative sequences in many critical ways (including image motion, non-self-paced viewing, and audio).

The Scene Perception and Event Comprehension Theory, or SPECT for short, is a new integrative theoretical framework for understanding both visual narrative perception and comprehension (Loschky et al. 2014, June, Loschky, Magliano, and Smith 2016, July, Loschky et al. in preparation). It incorporates theories from the domains of scene perception (Henderson and Hollingworth 1999, Wolfe et al. 2011, Irwin 1996, Oliva 2005), event perception (Kurby and Zacks 2008a, Zacks, Speer, and Reynolds 2009, Newtson 1973), and narrative comprehension (McNamara and Magliano 2009, Kintsch 1998, Gernsbacher 1990). SPECT has been developed to create bridges across these normally siloed (or, encapsulated) research areas, as a framework for understanding visual narrative perception and comprehension from the onset of the first image in a visual narrative until recall of the entire narrative years later. To our knowledge, SPECT is the first theory to attempt such a grand synthesis across these disparate areas of research and theory. Although we only briefly outline SPECT here, its utility in developing and testing hypotheses within the BDF studies below indicates the importance of developing a comprehensive theory of visual narrative perception and comprehension.

As shown in Figure 1, SPECT draws a key distinction between front-end and back-end mechanisms (see also Magliano et al. 2013). Front-end mechanisms occur during single eye

fixations, whereas back-end mechanisms occur across multiple fixations in working memory (WM) and long-term memory (LTM) over extended periods of time. Importantly, SPECT assumes that front-end and back-end mechanisms interact and influence each other bidirectionally. There are two primary front-end mechanisms that occur during each eye fixation: 1) information extraction from the image, and 2) attentional selection for where to send the eyes for the next fixation. These mechanisms are assumed to occur in parallel (Findlay and Walker 1999, Laubrock, Cajar, and Engbert 2013).

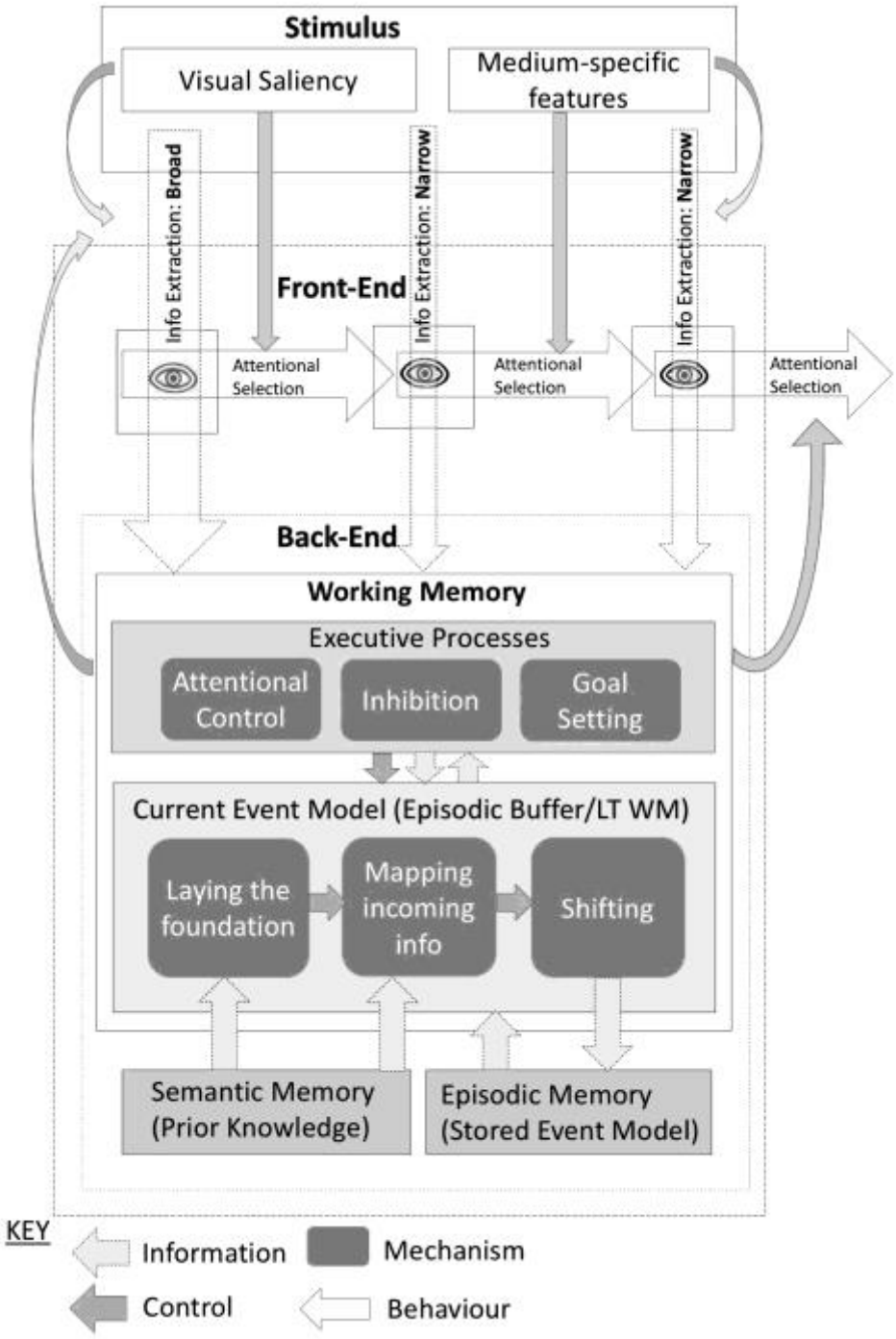

Figure 1: Box model of the Scene Perception & Event Comprehension Theory (SPECT). We thank Adam M. Larson and Morton Gernbsbacher for helping us formulate this box model. All rights reserved for the authors and Adam M. Larson.

The front-end information extraction mechanisms include everything that happens *within a single eye fixation*, from the moment that light from the image first hits the retina, until the recognition of a scene, person, object, or event, roughly 150-300 milliseconds (ms) later (VanRullen and Thorpe 2001, Ramkumar et al. 2016, Cichy et al. 2016, Glanemann 2008, Fei-Fei et al. 2007), which is entered into working memory (WM) in the back-end. Importantly, information extraction can either be *broad*, encompassing the entire scene, and arriving at a holistic semantic representation of it, called *scene gist* (Oliva 2005, Loschky and Larson 2010), or *narrow*, including only information about a specific person, animal, object, or event (Cichy et al. 2016, Glanemann 2008, VanRullen and Thorpe 2001). Broader (coarser) information is typically acquired more quickly than narrower (finer) information (Fei-Fei et al. 2007, Loschky and Larson 2010, Hegde 2008, Schyns and Oliva 1994). For example, broader basic level scene category information (e.g., kitchen) can be extracted in less than a single eye fixation, but narrower basic level actions by a person (e.g., cooking) require two fixations to extract (Larson 2012). Thus, increasingly detailed information can be extracted on a fixation-by-fixation basis in the front-end, which is accumulated *across multiple fixations in WM* (Hollingworth and Henderson 2002, Pertzov, Avidan, and Zohary 2009) in the event model in the back-end.

The front-end attentional selection mechanisms occur in parallel during each fixation, based both on the visual saliency of the stimulus features (Itti and Koch 2001, Mital et al. 2010), medium-specific features (such as the organization of panels in a comic, Cohn 2013, or the compositional features of edited shot sequences in movies, Smith 2012), and back-end guidance (to be described in more detail below).

