Young Children's Information Processing of Digital Media Content: The Impact of Media and Child Characteristics

by

Mengguo Jing

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctoral of Philosophy

(Human Development and Family Studies)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 05/28/2021

The dissertation is approved by the following members of the Final Oral Committee:

Heather L. Kirkorian (Chair), Associate Professor, Human Development & Family Studies
Robert Nix, Professor, Human Development & Family Studies
Janean E. Dilworth-Bart, Professor, Human Development & Family Studies
Percival G. Matthews, Associate Professor, Educational Psychology
John Franchak, Assistant Professor, Psychology, UC Riverside

@ Copyright by Mengguo Jing 2021

All Rights Reserved

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to my advisor Dr. Heather Kirkorian, who guided me through the journey of graduate school with unconditional, endless support over the past five years. Your expertise and pure passion for science have shaped me as a researcher. You have always held me up when I made mistakes and told me “this is a learning process”. You listened to my stress and confusion, and constantly spent extra time meeting with me. You were even more excited than myself at every single progress I made.

Thank you to my collaborators on this project, Dr. John Franchak, Dr. Tiffany Pempek, and Kellan Kadooka, for providing invaluable resources and being knowledgeable and generous mentors/colleague. Thank you to my other committee members: Dr. Percival Matthews for your insightful perspectives on spatial representation in young children, Dr. Robert Nix for your continuous guidance and feedback on this project, and Dr. Janean Dilworth-Bart for your mental and academic support along the way. Your time and attention during busy semesters is much appreciated. My gratitude also goes to Dr. Alvin Thomas, for your continuous inspiration and confidence in me as a researcher.

Special thanks to the members of Cognitive Development & Media Lab. Thank you to Seung Heon Yoo for always being there when I needed you, sharing your experience and advice through these years. Thank you to Koen Choi for being a wonderful academic sister, listening to and encouraging me. Thank you to Elizabeth Skora Horgan, Roxanne Etta, and the community of undergraduate students for being supportive and lovely lab mates and helping make the Ph.D. path less daunting and more joyful.

I gratefully acknowledge the children, families, and preschool partners for participating in this research and the funding provided by the Department of Human Development and Family

Studies at the University of Wisconsin-Madison. Without the generosity of these people, this work would not have been possible.

And finally, my heartfelt gratitude to my dear family, especially my dad, Song Jing, and my mom, Juan Zhang. Their immeasurable love and support have filled me with strength and fearlessness through the road of graduate school and my lifetime. I would like to give my sincere thanks to my fantastic fiancé, Qian Yao, for everything you helped during my intensive thesis writing -- from replacing my slow old laptop, to taking care of all the cooking and laundry, to bringing me ease with your guitar during the long hours. Thank you for seeing and bringing out the best in me as a learner, a researcher, and a life partner.

Young Children's Information Processing of Digital Media Content: The Impact of Media and Child Characteristics

ABSTRACT

A critical development in human cognition is growing to understand information presented in one context and use that information to solve problems in another context (DeLoache, 1995; Chen & Siegler, 2013). Digital media, as an increasingly important source of information in children's lives, could provide a relatively naturalistic context for studying young, preliterate children's behaviors in the laboratory environment. This dissertation investigated children's behaviors at two stages of information processing of screen media, including the initial stage of attending to information and the later stage of retrieving information.

Study 1 examined the effect of video comprehensibility and age on viewers' visual attention while watching a *Sesame Street* episode with a complete story arc. Results suggest that 1) adults' eye movements were more likely to be predicted by low-level visual salience on the screen compared to 4-year-olds, and 2) while adults' salience-based eye movements increased while watching a less comprehensible video narrative, there was limited impact on children. Study 2 tested the effects of an interactive feature on toddlers' symbolic transfer across different situations, particularly when using video as a symbolic medium. Interactivity was found to enhance children's overall errorless transfer performance in and only in the video, not the live, situation; however, it did not disproportionately affect the rate of perseverative errors, nor did it affect the latency of children's searches. Results have implications for the particular mechanisms by which interactivity affects toddlers' symbolic transfer.

Together, the findings provide direct evidence for the effectiveness of two important media characteristics (narrative comprehensibility, touchscreen interactivity) on screen-mediated

learning, and add to the literature about how the learning is supported by fundamental cognitive systems, such as selective attention and encoding and symbolic representation. This dissertation 1) demonstrates the active engagement of multiple cognitive processes in children's interaction with screen media, 2) provides empirical evidence that visual attention is a key process for young children's information processing of screen information, and 3) highlights a holistic view in understanding the relationship between children and screen media, particularly regarding cognitive processes engaged in media use.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

One of the most important developments in human cognition is growing to understand information presented in one context and use that information to solve problems in another context (DeLoache, 1995; Chen & Siegler, 2013). On the one hand, as with many sources of information (e.g., gestures, numerals, written text, maps, pictures), digital media are symbolic and require representational competence and cognitive flexibility to efficiently understand and learn from them (Troseth et al., 2003; Barr, 2013). While adults can quickly learn to solve many problems or accomplish many tasks by watching a tutorial video in YouTube (e.g., solve a Rubik's cube, hike to the Hollywood sign, cook a new recipe), it is well established that processing information on the screen is cognitively taxing and could be particularly challenging to young children who have limited cognitive maturation and video viewing experience (Anderson & Pempek, 2005). On the other hand, digital media have many unique features (e.g., video editing techniques, interactive interface) that distinguish them from other types of symbolic media, thereby posing new questions regarding children's learning from them (Troseth et al., 2019; Sheehan & Uttal, 2016). Moreover, given their ubiquity and availability to children at an increasingly early age, digital media could provide a relatively naturalistic context for studying young, preliterate children's behaviors in the laboratory environment (e.g., Shepherd et al., 2010; Wang et al., 2017).

This dissertation aimed at understanding children's information processing of digital media content by examining their behaviors at two stages of information processing of video-based screen media (henceforth, screen media), including the initial stage of attending to information and the later stage of retrieving information. At the theoretical level, the whole

dissertation was underpinned by the information-processing theories and the 3 C's framework. Drawing on the metaphor that the mind is a computer (Colombo & Mitchell, 2009; Sokolov, 1963), information-processing theories assume that information is represented internally by particular mental processes and flows through the information-processing systems of human cognition (Kail & Bisanz, 1992; Siegler & Abibali, 2005). At the core of this dissertation, children's behaviors (i.e., eye movements during video watching, retrieving an object) involved in screen media activities were conceptualized as mental processes in which the information of media content is processed at the cognitive level, from attending to information, encoding information into mental representation, to retrieving information from memory. Meanwhile, a growing body of research emphasizes the critical importance of considering the child, the content, and context – the 3 C's framework -- while understanding children's behaviors involved in using screen media (e.g., Guernsey, 2007; Valkenburg & Peter, 2013). While information-processing theories provided a theoretical account for the moment-to-moment processing of media information at the cognitive level, the 3 C's account provided a framework to contextualize the interaction between children and media. As such, the current dissertation examined the influence of factors from the three contextual levels on children's behaviors involved in screen media activities. Taken together, drawing from information-processing theories and 3 C's framework, this dissertation took a holistic approach to understand the mechanisms through which different cognitive systems and contextual factors interacted to influence children's behaviors involved in using screen media. Also note that this dissertation leveraged existing data sets to address key questions related to information encoding and retrieval from screen media while accommodating limitations on primary data collection during the COVID-19 pandemic.

Attending to Information

As the inlet of information, attending to and encoding specific objects in complex naturalistic contexts is crucial to children's learning. Similarly, selectively attending to and maintaining attention to particular content or information on the screen is the first step in acquiring information from screen media and plays an important role in learning the media content (Kirkorian et al., 2017). Screen media seem to be a transparent source of information that anyone could learn from. However, ample evidence exists that children may attend to and encode information differently when observing live demonstrations versus watching video demonstrations, or even when using different types of media (e.g., watching TV vs. using touchscreen). For example, young children appear to process information more slowly when viewing video demonstrations than when viewing the same demonstrations in person (Carver et al., 2006; Kirkorian et al., 2015). Therefore, a better understanding of children's attention to screen media at the attending and encoding stage would help to elucidate why children typically learn less from symbolic media than from equivalent real-life experiences (Anderson & Pempek, 2005; Troseth, 2010). To this end, this dissertation investigated the influence of two media characteristics on children's attention to and encoding of video information: the impact of narrative structure on children's and adults' visual attention during video watching (Study 1) and the impact of an interactive interface at the attending and encoding stage on toddlers' later memory retrieval (Study 2).

Research has consistently suggested that child do not share the apparent ease of adults in comprehending on-screen information (Barr, Muentener, & Garcia, 2007; Strouse & Troseth, 2008). Meanwhile, accumulated evidence from eye-tracking and overt-looking research shows that visual attention patterns during video viewing differ for children and adults (e.g., Franchak

et al., 2016; Kirkorian et al., 2012), which might help to explain young children's relatively low video comprehension of television (Anderson & Lorch, 1983). While research has demonstrated that young children's attention could be driven by bottom-up salience (e.g., luminance, contrast, color, orientation, and motion), top-down influences (e.g., children's comprehension of the story) are also important forces in attention allocation, particularly in older children. Study 1 explored the role of top-down influences in visual attention and its age-related change by examining preschoolers' and adults' eye movements while watching video. Specifically, this study examined the effect of video comprehensibility and age on viewers' visual attention while watching a *Sesame Street* episode with a complete story arc. The experiment manipulated the comprehensibility of the television show by rearranging the order of the shots, allowing us to test the effects of top-down attentional control while children and adults watch television. In the child sample, daily exposure of foreground television and background television were also examined as socio-contextual factors that could influence child media behaviors. Thus, this study was designed to examine the impact of both content characteristics and child characteristics on attention and encoding during television viewing.

While Study 1 directly examined attention during television viewing, Study 2 was designed to indirectly examine the effects of attention on learning from interactive screen media. Some scholars have proposed that interactive features, such as those afforded by video chat and touchscreen mobile applications, could increase young children's selective attention and encoding of information on screen (e.g., Choi & Kirkorian, 2016; Kuhl, 2007). However, research has been mixed regarding the effectiveness of interactivity (e.g., Alade et al., 2016; Schroeder & Kirkorian, 2016; Kirkorian et al., 2016). Thus, Study 2 was designed to understand the information attending/encoding mechanism, among others, underlying the

interactivity effect on children's video-mediated learning. In this way, Study 2 considered attention/encoding as one potential mechanism underlying an impact of interactivity on toddlers' learning from digital media.

Representing and Retrieving Information

Once attending to target information, how children mentally represent it directly affects learning of that information and transferring it to another context. To transfer the information encoded from screen media to corresponding objects or events in real life, children must remember the information in the video and understand how it maps onto a real-life situation. In other words, children must demonstrate *symbolic competence* to first achieve the insight that on-screen information represents objects or events in real life and then to make the mapping between the screen symbol and its referent in real life. Prior work has demonstrated that this is not an easy task for young children. During the first few years of life, children are unable to recognize the symbolic relation between the symbol and its referent. However, despite the symbolic insight at the conceptual level, symbolic transfer from video to real life could be particularly challenging. The reason could be, even when children gain some symbolic insight, their mental representation of video is quite fragile and easily disrupted (Troseth, 2003, 2010), which could be due to degraded perceptual and social experience (Schmitt & Anderson, 2002; Barr, 2010) and lower arousal and engagement as compared to in-person experience (Kuhl, 2003, 2007). Therefore, an in-depth examination of different accounts is necessary to understand potential encoding mechanisms underlying children's learning from screen media.

There are several ways in which media interactivity might influence children's symbolic competence. For example, interacting with the screen may foster children symbolic insight and increase arousal, engagement, and attention (Kuhl, 2003, 2007; Troseth et al., 2019; Kirkorian et

al., 2016), making it easier for children to retrieve the representation of symbolic media in a real-life situation. On the contrary, interacting with the screen may distract children from the representational function of symbolic media and even add extra cognitive demands (Alade et al., 2016; Schroeder & Kirkorian, 2016). Thus, Study 2 was designed to test the effects of an interactive feature on children's memory retrieval across different situations, particularly when using video as a symbolic medium. Through an investigation on how interactivity influences information retrieval across different situations (e.g., live vs. video, symbolic vs. non-symbolic), we could speculate about the underlying mechanism underlying the information processing of screen media at different information-processing stages. In addition, child characteristics at the socio-contextual level, including age and naturalistic media use, were also included to examine their relations with learning from video.

Summary

Underpinned by the information-processing theories and 3 C's framework, this dissertation examined the role of factors, at content, child, and context level, in children's learning from screen media and explored underlying mechanisms at different information-processing stages. Study 1 focused on the relation between video comprehensibility and visual attention, while Study 2 focused on the relation between video interactivity and memory retrieval. Both studies took into child characteristics at the developmental and socio-contextual level. The two empirical studies collectively contributed to understanding young children's processing of information in the digital context.

The findings would add to the literature by shedding light on how screen-mediated learning is supported by fundamental cognitive systems, such as selective attention and encoding and symbolic representation. Practically, this dissertation could help to identify specific child

(e.g., media experience) and content features (e.g., narrative structure, media interactivity) that are likely to support early learning from screen media, as an effort to create scalable and cost-effective media that facilitate early learning and development.

References

Aladé, F., Lauricella, A. R., Beaudoin-Ryan, L., & Wartella, E. (2016). Measuring with Murray: Touchscreen technology and preschoolers' STEM learning. *Computers in Human Behavior*, *62*, 433-441.

Anderson, D. R., Lorch, E. P., Field, D. E., Collins, P. A., & Nathan, J. G. (1986). Television viewing at home: Age trends in visual attention and time with TV. *Child development*, 1024-1033.

Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist*, *48*(5), 505-522.

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review*, *30*(2), 128-154.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, *4*, 328.

Barr, R., Muentener, P., Garcia, A., Fujimoto, M., & Chávez, V. (2007). The effect of repetition on imitation from television during infancy. *Developmental Psychobiology*, *49*(2), 196-207.