There are several major classes of back-end mechanisms incorporated in the model: executive mechanisms; mechanisms involved in creating and updating the current event model[1]

in WM; and mechanisms in LTM, including previously stored event models in episodic LTM, and schemas, scripts, and other background knowledge accessed from semantic LTM. The *current event model* is the viewer's understanding of what is happening now in the visual narrative, and is thus of particular interest in SPECT. Key information in the event model is in the form of event indices, including the event (i.e., meta-actions), entities (people, animals, objects), location, time, goals, and causal relationships (Zwaan and Radvansky 1998, Magliano et al. 2011). There are three primary mechanisms involved: 1) laying the foundation for the current event model, 2) mapping new incoming information to it, and 3) shifting to create a new event-model (Gernsbacher 1990).

Laying the foundation is the process of creating a new event model from scratch. Take, for example, the top left image in Figure 2. Laying the foundation begins with recognizing the gist of a scene (i.e., the spatial context, e.g., a forest), what sort of event is occurring (e.g., riding), and the main entities involved, including who is the agent and the patient (e.g., two frogs, on a turtle). These event indices (space, action, agent(s), patient) can generally be extracted within 1-2 fixations (Larson 2012, Larson and Loschky submitted, Glanemann 2008). This information activates associated knowledge from semantic long-term memory, which informs the new event model, producing expectations for upcoming information (Eckstein, Drescher, and Shimozaki 2006, Torralba et al. 2006).

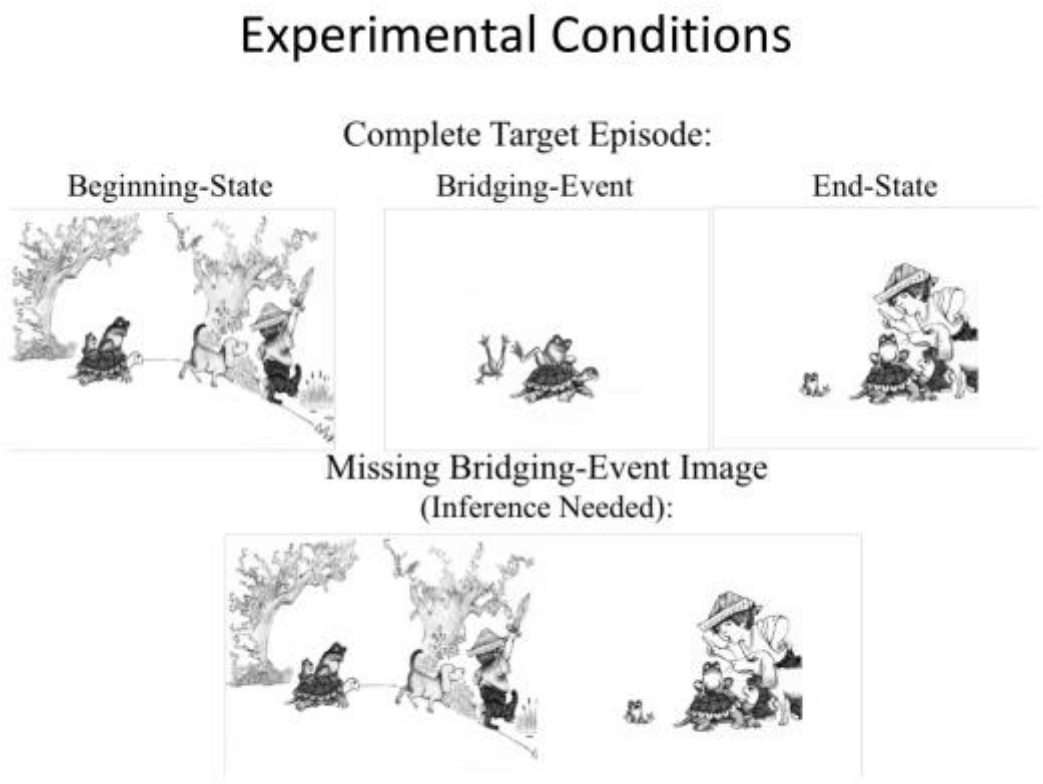

Figure 2. Bridging inference experimental manipulation conditions. A complete 3-image target episode from Figure 2 is shown, including beginning state, bridging event, and end-state images. The missing bridging-event condition requires the viewer to infer the bridging event when they see the end-state image.

Mapping further information to the current event model occurs so long as it is coherent. This involves adding information about other event indices, such as other entities (people, animals, objects), their goals (e.g., the big frog wants to dispose of the little frog), and causal relationships (e.g., the little frog is crying, because it was kicked off of the turtle).

Shifting to create a new event model occurs either when predictions of the current model fail (Zacks et al. 2007), and/or when in-coming information is incoherent with the current model (Gernsbacher 1997, Zwaan, Magliano, and Graesser 1995, Magliano, Zwaan, and Graesser 1999), or one or more important event indices change (Magliano et al. 2011, Huff, Meitz, and Papenmeier 2014). When this happens, the current event model is stored in episodic LTM (Gernsbacher 1985, Swallow, Zacks, and Abrams 2009, Zwaan and Radvansky 1998), becoming a *stored event model*. Stored event models in episodic LTM can then be used to inform the new event model.

Finally, executive processes in WM involve conscious control of processing through goal setting (e.g., be able to summarize the visual narrative), attentional control (e.g., consciously searching for specific information in the narrative), and inhibition (e.g., consciously ignoring certain information deemed irrelevant). Some of these executive processes can be effortful (e.g., goal setting and execution of strategies to achieve those goals), and may not be much involved in normal, seemingly effortless comprehension of visual narratives. However, they can exert strong effects on comprehension to the degree that viewers think of using them, and have the executive resources available to do so (Moss et al. 2011).

An important contribution of SPECT is that it enables a principled investigation of the interactions between the front-end and back-end mechanisms that have not been studied in scene comprehension before, due to artificial boundaries between the research areas of scene perception, event perception, and narrative comprehension. For example, SPECT provides important details to help understand how specific front-end mechanisms, such as gist recognition as a type of broad information extraction, influences specific back-end mechanisms, such as laying the foundation of the event model. SPECT also proposes how specific back-end

mechanisms, for instance, those involved in mapping incoming information to the current event model in WM, can affect specific front-end mechanisms, such as attentional selection. While these sorts of questions have been investigated in the area of reading, they have rarely been investigated in the areas of scene perception, event perception, or the comprehension of visual narratives (but see Foulsham, Wybrow, and Cohn 2016). Other important questions raised by SPECT have to do with the degree to which back-end processes operate similarly in the comprehension of visual narratives as they do in the comprehension of textual narratives, despite the large differences in media, or whether they differ, and if so, how.

2. What is the Nature of Back-end Mechanisms in "Reading" Picture Stories?

SPECT assumes that the back-end processes involved in the construction of the event model are general cognitive mechanisms, which extend beyond the language based modality to include visual narratives. This assumption of SPECT has been supported by the data from our studies, which have shown evidence consistent with three stages of building the current event model: laying the foundation, mapping, and shifting (Hutson, Magliano, and Loschky in preparation, Magliano et al. in press, Magliano et al. 2011, Magliano et al. 2016, Smith et al. in preparation-b, a). There is a growing body of research that has been conducted on each of these back-end mechanisms in the context of visual narratives. We describe studies that we have conducted, primarily using Mayer's BDF stories.

There are two distinct processes we have studied that support the three phases of building the current event model, namely event segmentation and bridging inference generation (Magliano et al. 2013). Event segmentation is the process of detecting boundaries between large events, or between sub-events that comprise larger events as illustrated in Figure 3. Figure 3

shows eight pictures from the beginning of a BDF story, three of which have dashed lines around them indicating that they were normatively segmented. The first two segmented pictures and the third segmented picture appear to be at different levels of granularity. The third segmented picture appears to a coarse event boundary marking the beginning of a new large event, because there has been a large change in location, time, actions, and goals (i.e., from inside a house to a forest, at a later time needed to get there, with the characters traveling, for some purpose unknown). The first two segmented pictures appear to be sub-events within the larger initial event of receiving a present and ensuing interactions among the characters, with the sub-events being the opening of the present, and the discovery of a new frog. Segmenting events in this way enables the reader to treat each sub-event or larger event as a single chunk of information, thus saving processing resources and improving comprehension when done well (e.g., Kurby and Zacks 2008b, Zacks, Speer, and Reynolds 2009, Magliano et al. 2011). It appears to be an obligatory aspect of event cognition in both real world and fictive events (e.g, Radvansky and Zacks 2014, Newtson 1973). Segmentation can be incremental or global in nature (Kurby and Zacks 2012). If a person or character is engaged in a goal-plan, changes in their behaviors to achieve that goal indicate incremental boundaries that move the agent towards completion of the goal (e.g., opening the present, and discovering the new frog). However, behaviors that suggest that a prior goal has been completed and a new, unrelated goal, has been initiated (e.g., the boy, dog, frogs, and turtle go for a walk in the woods) indicate global shifts (Magliano et al. 2011). Incremental boundaries may signal that mapping processes should be initiated because the boundaries still reflect events that are part of the current event, whereas global boundaries should initiate shifting to a new event model. Inferences support mapping (Gernsbacher 1990, Graesser, Singer, and Trabasso 1994), and therefore incremental boundaries may be a signal that that

inferential processes that support mapping should be engaged. However, this claim has yet to be directly tested, to our knowledge.

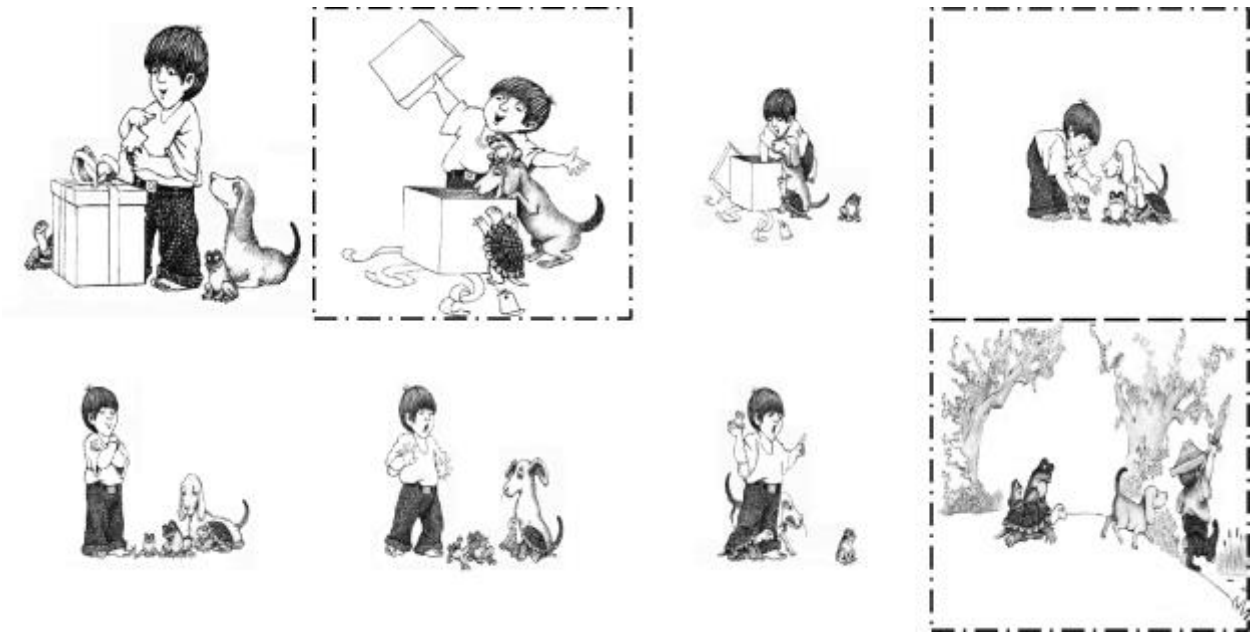

Figure 3: Illustration of event segmentation task in a BDF story (adapted from "One Frog Too Many," Mayer 1975). After participants read the entire story, they saw thumbnails of all the images in sequential order. Their task was to click on each picture when they thought a change in the story situation occurred. Participants were not given an example of a change in situation, but were informed that stories contain many changes although not every picture corresponds to one. Upon clicking a thumbnail image, the participant sees an outline around it. Normative event segmentation for this story is indicated in this illustration by dashed lines around the most frequently segmented images.

3. The Mechanism of *Laying the Foundation*: The Role of Front-end Processes

When someone sees the first image in a visual narrative, according to SPECT, they must lay the foundation for a new event model (i.e., create a new event model). Research has found

that when reading a textual narrative, reading times on the first sentence are longer than on subsequent sentences (Haberlandt and Graesser 1985), and the same is true when viewing short comic strips (Foulsham, Wybrow, and Cohn 2016). In studies using the BDF stories, results have shown that the longest viewing times of any image are on the very first image in a story (Gernsbacher 1983, Smith et al. in preparation-b), which is similar to what has been found for first sentences of a textual story episode (Haberlandt, Berian, and Sandson 1980). According to SPECT, this is due to the process of laying the foundation for the new event model. Importantly, SPECT goes into some detail about how front-end processes lay the foundation. As noted earlier, SPECT argues for two chief front-end mechanisms that occur in parallel, *information extraction* and *attentional selection*. In laying the foundation for a new event model, on the very first fixation, the information extraction process plays a critically important role. On that first fixation, the viewer perceives holistic semantic information about the scene in a coarse-to-fine manner: going from its subordinate level scene category (e.g., an indoor scene), to its basic level category (e.g., a kitchen). This information activates expectancies for where important information is in such a scene, which affects attentional selection on that first fixation (Eckstein, Drescher, and Shimozaki 2006). According to SPECT, recognizing the spatial context of the event (e.g., a kitchen) is a foundational event index for the new event model. Following that first fixation, the first eye movement in a image is usually made to something highly informative, particularly a person (if there is one), and by the end of the second fixation, the viewer will have recognized what that person is doing (i.e., the basic level action category, e.g., cooking) (Larson 2012, Larson and Loschky submitted). Having recognized the event indices of the spatial context, an entity (e.g., a potential protagonist), and the entity's action (the core event index), the viewer is well on their way to having laid the foundation for the new event model, and all within

the time span of two eye fixations. Importantly, the three back-end event model mechanisms of laying the foundation, mapping, and shifting, occur cyclically. Thus, after shifting to create a new event model, the viewer begins the cycle again by laying the foundation for the next event model.

When viewers look at a new, randomly selected, static image, their eye movements progress through two processing stages: 1) the *ambient mode*, characterized by longer saccade lengths and shorter fixation durations as viewers begin by exploring the image to figure out where the major scene constituents are (Pannasch et al. 2008, Smith and Mital 2013); and next, 2) the *focal mode*, characterized by shorter saccade lengths, and longer fixation durations, as viewers extract detailed information from specific scene constituents of interest. A prediction of SPECT is that viewers will go through ambient-to-focal stages of processing at event boundaries, with the ambient mode used in laying the foundation of the new event model. In our own studies with the BDF stories, we have found that saccade lengths were longer and fixation durations were shorter on images identified as event boundaries, consistent with a switch to the ambient mode at boundaries (Smith et al. in preparation-b). Specifically, both saccade lengths and fixation durations showed significant interactions between boundary/non-boundary and viewing time on an image. For boundary images, mean saccade lengths remained at a constant 5 degrees of visual angle over the time course of viewing, but for non-boundary images they showed roughly a 20% decrease (from 5-4 degrees). For boundary images, mean fixation durations remained shorter, increasing 22% more slowly over the first 20 seconds of viewing (going from 200-350 ms) than for the non-boundary images (going from 200-450 ms). This suggests that over the time period of viewing an image identified as an event boundary, viewers remained in the ambient mode for longer compared to the same time period when viewing non-boundary

images. The latter results are consistent with the hypothesis generated by SPECT that at the beginning of a new event, viewers explore an image in the ambient mode to lay the foundation for the new event model. In terms of the two primary front-end mechanisms of SPECT, the longer saccades would be evidence of a change in attentional selection, and the shorter fixations would be attributed to a change in information extraction, both of which would be related to changes occurring in the current event model, namely shifting and laying the foundation. Nevertheless, saccade lengths and fixation durations, while generally independent, can show interdependencies (Findlay and Walker 1999), and so further studies are needed to test these speculations.