Carver, L. J., Meltzoff, A. N., & Dawson, G. (2006). Event-related potential (ERP) indices of infants' recognition of familiar and unfamiliar objects in two and three dimensions. *Developmental Science, 9*, 51–62.

Chen, Z., & Siegler, R. S. (2013). Young children's analogical problem solving: Gaining insights from video displays. *Journal of experimental child psychology*, *116*(4), 904-913.

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of learning and memory*, *92*(2), 225-234.

Choi, K., & Kirkorian, H. L. (2016). Touch or watch to learn? Toddlers' object retrieval using contingent and noncontingent video. *Psychological science*, *27*(5), 726-736.

DeLoache, J. S. (1995). Early understanding and use of symbols: The model model. *Current Directions in Psychological Science*, *4*(4), 109-113.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. Cognition, 110, 160–170.

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy*, *21*(3), 262-287.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements* (pp. 537-III). Elsevier.

Guernsey, L. (2007). *Into the minds of babes: How screen time affects children from birth to age five*. Basic Books.

Kail, R., & Bisanz, J. (1992). The information-processing perspective on cognitive development in childhood and adolescence. Intellectual development, 1, 229- 260. Kirkorian, H. L. (2018). When and how do interactive digital media help children connect what they see on and off the screen?. *Child Development Perspectives*, *12*(3), 210-214.

Kirkorian, H. L., Anderson, D. R., & Keen, R. (2012). Age differences in online processing of video: An eye movement study. *Child development*, *83*(2), 497-507.

Kirkorian, H. L., & Anderson, D. R. (2018). Effect of sequential video shot comprehensibility on attentional synchrony: A comparison of children and adults. *Proceedings of the National Academy of Sciences, 115*, 9867–9874.

Kirkorian, H. L., Choi, K., & Pempek, T. A. (2016). Toddlers' word learning from contingent and noncontingent video on touch screens. *Child development*, *87*(2), 405-413.

Kirkorian, H. L., Lavigne, H. J., Hanson, K. G., Troseth, G. L., Demers, L. B., & Anderson, D. R. (2016). Video deficit in toddlers' object retrieval: What eye movements reveal about online cognition. *Infancy*, *21*(1), 37-64.

Kirkorian, H., Pempek, T., & Choi, K. (2017). The role of online processing in young children's learning from interactive and noninteractive digital media. In *Media exposure during infancy and early childhood* (pp. 65-89). Springer, Cham.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, 10, 110–120. doi:10.1111/j.1467- 7687.2007.00572.x

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, USA, 100, 9096–9101. doi:10.1073/pnas.1532872100.

Schroeder, E. L., & Kirkorian, H. L. (2016). When seeing is better than doing: Preschoolers' transfer of STEM skills using touchscreen games. Frontiers in Psychology, 7, 1377. https://doi.org/10.3389/fp syg.2016.01377.

Schmitt, K. L., & Anderson, D. R. (2002). Television and reality: Toddlers' use of visual information from video to guide behavior. *Media Psychology*, *4*(1), 51-76.

Schroeder, E. L., & Kirkorian, H. L. (2016). When seeing is better than doing: Preschoolers' transfer of STEM skills using touchscreen games. *Frontiers in psychology*, *7*, 1377.

Sheehan, K. J., & Uttal, D. H. (2016). Children's learning from touch screens: a dual representation perspective. *Frontiers in psychology*, *7*, 1220.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human–monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology, 20*, 649–656.

Siegler, R. S., & Alibali, M. W. (2005). Information-processing theories of development. *Children's thinking*, 65-106.

Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual review of physiology*, *25*(1), 545-580.

Strouse, G. A., & Troseth, G. L. (2008). "Don't try this at home": Toddlers' imitation of new skills from people on video. *Journal of Experimental Child Psychology*, *101*(4), 262-280.

Troseth, G. L. (2003). TV guide: Two-year-old children learn to use video as a source of information. *Developmental psychology*, *39*(1), 140.

Troseth, G. L., Flores, I., & Stuckelman, Z. D. (2019). When representation becomes reality: Interactive digital media and symbolic development. In *Advances in child development and behavior* (Vol. 56, pp. 65-108). JAI.

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development, 77*, 786–799. https://doi.org/10.1111/j.1467- 8624.2006.00903.x

Valkenburg, P. M., & Peter, J. (2013). The differential susceptibility to media effects model. *Journal of communication*, *63*(2), 221-243.

Wang, H. X., Freeman, J., Merriam, E. P., Hasson, U., & Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of vision*, *12* (1), 16-16.

CHAPTER 2

Abstract

Low-level visual features (e.g., movement, edges) drive eye gaze during video viewing, especially for older viewers. The current study investigated the effect of video comprehensibility on the extent to which eye movements are predicted by visually-salient features. Eye movements were recorded as 4-year-olds (n = 20) and adults (n = 20) watched a cohesive versus random video sequence of a 4.5-min full vignette from *Sesame Street*. Overall, salience-based gaze prediction was higher in adults than in children, especially when viewing a random video sequence. The impact of random-edit video on children's gaze was limited to the narrow window of time surrounding cuts to new video shots. The finding that adults had higher (not lower) salience-based gaze prediction when watching the random video sequences suggests that age-related increases in salience-based gaze prediction is not due to age-related increases in video comprehension. Implications for top-down versus bottom-up control of eye movements as well as children's developing attention during video viewing have been discussed.

Keywords: Eye movements, visual attention, salience, video viewing

The Effect of Comprehensibility on Salience-Based Gaze Prediction for Children and Adults Watching Video

Selective attention engages both bottom-up processes (e.g., stimulus-driven perceptual factors; Itti, 2000; Itti & Koch, 2001) and top-down processes (e.g., internally driven cognitive factors; Birmingham et al., 2008; Castelhano et al., 2007). Similar to real-world scenarios, television contains meaningful, audiovisual content that changes instantly over time. Television viewing, thus, provides an ideal context to study naturalistic viewing of complex, dynamic scenes. As with real-world viewing, a viewer's attention, during television viewing, may be driven by low-level perceptual features (e.g., perceptual changes between scenes, on-screen movements, visual effects) as well as their ongoing effort to comprehend the video content (e.g., integrating information from one scene to the next). Critically, bottom-up and top-down processes are not independent, insofar as perceptually-salient features may signal important content or be clustered around meaningful regions of the scene (e.g., a character's face) (Huston & Wright, 1983; Henderson et al., 2007; Wass & Smith, 2014).

The impact of perceptually-salient features on eye movements during video viewing has been well documented throughout lifespan. For example, low-level stimuli salience features, such as luminance, contrast, color, orientation, and motion, have been found to influence eye movements in adults ('t Hart et al., 2009; Itti, 2005; Mital et al., 2011; Frank et al., 2009), school-age children (Kadooka & Franchak, 2019; Rider et al., 2018), and infants (e.g., Kadooka & Franchak, 2019; Franchak, et al., 2016; Frank et al., 2009). Yet, research is limited during early childhood, a period of substantial change in children's comprehension of television content (Anderson & Hanson, 2010). Moreover, no study has directly examined the extent to which top-down, comprehension-driven processes contribute to the impact of salience on gaze. The purpose

of the current study was to determine the extent to which comprehensibility moderates the impact of visually-salient features on gaze allocation during free viewing of a complete television narrative in young children and adults.

Bottom-up and Top-down Influences on Gaze During Video Viewing

Several studies have examined the extent to which visually salient features predict eye gaze, which is referred to as *salience-based gaze prediction* in the current study, while adults view dynamic scenes (Itti, 2000; Itti & Koch, 2001). Saliency-based gaze prediction, in other words, denotes how well the visually salient features (e.g., color, contrast, luminance, and motion) on the screen predict where viewers look at on the screen. A relatively small number of studies have extended this work to younger viewers. For instance, Frank et al. (2009) observed infants and adults watching 4-second clips from an animated movie. Frank et al. defined a predictive model using visual features such as temporal luminance contrast and spatial luminance contrast. They found that the extent to which infants (3-9 months) fixated visually salient features on the screen was greater than would be expected by chance alone. In a similar study using a continuous 60-second clip from a children's television program, Franchak et al. (2016) calculated the salience of viewers' fixation areas based on five image channels, including color, contrast, orientation, flicker (i.e., luminance differentials across frames), and motion. The average salience of the fixated areas was found to be in the top 20% of the salience of the whole screen in both an infant (6-24 months) and adult group. Together, these studies and others (e.g., Kadooka & Franchak, 2019; Mital et al., 2011; Rider et al., 2018; Shepherd et al., 2010) demonstrate that visual salience predicts where eye gaze is directed in infants, school-age children, and adults alike. Yet, many studies use brief video clips that lack narrative context, potentially limiting the impact of top-down, comprehension-driven processes on eye gaze.

Research suggests that viewers may employ different processes to guide attention during structured tasks that contains a goal than during unstructured free-viewing situations. For instance, Smith and Mital (2013) compared adult viewers who freely viewed a naturalistic video to those who were instructed to identify the location depicted in each video scene. With a meaningful goal, viewers in the spot-the-location condition were less likely to direct gaze to perceptually-salient features, such as areas of high flicker and moving objects such as people. That is, when given an explicit task goal, the processes driving visual attention appear to have changed: Adult viewers relied less heavily on bottom-up gaze control and, by extension, more heavily on top-down gaze control.

When watching television, viewers rarely have an explicit task goal. Yet even when viewing short videos with relatively little narrative structure, viewers' gaze is at least partly driven by semantically-relevant stimuli, such as faces (e.g., Franchak et al., 2016; Frank et al., 2009; Rider et al., 2018). Further, in the case of narrative television, it is likely that comprehension-driven, top-down processes are involved as viewers seek to understand the story. Indeed, the comprehensibility of video content has been shown to influence the duration of looks toward the screen in viewers as young as 18 months of age (Anderson et al., 1981; Pempek et al., 2010). Moreover, comprehensibility affects attentional synchrony, or the consistency in gaze location across observers, at least as early as 4 years of age (Wang et al., 2012; Kirkorian & Anderson, 2018). Importantly, these prior studies experimentally manipulated the comprehensibility of video content while holding low-level visual and auditory features constant. Therefore, the prior work provides causal evidence of top-down influences on attention.

Importantly, bottom-up and top-down processes do not work in isolation. For instance, perceptually-salient features (e.g., motion in dynamic scenes; Mital et al., 2011; Smith & Mital,

2013) and semantic information (e.g., faces; Franchak et al., 2016; Shepherd et al., 2010; Smith & Mital, 2013) are both related to common gaze patterns, such as attentional synchrony. Moreover, perceptually-salient information may be a useful indicator of meaningful, comprehension-relevant information. For example, in a study where adult participants were either instructed to free view or to determine which player earned a point while watching a tennis match, no difference in eye-movement patterns between the two viewing situations is likely due to the nature of the tennis match where the most visually-salient regions (i.e., movement) overlapped with regions that provided task-relevant information (i.e., players earning points) (Taya et al., 2012). Particularly, in the case of children's television, meaningful information may be especially likely to overlap with perceptually-salient information, such as feature congestion and flicker (Wass & Smith, 2014). Salience may also serve as a cue to draw attention towards meaningful areas in a scene. Infants, children, and adults were all more likely to attend to faces when faces were more salient compared to when faces were less salient in a series of television clips (Franchak & Kadooka, under review).

Age-Related Changes in Attention During Video Viewing

Evidence from several studies implies age-related change in the extent to which bottom-up and top-down processes predict eye gaze. The impact of salient features and semantics features on eye movements have been typically found to increase with age, even though emerging evidence suggests this is not always true for all videos (Kadooka & Franchak, 2020). For example, in a study of infants (6-24 months) and adults, the overall salience of fixated areas increased with age, indicating an age-related increase in salience-based gaze prediction (Franchak et al., 2016). Similar results have been reported by others (Frank et al., 2009; Rider et al., 2018), even though emerging evidence suggests this is not always true for all videos

(Kadooka & Franchak, 2020). Moreover, the impact of semantic features (i.e., faces) on fixations was also found to increase with age throughout early infancy (3-9 months) and was even higher in adults, suggesting an age-related increase in the influence of top-down processes (Frank et al., 2009). However, another investigation found no age-related increases in face looking among infants and children (6 months to 12 years) who watched 7 different video clips (Kadooka & Franchak, 2020), suggesting that gaze predictability, whether using a salience-based or object-based predictive model, depends on the nature of the video content.

When salience-based gaze prediction increases with age, there are at least three hypotheses for this age-related increase. First, older viewers could have increased sensitivity to bottom-up features, and their eye movements might be more driven by visually salient features as compared to younger viewers. We will refer to this as the *salience-based hypothesis*. While this hypothesis is supported by accumulated evidence that older viewers have increasingly more eye movements to visual salience (e.g., Franchak et al., 2016; Frank et al., 2009; Rider et al., 2018), conflicting findings from developmental research indicate the opposite across early life. For example, Gola and Calvert (2011) found that, as compared to 12-month-olds, children aged 6 and 9 months attended more to the video with faster pacing, and suggested a reducing influence of perceptual salience in visual attention with age.

Another possibility is that viewers' overall gaze predictivity, at both salience-based level and comprehension-based level, increases with age due to a range of factors, resulting in age-related increase in salience-based gaze prediction. We will refer to this as the *overall predictability hypothesis*. For instance, empirical evidence from eye-tracking research regarding selective fixation within a scene shows growing attentional synchrony in older viewers (Frank et al., 2009; Franchak et al., 2016; Kirkorian & Anderson, 2018). Older viewers' eye gaze is more

spatially and temporally consistent with each other within the same age group. In contrast, young viewers' eye movements tend to be more scattered and idiosyncratic, which could lead to noisy data and poor predictivity of eye gaze.