4. The Mechanisms of *Mapping* & *Shifting* in the Context of Picture Stories

After having laid the foundation, according to SPECT, the viewer continues to map new incoming information to the current event model. However, this depends on the degree to which the new information is coherent with the current event model (Gernsbacher 1997, Gernsbacher 1990), the degree of change to the various event indices for the current event model (e.g., new people, places, or goals)(Kurby and Zacks 2012, Magliano, Miller, and Zwaan 2001, Magliano, Taylor, and Kim 2005, Zacks, Speer, and Reynolds 2009), and the degree to which the new information fits with predictions generated by the current event model (Zacks et al. 2007, Zwaan, Langston, and Graesser 1995). New information is incorporated into the current event so long as the new information is coherent, does not change radically, and fits with expectations. However, if there is a lack of coherence, mapping stops, and a shift occurs, which marks the boundary at the end of the current event, and the start of the next one. The chief behavioral measure indicating that shifting has occurred is event segmentation.

Based on prior research showing that viewers perceive event boundaries when there are situational shifts in space, time, characters, goals, and causality, Magliano et al. (2011) conducted a study exploring the similarity in event segmentation across visual and text-based narratives using the BDF stories. Participants either viewed the BDF picture stories or read text-based versions of those stories written by the experimenters such that they conveyed content as close as possible to that in the pictures. They also performed a content analysis of the original illustrated stories to determine when the pictures reflected changes in the event indices of time, space, characters, or goals. While viewing the picture stories or reading the texts, participants also carried out an event segmentation task in which they decided whether a story unit (picture or sentence) was the start of a new event (e.g., Zacks, Speer, and Reynolds 2009)(see Figure 3). Magliano et al. (2011) used the situational change content analysis to determine if changes in the event indices were correlated with segmentation judgments and to determine if there were differences in these relationships as a function of modality (i.e., picture vs. text).

Magliano et al. (2011) found that participants' judgments of event boundaries were significantly correlated with all changes in event indices (i.e., shifts in situational continuity), which is consistent with SPECT in terms of the relationship between the mapping and shifting mechanisms. If there is a large enough change in situational continuity (i.e., changes in event indices), it signals a lack of coherence, which leads to shifting, marking an event boundary. They also found that the magnitude of the correlations were similar across the visual and text-based versions, which indicated that segmentation is similar across visual and text-based narratives. However, there was one interesting difference between the modalities. Viewers' event segmentation decisions were less influenced by changes in the time and space event indices in the picture stories than in the text versions. While this particular result is in need of replication,

one possible explanation for it is that viewers of visual narratives give greater weight to changes in goals than to changes in time and space when judging event boundaries, which has also been found for viewers of a film narrative largely lacking dialog ("The Red Balloon"; Magliano and Zacks 2011).

This raises the question of why such spatio-temporal changes would be weighted more highly by readers of texts than visual narratives. Magliano et al. (2011) speculated that it may have been an artifact of the materials and their translations. Many of the narrative events take place in similar settings (e.g. different parts of a forest, different parts of a park), and therefore the products of the front-end mechanism of broad information extraction (i.e., scene gist processing) could have been very similar across pictures that depicted characters changing locations in these settings, thus attenuating the perceived shifts in space. However, these changes were explicitly marked with temporal adverbs (e.g., a little while later, the next day) and spatial prepositions (e.g., into the woods, into the open field, on the log) in the text-based adaptation. In the context of SPECT, the results of Magliano et al. (2011) confirm the importance of segmentation to both the mapping and shifting mechanisms, and that changes in situational continuity signal changes in the current event model (see also Kurby and Zacks 2012, Magliano, Miller, and Zwaan 2001, Magliano, Taylor, and Kim 2005, Zacks, Speer, and Reynolds 2009).

There are various processing costs that accompany shifting to create a new event model. In reading, when incoming information is coherent with the current model, it is readily mapped onto the event model (Zwaan and Radvansky 1998), requiring minimal processing resources. However, when there is a change on one or more event indices, or an event boundary is passed, processing becomes much more intensive, producing slower reading times (Zacks, Speer, and

Reynolds 2009) and longer fixation durations (Swets and Kurby 2016). Part of this is that, when a shift occurs, the current event model is stored in episodic LTM, analogous to processes involved in *sentence wrap-up*, which have also been found to produce longer readings times at the end of sentences and clauses (Just and Carpenter 1980, Rayner et al. 1989, Rayner, Raney, and Pollatsek 1995). In our studies using the BDF stories, visual narrative "readers" viewed pictures identified as boundaries longer than non-boundary pictures (Smith et al. in preparation-b). Given that SPECT distinguishes information extraction and attentional selection front-end mechanisms, an interesting question is whether these longer viewing times were due to viewers making longer fixation durations (e.g., more intensive information extraction) or making more fixations (more rapid shifts of attention), or both. In fact, as described above, in an eye movement study we found that fixation durations on images identified as boundaries were shorter, not longer, compared to non-boundaries (Smith et al. in preparation-b). In contrast, more fixations were made on those images. This suggests that the back-end processes of shifting and laying the foundation are associated with greater exploration of boundary images as reflected by changes in the front-end processes of information extraction and attentional selection. An important remaining challenge for further research is to firmly distinguish effects due to event wrap-up processes versus those due to laying of the foundation for the next event.

5. The Mechanisms of *Mapping* & *Shifting*: Mapping & Bridging Inference Generation in the Context of Picture Stories

As noted earlier, whether new incoming information from the front-end is mapped onto the current event model depends on how well that new information coheres with it. However, visual narratives, like textual narratives, cannot show everything in their narrative worlds.

Instead, only information that is crucial to the narrative is shown, which creates gaps in the narrative world that the visual story teller assumes will not create comprehension problems for the viewer.[2] Thus, when the viewer of a visual narrative is faced with such a narrative gap, they must decide whether it creates a coherence gap sufficient to warrant shifting to create a new event model, or if a reasonable inference can bridge the gap and maintain the coherence of the current event model (Graesser, Singer, and Trabasso 1994). Thus, the mapping mechanism in SPECT crucially involves generating bridging inferences to fill the gaps between explicitly shown events, as has been consistently demonstrated with text materials (Clark 1977, Graesser, Singer, and Trabasso 1994, Magliano et al. 1993, Singer and Halldorson 1996).

In the context of this chapter, bridging inferences are important for making connections between panels and pictures in visual narratives. Consider Figure 2, which depicts a three-picture sequence from one of the BDF stories. The first panel shows the boy with his pets, including a big frog and a little frog who are riding on the back of a turtle. As shown in Figure 3, we learn that the boy got a new little pet frog, whom his older bigger frog is jealous of. Thus, in the second picture, the big frog kicks the little frog off of the turtle's back, and in the third picture the boy scolds the big frog. However, what would happen to comprehension of this episode if the second (middle) picture were missing? The viewer would have to infer why the little frog was on the ground crying and the boy was scolding the big frog in order to understand how that end-state picture is related to the current event model. Readers routinely make these bridging inferences to support constructing a coherent event model (i.e., an event model in which story constituents are semantically connected).