Third, it is also possible that the increase of salience-based gaze prediction with age is a coincidental consequence of an age-related increase in comprehension-related gaze prediction. We will refer to this as *comprehension-based hypothesis*. On one hand, developmental research has consistently showed that children's visual attention to television increases with age, suggesting that older children with better comprehension pay greater attention to video (Anderson et al., 1981; Richards & Cronise, 2000). As more direct evidence, beginning as early as 18 months of age, children pay greater attention to comprehensible child-directed television than to random audio-visual displays (Richards & Cronise, 2000), adult-directed television (Schmidt et al., 2008; Schmitt et al., 1999), or the same child-directed television shots presented in a random sequence (Anderson et al., 1981; Pempek et al., 2010). On the other hand, since perceptually-salient regions may often overlap with meaningful ones (Wass & Smith, 2014), an increased salience-based gaze prediction may be observed due to an increased meaning-based gaze prediction. Indeed, young viewers may learn that certain features are associated with comprehensible child-directed content, while other features tend to signal less comprehensible adult-directed content (Huston & Wright, 1983). Research suggests that preschool-age children do respond differently to different types of formal features, paying greater attention to television containing features associated with child-directed programs (e.g., animation, puppets, child actors, fast-paced music) and less attention during features associated with adult-directed programs (e.g., live-action, adult male actors, slow-paced music) (Alwitt et al., 1980; Valkenburg & Vroone, 2004). While most formal-feature research considers overt looks at the

screen, a similar pattern may be observed for changes in gaze location within the same scene in response to visually salient features (e.g., movement, contrast).

Previous findings, based on correlation research, fails to provide direct evidence to an age-related increase in salience-based gaze prediction. To tease apart the multiple hypothesis, an experiment that manipulates viewing content's comprehension-related semantic feature (i.e., narrative comprehensibility) within same age groups would disentangle the confounding between visually-salient features and comprehension-related features. Moreover, the majority of research to date has focused on adults and infants 24 months old and younger. Given the rapid development of young children's video comprehension and film literacy during early childhood (Anderson & Hanson, 2010; Schmitt et al., 1995), research with young children is needed to better understand how attention to and comprehension of video emerge during this period.

Impact of Scene Changes in Edited Video Sequences on Attention

Television viewing is more complex than viewing unedited videos of naturalistic scenes or brief film excerpts due to *filmic montage*, or video editing techniques that convey concepts through relations between shots. For example, television programs typically convey complex narratives through transitions across time, space, and character perspective. Comprehending filmic montage requires experience (Ildirar & Schwan, 2015) and well-developed cognitive abilities (Smith et al., 2012). As such, while television comprehension may feel like an automatic, mindless process to experienced adult viewers, visually processing and comprehending video is more challenging for young children (e.g., Anderson et al., 2006; Smith, Anderson, & Fischer, 1985; Lorch et al., 1987). Among the different montage techniques, of particular interest in the current study are the transitions from one video shot to another, often called *jump cuts*, such as from one camera angle to the next or one scene to the next.

Jump cuts may be a unique type of formal feature that guides attention in specific ways. For instance, jump cuts have been shown to elicit overt looks from inattentive viewers (Alwitt et al., 1980), likely due to the abrupt change in visual and auditory cues. Jump cuts also have unique effects on the visual fixation of attentive viewers. Adult viewers tend to fixate the center of the screen immediately after a cut to a new scene (Tseng et al., 2009; LeMeur et al., 2007; Mital et al., 2011; Tosi et al., 1997; Wang et al., 2012). As a result, attentional synchrony in adults is higher immediately following cuts to new scenes than later on in those scenes (Kirkorian & Anderson, 2018; Kirkorian et al., 2012). Fixating the center of the screen may be strategic, allowing viewers to orient quickly to new scene content. Indeed, adult viewers are more likely to fixate the center of the screen following a cut to a brand-new, unfamiliar scene than following a cut to a different camera angle within the same, familiar scene (Kirkorian et al., 2012).

While cuts clearly have an effect on eye movement patterns in adults, the specific impact of cuts on salience-based gaze prediction remains unclear. A comparison across different studies suggests that the impact of cuts may depend on the semantic relation between consecutive video shots. When adult viewers watched a series of short (4.5 to 30 sec), unrelated video shots, salience-based gaze prediction peaked within 250 ms of jump cuts and gradually decreased over the subsequent 2500 ms (Carmi & Itti, 2007). The authors posited that viewers first oriented to perceptually-salient features in the new shot before identifying and attending to semantically meaningful objects. In contrast, in a recent study using longer (3-minute) excerpts from movies, Rider et al. (2018) found an immediate decline in salience-based gaze prediction after jump cuts, followed by a recovery at around 500 ms. Together, this research suggests that the impact of cuts

on salience-based gaze prediction may differ for short, unrelated video clips versus a coherent sequence of shots representing a continuous action or story.

One reason attention patterns may differ for coherent shot sequences versus disconnected shots is adult viewers' tendency to anticipate the reappearance of an object based on its trajectory before the cut (Kirkorian & Anderson, 2017). If a viewer comprehends a coherent action sequence, they have the opportunity to make anticipatory eye movements following a cut rather than reactive eye movements toward salient regions. Yet young viewers' comprehension of filmic montage emerges gradually throughout early and middle childhood (Kirkorian & Anderson, 2017; Calvert & Scott, 1988; Pempek et al., 2010; Smith et al., 1985; Smith & Henderson, 2008; Smith et al., 2012). Children as young as 18 months show emergent comprehension of cuts and shot sequences (Pempek et al., 2010); however, young children often fail to make inferences about scene continuity and discontinuity across a sequence of visually distinct shots (e.g., across space, time, action, character intention, character psychology) (see Anderson & Hanson, 2010). For instance, 4-year-olds were found to have poor comprehension of two parallel, simultaneous actions conveyed through a sequence of shots alternating between the two scenes (Smith et al., 1985). Moreover, some children as old as 10-12 years showed poor comprehension of more complicated transitions, such as flashbacks (Calvert & Scott, 1988). Together, the research suggests that children begin to comprehend edited sequences of video shots during the second year of life, but this skill continues to improve through middle childhood. Given such protracted development of video comprehension, the impact of comprehension on eye movement patterns like salience-based gaze prediction may be markedly different in young children than adults.

Overview of the Current Study

The present study is based on a secondary analysis of a dataset described elsewhere [citation redacted for blind review] to address new research questions about salience-based gaze prediction (SBGP). Our aim was to examine the extent to which video comprehensibility affects salience-based gaze prediction in adulthood as well as in early childhood, a period of rapid development in cognitive skills generally as well as video comprehension in particular. Prior research examining top-down influences on eye gaze during children's video viewing has been largely correlational, using isolated video clips that lack a narrative context. To more directly test the impact of top-down, comprehension-related processes on eye movements, we compared viewers watching a complete, comprehensible video sequence (i.e., normal video) to those watching a less comprehensible, random sequence of the same video shots (i.e., random-edit video). This manipulation has been used in several prior studies to examine young children's comprehension of television, revealing that sensitivity to random shot sequences emerges during the second year of life (Pempek et al., 2010) and increases through the preschool years (Anderson et al., 1985; Hawkins et al., 1995).

Our first analysis focused on overall effects of comprehensibility (normal vs. random sequence) and age group (4-year-olds vs. adults) across the whole video. Based on prior research with infants and with older children (Franchak et al., 2016; Frank et al., 2009; Rider et al., 2018), we expected gaze to be more predictable in adults than in young children, as evidenced by higher salience-based gaze prediction. The impact of comprehensibility on salience-based gaze prediction in adults was an open research question. On the one hand, random-edit video could decrease opportunities for top-down attentional control, causing adults to rely more heavily on bottom-up attentional control as evidenced by greater salience-based gaze prediction in adults. On the other hand, reduced comprehension could result in less systematic and predictable eye

gaze in adults. Unsynchronized gaze patterns would add noise to eye movements at both the spatial and temporal level, yielding lower salience-based gaze prediction. Given that young children have relatively limited comprehension of edited video sequences, we expected that any effect of comprehensibility on salience-based gaze prediction (whether positive or negative) would be smaller in children than adults.

Our second analysis focused on salience-based gaze prediction immediately following jump cuts and other transitions to new shots, given the importance of these transitions for guiding fixations on the screen (e.g., Mital et al., 2010; Kirkorian et al., 2012; Rider et al., 2018) and in young children's comprehension of video (Anderson & Hanson, 2010). Based on prior research with cohesive video sequences (Rider et al., 2018), we expected an initial drop in salience-based gaze prediction immediately after cuts to new scenes, accompanied by an increased likelihood of fixating the center of the screen (Kirkorian et al., 2012; LeMeur et al., 2007; Tseng et al., 2009). To the extent that these timebound effects reflect viewers' ongoing comprehension of a video sequence (e.g., anticipating a new scene), we expected smaller timebound effects, as indicated by smaller decrease in salience-based gaze prediction, for children (versus adults) and for the random (versus normal) shot sequence.

Method

Participants

The current study constitutes secondary data analysis. The original sample included 33 4-year-old children and 44 adults with normal or corrected-to-normal vision. Participants were assigned at random to one of two comprehensibility groups: normal or random video sequence. From the original sample, 3 children and 12 adults were dropped due to inability to calibrate the eye tracker, for instance due to head movements disrupting the head tracker or reflective glasses

distorting the corneal reflection. From the remaining 62 participants, this secondary analysis included data from 10 participants per cell with the highest data quality, as described later (see Data Pre-processing and Inclusion Criteria). Thus, the final sample used in this secondary analysis included 20 children (7 females; M = 4.51 years, SD = 0.10, range 4.36 to 4.74 years) and 20 adults (15 females; M = 20.46 years, SD = 1.14, range 18.47 to 22.21 years), divided equally into the two comprehensibility groups.

The original study was approved by the Institutional Review Board at [university name redacted for blind review]. Data were collected in 2008. Child participants were recruited through letters and phone calls based on a local database of birth records. The majority (90%) of the 4-year-olds were described as White/Caucasian. As a proxy for socio-economic status, parents reported the number of years of education they completed, with 12 years typically indicating a high-school diploma, 16 years typically indicating a 4-year college degree, and so on. The average number of years of education per parent was 17.73 (SD = 3.88, range 12 to 25 years). Adult participants were recruited from undergraduate psychology courses.

Stimuli

The current analysis is based on a complete vignette from the children's program *Sesame Street*. In the original study, participants viewed the 20-second opening scene for the show, followed by the 4.5-minute *Journey to Ernie* vignette used for the current analysis. The video presents a full story arc from a recurring vignette in which the character Ernie hides and other characters search for him. In the specific vignette used in the current study, Ernie says he will hide behind something that grows. Another character, Big Bird, searches for Ernie behind several plants (e.g., flowers, pumpkin, acorn) before ultimately finding Ernie behind a leaf at the top of a beanstalk. Most of the video consists of live-action puppets superimposed on a computer-

animated environment. The vignette contains 28 distinct shots with an average length of 6 seconds (range 3.08 to 37.6 seconds). Most transitions between shots were abrupt jump cuts (e.g., shifting from one camera angle to another or one scene to another). As such, we use the term *cut* to refer to any transition between distinct shots.

The only difference between the two experimental conditions was the order of shots in the sequence: In the normal condition, the shots were played in their original order, presenting a cohesive story. In the random-edit condition, the 28 shots (both video and audio) were reordered in a random sequence (see Figure 1). To create the random sequence, the shots were separated either at the exact moment of an abrupt jump cut or at the midpoint of a wipe across the screen. Thus, to the extent possible, the visual and auditory characteristics within each shot remained, but the sequence of events across the vignette was disrupted, rendering the narrative less comprehensible. This manipulation has been used in several prior studies to test the impact of comprehensibility on young children's attention to television (e.g., Anderson et al., 1981; Hawkins et al., 1995; Pempek et al., 2010).

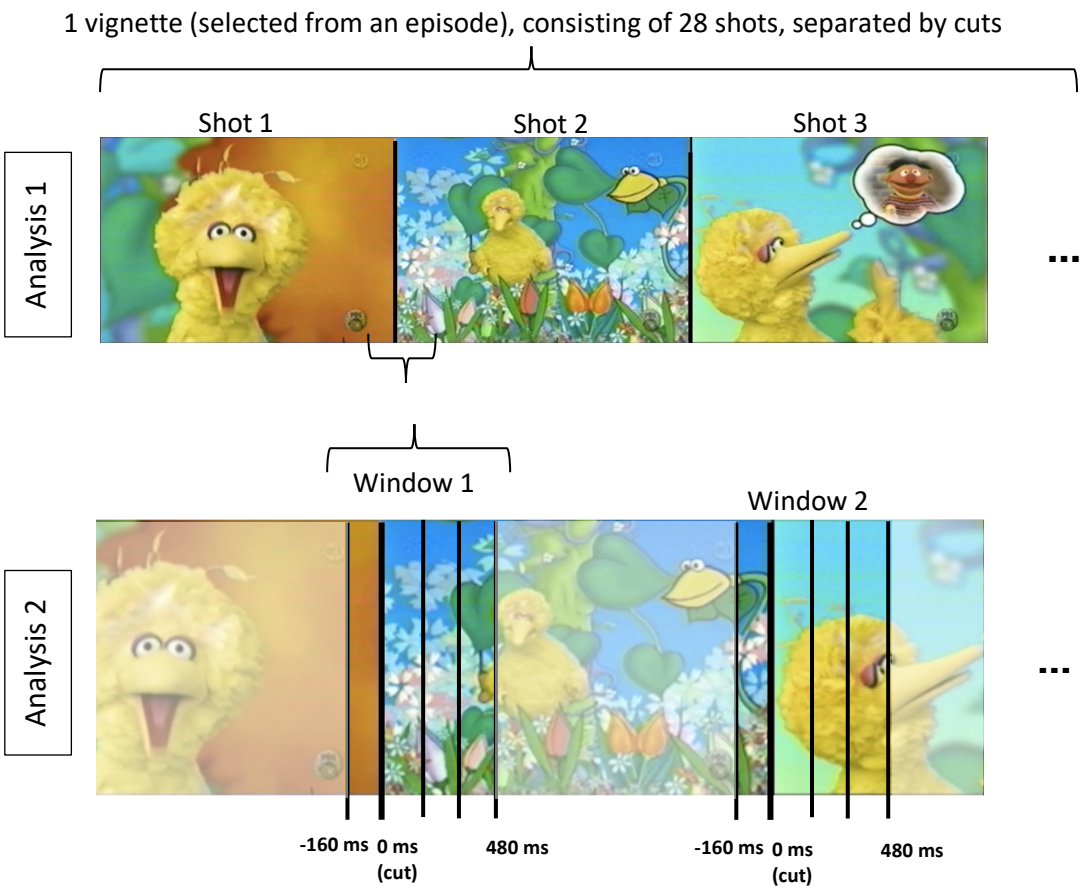

Figure 1. Graphical representation of the data structure in data pre-processing. Analysis 1 examined SBGP for each shot ($N = 28$) nested within each participant ($N = 40$). Analysis 2 examined SBGP and DTC for each 160-ms bin ($N = 4$) nested within each window ($N = 28$), pooling across all participants within each of the four groups.