In a recent study (Magliano et al. 2016), we showed that, indeed, viewers generate bridging inferences to connect pictures when processing visual narratives. Participants viewed

the six BDF stories, and picture viewing times were collected. An assumption was that viewing times for picture sequences where bridging inferences were needed would be longer than when they were not. To create this situation, we identified 24 three-picture sequences like the one shown in Figure 2, which consisted of a beginning-state, a bridging-event, and an end-state. We then manipulated whether the bridging-event picture was present. In a pilot study, we had participants think aloud following the end-state pictures, and we found that, consistent with a counter-intuitive prediction, viewers were more likely to mention the bridging-event actions when the bridging-event pictures were *missing* than when they were present. Generating the inferred bridging action results in it being more highly activated in the viewer's event model than if it were simply seen. This suggests that when viewers "read" visual narratives silently, they infer bridging events when they are missing (Magliano and Graesser 1991). Also consistent with our predictions, we found that viewing times for the end-state pictures were longer when the bridging-event pictures were absent (on average, ~2,800 ms) than when they were present (on average, ~2,450 ms). This prediction was based on research showing that sentence-reading times are longer when viewers need to generate bridging inferences (e.g., Clark 1977). Together with the think aloud data, this provided converging evidence that viewers do indeed regularly generate bridging inferences online to map across gaps between pictures in visual narratives (c.f., Cohn and Wittenberg 2015). This is also further consistent evidence in favor of the claim that back-end mechanisms operate similarly between visual and textual narratives.

An interesting question pertains to whether and how bridging inference generation processes relate to situational continuities and event segmentation. According to SPECT, this is conceptualized in terms of the relationship between the mapping and shifting mechanisms. If a gap in narrative coherence occurs, and the coherence can be maintained by generating a bridging

inference, then there is no need to shift to create a new event model, thus viewers would not be expected to give a segmentation response. Conversely, if the gap cannot be bridged by an inference or if the gap is judged to be too large, then there will be a shift to create a new event model, and the viewer will make a segmentation response.

We have recently investigated this question in an event segmentation study with the BDF stories (Smith et al. in preparation-a). Using the same self-paced viewing time paradigm as Magliano et al. (2016), we manipulated the presence/absence of the bridging event images. After viewing each of the stories, participants identified images in the story where they felt there was a change in the story's situation (i.e., an event boundary), as illustrated in Figure 3. Interestingly, we found that participants were more likely to identify the end-state image as a boundary (likely a global boundary) when the bridging event was absent than when it was present. This is consistent with the idea that viewers were reacting to a perceived lack of coherence.

However, we have already presented evidence consistent with the idea that viewers were generating bridging inferences when pictures were removed (e.g., Magliano et al., 2016), which would allow them to maintain coherence with the current event model, and therefore they would not be expected to make a segmentation response as was found in Smith et al. (in preparation-a). Thus, the result showing increased event segmentation in the bridging event-absent condition presents a theoretical puzzle. One possible solution to this puzzle is to assume that our results are due to a mutually exclusive mixture of different participants making one or the other response (inference, or event segmentation), but not both. In this context, the character goals are still maintained when the bridging event is missing, and therefore these judgments reflect incremental boundaries. As mentioned above, incremental boundaries may be a signal that the

coherence break should be resolved via mapping processes. This is clearly a testable hypothesis that warrants further exploration.

6. How Back-end Mechanisms Affect the Front-end Mechanism of *Attentional Selection* in the Context of Picture Stories

As noted earlier, a unique contribution of SPECT is that it allows us to test novel hypotheses about the interactions between front-end and back-end mechanisms. As shown in Figure 1, SPECT assumes that these interactions are bidirectional. Thus, it is assumed that the front-end mechanisms of information extraction and attentional selection influence back-end mechanisms involved in creating the current event model. That assumption is "baked into" most theories of narrative comprehension without comment. SPECT also generated the more specific hypothesis, described above, that viewers would enter the ambient mode of eye movements at the beginning of a new event, to facilitate laying its foundation. And, as reported above, we found empirical support for this by finding that viewers remain in the ambient mode of processing longer while viewing images identified as boundaries than non-boundaries (Smith et al. in preparation-b). The flip side is the assumption in SPECT that back-end mechanisms involved in creating the current event model influence moment-to-moment processes involved in information extraction and attentional selection in the front-end during eye fixations. These predictions are very novel within the areas of scene perception, event perception, and the comprehension of visual narratives (see also Foulsham, Wybrow, and Cohn 2016).

Hutson, Magliano, and Loschky (in preparation) tested the assumption that front-end information extraction and attentional selection processes are sensitive to the back-end mapping mechanism sub-process of inference generation. Magliano et al. (2016) showed that generating a

bridging inference during picture story viewing resulted in longer viewing times for end-state images. Using the same manipulation of bridging event presence/absence to induce the mapping sub-process of inference generation, viewers' eye movements were measured to test a series of research questions to get at the influence of the back-end on the front-end.

The first question was what eye-movement processes accounted for the increased viewing time during inference generation in Magliano et al. (2016). As noted earlier, increased viewing times can be accounted for by eye movements in terms of either increased fixation durations or more fixations, or both. We proposed a *Computational Load hypothesis* that if viewers were under greater computational load while generating the inference, they would produce longer fixation durations (Just and Carpenter 1980). The competing *Visual Search hypothesis* was that if viewers felt the need to search the scene for information relevant to making a bridging inference, they would produce more fixations.

Somewhat surprisingly, there was essentially zero difference in fixation durations between the bridging event present and absent conditions. There was however a significant and meaningful 20% increase in the number of fixations produced, with viewers on bridging event-absent trials making two additional fixations on the end-state image compared to bridging event-present trials (on average, 11 vs. 9 fixations, respectively). These results were therefore consistent with the Visual Search hypothesis. This raised the follow-up question, what were viewers fixating on those extra fixations? This led to the follow-up *Inferential-Informativeness hypothesis*, that viewers needing to generate a bridging inference to maintain coherence between pictures would preferentially fixate regions that were more informative for drawing the inference.

To measure the inferential-informativeness of scene regions, we ran a rating study with new participants to identify inference relevant regions. In that study, we fully informed participants about the bridging event present/absent manipulation in the eye-tracking experiment, and asked them to click the areas of the end-state scene they thought were relevant for drawing the inference if a participant had not seen the bridging event image. We used these click locations to quantify the inferential-informativeness of regions, and found that for bridging event-absent trials (which needed a bridging inference), viewers were more likely to look at inference relevant regions than in the bridging event-present trials (which needed no inference). Thus, when participants needed to draw an inference, they used additional eye-movements to pick up information relevant to generating that inference. For SPECT, this shows that the back-end mapping mechanism's sub-process of bridging inference generation has an online impact on the front-end mechanism of attentional selection.

Importantly, support for the *Visual Search hypothesis* shows a potential departure of comprehension in scenes from text, which supports the importance of SPECT. In visual narratives, when there is a break in coherence, the pictures available allow for a visual search. In text, regressive eye movements could be considered to be analogous to visual search in a scene. However, in text when readers need to generate a bridging inference, they typically don't use regressive eye movements to search for the relevant information for drawing the inferences; rather, they are more likely to rely simply on activating knowledge in LTM (Singer and Halldorson 1996), producing longer gaze durations on the target items (Myers et al. 2000, O'Brien et al. 1988).[3] This asymmetry in the likelihood of using different processes to generate an inference between textual and visual narratives is likely due to what we refer to as the influence of *medium-specific features* of the stimulus on attentional selection (see Figure 1, top

right). For example, the essential spatiotemporal linearity of attentional selection across multiple fixations while reading text (i.e., mostly rightward, but occasionally leftward) stands in contrast to the much less constrained spatiotemporal dynamics of attention in an image, even if that image is embedded in an ordered image sequence in a visual narrative. This shows the importance of the differences between media types, and how SPECT can benefit future research on visual narratives.

7. Characteristics of the *Stored Event Model* in Long-term Memory in the Context of Picture Stories.

According to SPECT, once a viewer decides to shift to create a new event model, they store the current event model in LTM. This then becomes part of a linked set of stored event models that represent the entire visual narrative, which can be recalled later, for example when retelling the story to a friend. SPECT assumes that the set of stored event models for a given narrative are structured around the goal episodes of characters for both textual and visual narratives (Baggett 1979). We have investigated this issue in terms of understanding the implications of generating inferences on LTM for explicit versus inferred content. It is well documented that memory for narrative texts becomes distorted over time such that memory for explicit content becomes weaker and memory for inferred content becomes stronger (Bartlett 1932, Graesser and Nakamura 1982, Schmalhofer and Glavanov 1986). Magliano and colleagues conducted a study that shows that bridging inferences that connect pictures in a visual narrative distort LTM relatively quickly (Magliano et al. in press).

Magliano et al. (in press) had participants view the BDF stories and manipulated whether the beginning-state, bridging-event, and end-state pictures of the target episodes of Magliano et

al. (2016) where shown. We measured participants' picture viewing times during viewing, and after participants had viewed all six stories, gave them a picture recognition memory task.[4] In this study, we measured viewing times for the pictures following beginning-state, bridging-event, and end-state pictures, and like Magliano et al. (2016), assumed that finding longer viewing times on those picture when the preceding picture was missing than when it was present indicated that the missing event was inferred. We found that there were longer viewing times when the beginning- and bridging-state pictures were missing than when they were present, but this was not the case for end-state pictures, suggesting that viewers were less likely to infer end-states. Importantly, performance on the recognition memory task showed that participants were also more likely to falsely remember having seen missing beginning-state and bridging-event pictures than end-state pictures (Magliano et al. in press). These results suggest that generating bridging inferences distorts memory for the content of visual narratives. In terms of SPECT, this shows the workings of processes within the stored event models in LTM.

8. Conclusions

We all experience visual narratives in many different formats, including, but not limited to, comics, picture stories, and movies. The theory and practice of creating visual narratives has far outpaced their empirical study. SPECT introduces a novel theoretical framework for those interested in the growing field of empirical research in visual narratives (See also Cohn 2013 for an alternative, but not mutually exclusive framework for comics). SPECT bridges well developed theories on scene perception, event cognition, and narrative comprehension which have thus far been siloed, with little interchange between areas. By providing an integrative framework encompassing these three areas of research and theory, SPECT offers testable

predictions about the bidirectional relationship between the front-end mechanisms of information extraction and attention selection, and the back-end mechanisms in WM and LTM involved in constructing the current and stored event models.

One novel aspect of SPECT, which makes it a useful theoretical framework for those interested in visual narratives, is that it relies on well-established comprehension mechanisms identified in theories of reading comprehension, and generates predictions for how they operate in visual narrative comprehension. Importantly, while the comprehension processes themselves are based on general cognitive mechanisms, they may require and rely on different front-end mechanisms given the unique characteristics of the visual scene stimuli (Loughlin et al. 2015). As such, SPECT allows for tests of classic comprehension mechanisms (e.g., laying the foundation, shifting, and mapping), but asks how well-established perceptual and attentional mechanisms identified in theories of scene perception may induce or even necessitate visual-narrative-specific processes. Within the SPECT framework, predictions for these relationships are generally presented as the interaction between front-end information extraction and attentional selection mechanisms and back-end event-model construction mechanisms.

The development of SPECT has been carried out through tests of major assumptions of the framework. The Boy, Dog, Frog picture stories are an ideal stimulus set for these tests, because they are purely visual narratives that are fairly easily manipulated (e.g., to require bridging inference generation). (The difficulty of manipulating visual narratives compared to text is one of the major limitations of research in this area.) The Boy, Dog, Frog studies presented here tested how generalizable some important back-end mechanisms, which have been primarily studied in the context of textual narrative comprehension, are to the context of visual narrative comprehension. We found strong generalizability for the relationship between

mapping and shifting mechanisms as a function of the degree of change in event indices, and likewise for the mapping mechanism sub-process of bridging inference generation. Nevertheless, each of these back-end mechanisms showed interesting differences in the context of visual narrative comprehension that could be related to specific front-end visual processes. For example, the front-end mechanism of broad information extraction, or scene gist, was proposed to be behind the difference in the relative importance of spatial changes for event segmentation between text and visual narratives. Likewise, we have proposed that pictorial versus textual medium-specific features place constraints on attentional selection that may explain differences in eye movements during bridging inference generation while viewing visual versus textual narratives. Thus, the above work shows that the back-end event model mechanisms tested (i.e., laying the foundation, inference generation, shifting) are the same ones identified in research on text comprehension, but that due to the differences in the visual narrative stimuli compared to text, research on scene perception introduces important new considerations. This is precisely where SPECT offers unique contributions to further our understanding of comprehension processes in visual narratives, which have become the modal form of narrative consumed for much of the population. The work presented only scratches the surface of how visual narrative comprehension, and there are many fundamental questions SPECT poses that still need to be asked (Loschky et al. in preparation).

Importantly, we have also tested important hypotheses about the proposed bidirectional relationships between the front-end and back-end processes, such as how the front-end processes of attentional selection and information extraction influence laying the foundation for a new event model (Smith et al. in preparation-b), and whether/how the back-end mapping mechanism sub-process of inference generation influences the front-end mechanism of attentional selection

(Hutson, Magliano, and Loschky in preparation). Nevertheless, this work has only begun to touched on these issues, and the need for further research on them is now highlighted by SPECT. For example, how exactly do does the front-end mechanism of information extraction influence the back-end mechanism of laying the foundation for a new event model? Initial investigations of this question (Larson 2012, Larson and Loschky submitted) showed that gist processing of different levels of scene categories (i.e., superordinate level indoor vs. outdoor, and basic level kitchen vs. office) and actions (basic level cooking, vs. washing dishes) occurred at different times scales, going from the coarse superordinate scene category level, within 150 ms processing time, to the fine basic level actions, which required two full eye fixations. Further research has shown that extracting the gist of a scene facilitates recognizing a person's action within that scene (Larson and Lee 2015). These studies illustrate the kinds of studies motivated by hypotheses generated by SPECT. Many more such studies (Loschky et al. in preparation), and eventually computational models will be needed to fully test the theory.

References

Baggett, P. 1979. "Structurally equivalent stories in movie and text and the effect of the medium on recall." *Journal of Verbal Learning & Verbal Behavior* 18 (3):333-356.

Bartlett, F. C. . 1932. *Remembering*. London: Cambridge University Press.

Calvo, M. G., E. Meseguer, and M. Carreiras. 2001. "Inferences about predictable events: Eye movements during reading." *Psychological Research* 65 (3):158-169. doi: 10.1007/s004260000050.

Cichy, Radoslaw Martin, Aditya Khosla, Dimitrios Pantazis, A. Torralba, and A. Oliva. 2016. "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence." *Scientific Reports* 6:27755. doi: 10.1038/srep27755.

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.

Cohn, N. 2013. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*: A&C Black.

Cohn, N., and E. Wittenberg. 2015. "Action starring narratives and events: Structure and inference in visual narrative comprehension." *Journal of Cognitive Psychology* 27 (7):812-828. doi: 10.1080/20445911.2015.1051535.