The random-edit manipulation did not reduce the overall perceptual similarity of adjacent shots, given that all shots occurred within the same general scene (i.e., computer-animated environment with live-action puppets). As evidence, the average visual activity index (Cutting et al., 2011), which describes the similarity in luminance in two separate images (i.e., video frame immediately before versus after each cut) did not differ for the normal versus random video sequence. See [citation redacted for blind review] for a full description of this

analysis. The visual activity index analysis suggests that, despite potential differences in other visual metrics than luminance, any condition effects observed in the current study are not likely to be due to systematic differences in the visual similarity of adjacent shots in the normal versus random sequence. This lends further support to the interpretation that any observed condition effects are due to differences in viewers' comprehension of the narrative.

Setting and Apparatus

The study took place on a university campus in an eye-tracking laboratory room in which dark curtains hung along the walls on all sides of the viewing area. The eye-tracking cameras sat on a table approximately 65 cm in front of the participants. The eye camera was the Applied Science Laboratories (ASL) Eye-Trac 6000, a near-infrared corneal reflection system with remote pan-tilt optics. Temporal resolution was 60 Hz. The ASL VH2 head-tracking camera used face-recognition software to locate and track the viewer's head. An ASL Digital Frame Overlay was also used to insert a digital frame number from the ASL Control Unit onto the video recordings of the sessions, allowing the experimenter to sync gaze data and video stimuli.

General Procedure

Upon entering the study room, the participant was seated in front of the video display screen. Children sat on a booster seat to approximate the height and viewing angle of adults. Parents sat in a chair to the right of the child participants. Parents remained in the room during the session but were asked to refrain from directing their child's attention to any particular area on the screen once the stimulus video began.

A two-point calibration procedure was used for all participants. Small animated characters appeared on the screen for 4 seconds each, alternating between the top-left and bottom-right corners of the screen. Adult participants were asked to look at each character;

children were asked to "play a guessing game" by identifying the characters (e.g., mouse, robot) as they appeared. Given that the same calibration procedure was used for all participants, any condition effects observed in the current study are not likely to be explained by systematic differences in the quality of calibration.

After calibration, the experimenter started the stimulus video and began recording the gaze file and the digital video. Throughout the session, the experimenter ensured that the eye-tracker remained focused on the participant's right eye.

Parent Survey

Parents of child participants completed a questionnaire on demographic information (e.g., parent's education, child's race and ethnicity). To gain a general sense of children's household television exposure, the parents also completed a retrospective viewing diary, recording their child's television exposure for each day (Monday through Sunday) in a typical week. Parents were asked to report on typical television exposure in the foreground (watching child-directed television) and background (being in the room with the television on but not watching a child-directed program). Television exposure data for adults were not collected, because we expected them to be experienced television viewers with relatively little variability in the adults' ability to comprehend the *Sesame Street* vignette.

Data Pre-Processing and Inclusion Criteria

The raw eye-tracking data contained horizontal and vertical gaze coordinates originally recorded at 60 Hz. To smooth the raw data and reduce noise, gaze coordinates were down-sampled to match the frame rate of the video (25 Hz) by taking the average gaze coordinate within each frame.

Participants with low quality data were excluded in an effort to minimize the impact of systematic data loss. Data exclusion occurred in two steps. In the first step, we excluded individual video shots within each participant if that participant was missing data (due to looking away for the screen, for example) for more than 50% of the shot. The first step removed 48% of shots across all children and 25% of shots across all adults. The key variables of interest (described later) are sensitive to sample size, so it was necessary to include the same number of participants in each of the four groups. Thus, in the second step, we included 10 participants in each cell with the least missing data, resulting in a final sample of 20 children and 20 adults (10 per condition in each age group). The mean proportion of shots with less than 50% missing data was 74% (range 50%–89%) for the child-normal group, 95% (range 89%–100%) for the adult-normal group, 69% (range 50%–93%) for the child-random group, and 91% (range 86%–100%) for the adult-random group. Notably, a robustness check showed that the general pattern of results was the same when selecting a random sample of 10 adults per condition rather than selecting the 10 adults per condition with the fewest excluded shots.

Our primary analysis (Analysis 1) examined overall age and condition effects on salience-based gaze prediction across the entire vignette. We calculated SBGP based on the gaze locations within each shot for individual participants; that is, the unit of analysis was each video shot, with each participant receiving one SBGP measurement per shot (see Figure 1). Thus, the dependent variable was calculated at the individual participant level.

The second analysis (Analysis 2) was conducted to address age and condition effects on SBGP and distance to center (DTC) during the specific window of time surrounding cuts to new shots. Based on prior research on SBGP after cuts in older children and adults (Rider et al., 2018), we focused on windows of time that began 160 milliseconds (4 video frames) before each

cut and ended 480 milliseconds (12 video frames) after each cut. To capture the temporal change within each window, we calculated SBGP and DTC within each of four 160-millisecond bins (see Figure 1), resulting in 4 data points per window. Note that, to ensure a sufficient amount of data samples for the SBGP calculation, we pooled the gaze locations of all participants within each group. As such, the dependent variables in Analysis 2 were at the group level rather than at the individual level. The scripts for data processing were written in Matlab (Mathworks, Natick, MA).

Data Reduction and Statistical Analysis

We first computed the two dependent variables, salience-based gaze prediction (SBGP) and distance to center (DTC). Then, descriptive analyses (i.e., t-tests, chi-square tests, bivariate correlations) were run to identify potential covariates at the individual level in child participants. Last, we fitted linear mixed-effects models to examine the age and condition effects on the dependent variables across the entire vignette (Analysis 1) and immediately following cuts (Analysis 2).

Salience-Based Gaze Prediction (SBGP). To estimate SBGP, we first conducted salience analyses. That is, for each gaze coordinate, we computed the overall salience of pixels within the screen region where eye gaze was directed (i.e., *gaze salience*). Next, we conducted receiver operating characteristic (ROC) analyses to obtain area under the curve (AUC) as the metric of SBGP.

Salience Analyses. Image frames of the video vignette for each condition were extracted from the video as JPEG files at the rate of presentation (25 Hz). For each frame image, a salience map was generated to calculate the relative salience of each pixel using the Itti & Baldi (2005) salience algorithm as implemented in GBVS MATLAB toolbox (Harel, Koch, & Perona, 2006).