Eckstein, M. P., B. A. Drescher, and S. S. Shimozaki. 2006. "Attentional cues in real scenes, saccadic targeting, and Bayesian priors." *Psychological Science* 17 (11):973-980. doi: 10.1111/j.1467-9280.2006.01815.x.

Fei-Fei, L., A. Iyer, C. Koch, and P. Perona. 2007. "What do we perceive in a glance of a real-world scene?" *Journal of Vision* 7 (1: 10):1–29. doi: doi:10.1167/7.1.10.

Findlay, J., and R. Walker. 1999. "A model of saccade generation based on parallel processing and competitive inhibition." *Behavioral and Brain Sciences* 22 (4):661-721.

Foulsham, T., Dean Wybrow, and N. Cohn. 2016. "Reading Without Words: Eye Movements in the Comprehension of Comic Strips." *Applied Cognitive Psychology* 30 (4):566-579. doi: 10.1002/acp.3229.

Frazier, L., and K. Rayner. 1982. "Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences." *Cognitive psychology* 14 (2):178-210.

Gernsbacher, M. A. 1985. "Surface information loss in comprehension." *Cognitive Psychology* 17 (3):324-363. doi: 10.1016/0010-0285(85)90012-x.

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

Gernsbacher, M.A. 1997. "Coherence cues mapping during comprehension." In *Processing interclausal relationships: Studies in the production and comprehension of text*, edited by Jean Costermans and Michel Fayol, 3-21. Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc,.

Gernsbacher, Morton A. 1983. "Memory for surface information in non-verbal stories: Parallels and insights to language processes."Ph.D. dissertation, Psychology, University of Texas.

Glanemann, R. . 2008. "To see or not to see--Action scenes out of the corner of the eye." PhD. Dissertation, Psychology, University of Münster.

Graesser, A. C., and G. V. Nakamura. 1982. "The Impact of a Schema on Comprehension and Memory." *Psychology of Learning and Motivation* 16:59-109. doi: http://dx.doi.org/10.1016/S0079-7421(08)60547-2.

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

Haberlandt, K. F., C. Berian, and J. Sandson. 1980. "The episode schema in story processing." *Journal of Verbal Learning and Verbal Behavior* 19 (6):635-650.

Haberlandt, K. F., and A. C. Graesser. 1985. "Component processes in text comprehension and some of their interactions." *Journal of Experimental Psychology: General* 114 (3):357-374.

Hegde, J. 2008. "Time course of visual perception: Coarse-to-fine processing and beyond." *Progress in Neurobiology* 84 (4):405-439. doi: 10.1016/j.pneurobio.2007.09.001.

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.

Hollingworth, A., and J. M. Henderson. 2002. "Accurate visual memory for previously attended objects in natural scenes." *Journal of Experimental Psychology: Human Perception & Performance* 28:113-136.

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

Hutson, J. P., J. P. Magliano, and L. C. Loschky. in preparation. Understanding Moment-to-moment Processing of Visual Narratives. edited by Kansas State University.

Irwin, D. E. 1996. "Integrating information across saccadic eye movements." *Current Directions in Psychological Science* 5 (3):94-100.

Itti, L., and C. Koch. 2001. "Computational modeling of visual attention." *Nature Reviews Neuroscience* 2 (3):194-203.

Just, M. A., and P. A . Carpenter. 1980. "A theory of reading: From eye fixations to comprehension." *Psychological Review* 87 (4):329-354.

Kintsch, W. 1998. *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

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

Kurby, C. A., and J. M. Zacks. 2012. "Starting from scratch and building brick by brick in comprehension." *Memory & Cognition* 40 (5):812-826. doi: 10.3758/s13421-011-0179-8.

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

Larson, A. M. 2012. "Recognizing the setting before reporting the action: Investigating how visual events are mentally constructed from scene images." Ph.D. Ph.D. Dissertation, Psychology, Kansas State University.

Larson, A. M., and M. Lee. 2015. "When Does Scene Categorization Inform Action Recognition?" *Journal of Vision* 15 (12):118-118. doi: 10.1167/15.12.118.

Larson, A. M., and L. C. Loschky. submitted. "From scene perception to event conception: How scene gist informs event perception."

Laubrock, J., A. Cajar, and R. Engbert. 2013. "Control of fixation duration during scene viewing by interaction of foveal and peripheral processing." *Journal of Vision* 13(12):11:1-20. doi: 10.1167/13.12.11.

Laubrock, J., and R. Kliegl. 2015. "The eye-voice span during reading aloud." *Frontiers in Psychology* 6 (1432). doi: 10.3389/fpsyg.2015.01432.

Loschky, L. C., J. P. Hutson, J. P. Magliano, A. M. Larson, and T. J. Smith. 2014, June. "Explaining the Film Comprehension/Attention Relationship with the Scene Perception and Event Comprehension Theory (SPECT)." The 2014 annual meeting of the Society for Cognitive Studies of the Moving Image, Lancaster, PA.

Loschky, L. C., and A. M. Larson. 2010. "The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing." *Visual Cognition* 18 (4):513-536.

Loschky, L. C., A. M. Larson, T. J. Smith, and J. P. Magliano. in preparation. The scene perception and event comprehension theory (SPECT). edited by Kansas State University.

Loschky, L. C., J. P. Magliano, and T. J. Smith. 2016, July. "The Scene Perception and Event Comprehension Theory (SPECT) Applied to Visual Narratives." International Conference on Empirical Studies of Literature and Media, Chicago, IL, USA.

Loughlin, S. M., E. Grossnickle, D. Dinsmore, and P. A. Alexander. 2015. ""Reading" Paintings: Evidence for Trans-Symbolic and Symbol-Specific Comprehension Processes." *Cognition and Instruction* 33 (3):257-293.

Magliano, J. P., William B. Baggett, B. K. Johnson, and A. C. Graesser. 1993. "The time course of generating causal antecedent and causal consequence inferences." *Discourse Processes* 16 (1-2):35-53. doi: 10.1080/01638539309544828.

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. Kopp, K. Higgs, and D. N. Rapp. in press. "Filling in the Gaps: Memory Implications for Inferring Missing Content in Graphic Narratives." *Discourse Processes*:1-14. doi: 10.1080/0163853X.2015.1136870.

Magliano, J. P., K. Kopp, M. W. McNerney, G. A. Radvansky, and J. M. Zacks. 2011. "Aging and perceived event structure as a function of modality." *Aging, Neuropsychology, and Cognition* 19 (1-2):264-282. doi: 10.1080/13825585.2011.633159.

Magliano, J. P., A. M. Larson, K. Higgs, and L. C. Loschky. 2016. "The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension." *Memory & Cognition* 44 (2):207-219. doi: 10.3758/s13421-015-0558-7.

Magliano, J. P., L. C. Loschky, J. 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.

Magliano, J. P., H. A. Taylor, and H-J. J. Kim. 2005. "When goals collide: Monitoring the goals of multiple characters." *Memory & Cognition* 33 (8):1357-1367.

Magliano, J. P., and J. M. Zacks. 2011. "The impact of continuity editing in narrative film on event segmentation." *Cognitive Science* 35 (8):1489-1517. doi: 10.1111/j.1551-6709.2011.01202.x.

Magliano, J. P., R. A. Zwaan, and A. C. Graesser. 1999. "The role of situational continuity in narrative understanding." In *The construction of mental representations during reading*, edited by S. R. Goldman and H. van Oostendorp, 219-245. Mahwah, NJ: Erlbaum.

Mayer, Mercer. 1967. *A boy, a dog, and a frog*. New York, NY, US: Dial Books for Young Readers.

Mayer, Mercer. 1969. *Frog, where are you?* New York, NY, US: Dial Books.

Mayer, Mercer. 1973. *Frog on his own*. New York, NY: Dial Press.

Mayer, Mercer. 1974. *Frog goes to dinner*. New York, NY: Dial Press.

Mayer, Mercer. 1975. *One frog too many*. New York, NY: Dial Press.

Mayer, Mercer, and Marriana Mayer. 1971. *A boy, a dog, a frog and a friend*. New York, NY: Dial Press.

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.

Mital, P. K., T. J. Smith, R. Hill, and J. M. Henderson. 2010. "Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion." *Cognitive Computation* 3 (1):5-24.

Moss, J., C. D. Schunn, W. Schneider, D. S. McNamara, and K. VanLehn. 2011. "The neural correlates of strategic reading comprehension: Cognitive control and discourse comprehension." *NeuroImage* 58 (2):675-686.

Myers, J. L., A. E. Cook, G. Kambe, R. A. Mason, and E. J. O'Brien. 2000. "Semantic and episodic effects on bridging inferences." *Discourse Processes* 29 (3):179-199.

Newtson, Darren. 1973. "Attribution and the unit of perception of ongoing behavior." *Journal of Personality and Social Psychology* 28 (1):28-38.

O'Brien, E. J., D. M. Shank, J. L. Myers, and K. Rayner. 1988. "Elaborative inferences during reading: Do they occur on-line?" *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14 (3):410-420.

Oliva, A. 2005. "Gist of a scene." In *Neurobiology of Attention*, edited by Laurent Itti, Geraint Rees and John K. Tsotsos, 251-256. Burlington, MA: Elsevier Academic Press.

Pannasch, S., J. R. Helmert, K. Roth, A. K. Herbold, and H. Walter. 2008. "Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions." *Journal of Eye Movement Research* 2(2):4:1-19. doi: 10.16910/jemr.2.2.4.

Pertzov, Y., G. Avidan, and E. Zohary. 2009. "Accumulation of visual information across multiple fixations." *Journal of Vision* 9 (10:2):1-12.

Poynor, David V., and Robin K. Morris. 2003. "Inferred goals in narratives: Evidence from self-paced reading, recall, and eye movements." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29 (1):3.

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

Ramkumar, P., B. C. Hansen, S. Pannasch, and L. C. Loschky. 2016. "Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG study." *NeuroImage* 134:295-304. doi: http://dx.doi.org/10.1016/j.neuroimage.2016.03.027.

Rayner, K. 1998. "Eye movements in reading and information processing: 20 years of research." *Psychological Bulletin* 124 (3):372-422. doi: 10.1037//0033-2909.124.3.372.

Rayner, K., Barbara R. Foorman, Charles A. Perfetti, David Pesetsky, and Mark S. Seidenberg. 2001. "How psychological science informs the teaching of reading." *Psychological Science in the Public Interest* 2 (2):31-74.

Rayner, K., Gary E. Raney, and A. Pollatsek. 1995. "Eye movements and discourse processing." In *Sources of coherence in reading*, edited by Robert Frederick Lorch Jr. and Edward J. O'Brien, 9-35. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Rayner, K., Sara C. Sereno, Robin K. Morris, A Rene Schmauder, and Charles Clifton. 1989. "Eye movements and on-line language comprehension processes." *Language & Cognitive Processes* 4:21-50.

Schmalhofer, F., and D. Glavanov. 1986. "Three components of understanding of a programmer's manual: Verbatim, propositional, and situational representations." *Journal of Memory & Language* 25 (3):279-294.

Schyns, P. G., and A. Oliva. 1994. "From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition." *Psychological Science* 5:195-200.

Singer, M., and M. Halldorson. 1996. "Constructing and validating motive bridging inferences." *Cognitive Psychology* 30 (1):1-38.

Smith, M. E., J. P. Hutson, J. P. Magliano, and L. C. Loschky. in preparation-a. Bridging the gap in coherence: Inference generation and event segmentation inform event model construction processes in visual narratives. edited by Kansas State University.

Smith, M. E., J. P. Hutson, J. P. Magliano, and L. C. Loschky. in preparation-b. Laying the foundation for event understanding in picture stories: Evidence from eye movements. edited by Kansas State University.

Smith, T. J. 2012. "The attentional theory of cinematic continuity." *Projections* 6 (1):1-27. doi: 10.3167/proj.2012.060102.

Smith, T. J., and P. K. Mital. 2013. "Attentional synchrony and the influence of viewing task on gaze behaviour in static and dynamic scenes." *Journal of Vision* 13(8):16:1-24. doi: 10.1167/13.8.16.

Swallow, K. M., J. M. Zacks, and R. A. Abrams. 2009. "Event boundaries in perception affect memory encoding and updating." *Journal of Experimental Psychology: General* 138 (2):236-257. doi: 2009-05547-006 [pii] 10.1037/a0015631.

Swets, B., and C. A. Kurby. 2016. "Eye Movements Reveal the Influence of Event Structure on Reading Behavior." *Cognitive Science* 40 (2):466-480. doi: 10.1111/cogs.12240.

Torralba, A., A. Oliva, M. S. Castelhano, and J. M. Henderson. 2006. "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search." *Psychological Review* 113 (4):766-786.

VanRullen, R., and S. J. Thorpe. 2001. "The time course of visual processing: From early perception to decision-making." *Journal of Cognitive Neuroscience* 13 (4):454-461.

Wolfe, J. M., M. L.-H. Võ, K. K. Evans, and M. R. Greene. 2011. "Visual search in scenes involves selective and nonselective pathways." *Trends in Cognitive Sciences* 15 (2):77-84. doi: 10.1016/j.tics.2010.12.001.

Zacks, J. M., N. K. Speer, and J. R. Reynolds. 2009. "Segmentation in reading and film comprehension." *Journal of Experimental Psychology: General* 138 (2):307-327. doi: 2009-05547-010 [pii] 10.1037/a0015305.

Zacks, J. M., N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. 2007. "Event perception: A mind-brain perspective." *Psychological Bulletin* 133 (2):273-293. doi: 2007-02367-005 [pii] 10.1037/0033-2909.133.2.273.

Zwaan, R. A., M. C. Langston, and A. C. Graesser. 1995. "The construction of situation models in narrative comprehension: An event-indexing model." *Psychological Science* 6 (5):292-297. doi: 10.1111/j.1467-9280.1995.tb00513.x.

Zwaan, R. A., J. P. Magliano, and A. C. Graesser. 1995. "Dimensions of situation model construction in narrative comprehension." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (2):386-397.

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

Endnotes

[1] We use the term "event model" consistently with Radvansky and Zacks (2011). They argue that "mental models" are broader, because they are not necessarily "tied to a specific event" (p. 609) and that "situation models" are a narrower sub-type of "event models" that are tied only to representations of text. "Event models" encompass event-specific representations of real life, movies, virtual reality, text, or, we would add, comics or picture stories.

[2] Of course, some visual storytellers want to challenge their viewers, and intentionally create large gaps in coherence for various aesthetic reasons (e.g., in film, "Memento" or "Pulp Fiction"; in comics, "The Sculptor" or "The Multiversity: Pax Americana #1").

[3] One exception to this pattern is that readers sometimes make regressive eye-movements when they are faced with information that is inconsistent with their predictive inferences (e.g., when reading about someone planning for a vacation, and knowing that he is a "sun bird," but then finding that he has decided to go to Alaska for his vacation)(Calvo, Meseguer, and Carreiras 2001, Poynor and Morris 2003). This is likely related to the error processing role that regressions in reading generally play (Frazier and Rayner 1982, Laubrock and Kliegl 2015).

[4] Although this memory retention interval was only a few minutes after having seen each picture story, it is well-beyond the limits of visual short-term memory (i.e., roughly 30 seconds), and is therefore clearly within the realm of LTM.