Relative salience of each pixel was determined using a combination of five feature maps that capture low-level image characteristics: color, intensity, orientation, flicker, and motion (flicker and motion were calculated by comparing each frame to the previous frame). Feature maps were weighted equally to create an overall salience map for each frame. Each pixel within a frame was ranked with a value between 0 and 1 which represented the salience relative to all pixels in this frame with the most salient pixel ranked 1. For every frame, an individual participant's gaze salience was calculated as the average salience value within a 24 pixel radius of the point of gaze. Gaze salience values for each participant were calculated for all frames with a valid gaze coordinate.

ROC Analyses. SBGP was estimated using the Receiver Operating Characteristic (ROC) curve as proposed for eye-tracking analysis by Tatler et al. (2005). The aim was to test the extent to which salience can discriminate gazed regions (i.e., where eye gaze lands) and ungazed regions (i.e., where eye gaze does not land). In the current study, on each video frame, there was a gazed region and a corresponding ungazed region that was randomly sampled from the participant's gaze locations across all frames of the entire video vignette. For a certain salience threshold, a region was classified as "gazed" if its salience was larger than the threshold and as "un-gazed" if its salience was below the threshold. By comparing this classification with actual gaze locations, we extracted the hit (i.e., classifying gazed region as "gazed") and false alarm (i.e., classifying "ungazed" as gazed) for each frame. By varying the threshold between 0 and 1 at a 0.001 interval, a ROC curve was plotted; the area under this curve (AUC) indicated how well salience discriminate eye locations from random locations, in other words, how well the gaze locations were predicted by salience. Thus, AUC was the measure for our main dependent variable, SBGP. Uncentered, AUC had a possible range of 0 to 1, with .5 as the chance level

indicating that salience equally predicted where participants looked at and where participant did not look at. In the current study, we centered AUC at the chance level, such that the centered range was -.5 to .5, and positive values indicated that gazed locations were on average more salient.

Distance to Center (DTC). DTC was calculated to quantify the extent to which viewers fixated the center of the screen following cuts. We computed the Euclidean distance between the screen center and the gaze location for each data point. DTC data were then reduced using the same procedure as SBGP for Analysis 2: Data were averaged across participants to generate one average DTC measurement per participant group for each of four 160-ms bins surrounding each of 28 cuts (Figure 1).

Statistical Analysis. We first conducted preliminary analyses, including randomization checks and bivariate correlations to identify potential covariates (e.g., gender, exact age within each age group, naturalistic television exposure in the child group). In Analysis 1 examining overall age and condition effects across the entire vignette, the dependent variable was SBGP measured as AUC, which was averaged within each video shot for each participant. To account for the potential clustered standard errors at both the participant level (i.e., an individual participant may display similar eye movement patterns across the shots) and the shot level (i.e., participants may show similar patterns with each other when watching the same shots), we used multilevel modeling with shots and participants as the random effects. The proportion of variance in the outcome variable explained by the participant-level (i.e., between shots within participants) and shot-level clustering (i.e., between participants within shots) added up to 20% of the total variance, as indicated by the intraclass correlation. We used the function glmer from

the package lme4 (Version 1.1-12; Bates et al., 2015) in the R software environment (Version 3.3.0; R Core Team, 2016) to estimate the models.

Specifically, a mixed-effects model was used to estimate the fixed effect of age and video condition on AUC in data clustered at the participant and shot level. Given that participants and shots were crossed factors nested within each other, a crossed-mixed-effects model was fitted (Raudenbush, 1993) with the participant ID and shot ID as the crossed random effects (level 2) nested within observations (level 1). The model specification was as follows:

Level 1 model:

$$auc_{ij} = \beta_{0i} + \beta_{0j} + \varepsilon_{ij}$$

Level 2 model:

$$\beta_{0i} = \gamma_{00.1} + \gamma_{01}(age_i) + \gamma_{02}(condition_i) + \gamma_{03}(age_i \cdot condition_i) + \eta_{0i}$$

$$\beta_{0j} = \gamma_{00.2} + \theta_{0j}$$

Combined model:

$$auc_{ij} = \gamma_{00} + \gamma_{01}(age_i) + \gamma_{02}(condition_i) + \gamma_{03}(age_i \cdot condition_i) + \eta_{0i} + \theta_{0j} + \varepsilon_{ij}$$

In these models, γ_{00} represents the (intercept) grand mean of the reference group (adult-normal) across participants and shots. γ_{0q} represents the fixed effect of variable q (q = 1, 2, 3 for age, condition, and age-by-condition interaction) on the participant-level intercept, β_{0i}, and η_{0i} adds a random effect to β_{0i}. θ_{0j} denotes the random intercept (i.e., β_{0j}) effect at the shot-level. The residual at the participant-cross-shot level is represented by ε_{ij}.

In Analysis 2 examining age and condition effects immediately following cuts to new shots, the dependent variables were AUC and DTC, averaged across participants within each 160-ms bin for each of the four groups. To account for the clustering in the nested data at the bin level (i.e., participants may display similar eye movement patterns when watching the same 160-

min bins) and the shot level (i.e., participants may display similar eye movement patterns when watching different bins within the same window), we again relied on multilevel modeling with bins and shots as the random effects. The total intraclass correlation at the shot and the bin level was 22% for AUC and 26% for DTC. A piecewise linear regression, with the split point being placed between the first and the second bin, was used to capture the non-linear change in AUC and DTC across bins.

Specifically, a three-level hierarchical mixed-effects model with groups of participants (Level 1) nested within bins (Level 2) nested within shots (Level 3) was fit to test the fixed effect of age and video condition on AUC change following a cut. Two splines were used to fit a piecewise linear regression with a knot fixed at the time of the cut to index the change in AUC across the cut (phase1: bin 1-2 or -160 to +160 ms) and after the cut (phase 2: bin 2-4 or 0 to 480 ms). The model specification was as follows:

Level 1 model:

$$Auc_{ijk} = \beta_{0jk} + \beta_{1jk}(age_{ijk}) + \beta_{2jk}(condition_{ijk}) + r_{ijk}$$

Level 2 model:

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}(bin1_{jk}) + \gamma_{02k}(bin2_{jk}) + u_{0ik}$$

$$\beta_{1jk} = \gamma_{10k}$$

$$\beta_{2jk} = \gamma_{20k}$$

Level 3 model:

$$\gamma_{00k} = \delta_{000} + v_{00k}$$

$$\gamma_{01k} = \delta_{100}$$

$$\gamma_{02k} = \delta_{200}$$

Combined model:

$$auc_{ijk} = \delta_{000} + \delta_{100}(bin1_{jk}) + \delta_{200}(bin2_{jk}) + \gamma_{10k}(age_{ijk}) + \gamma_{20k}(condition_{ijk}) + v_{00k} + u_{0ik} + r_{ijk}$$

In these models, δ_{000} represents the intercept for the reference group (adult-normal). δ_{100} and δ_{200} represent the linear and quadratic effects of the bin variable, which models the change in AUC over time, on β_{0jk}. The bin variable, which was indexed by the time order of bins, was treated as a continuous predictor centered at the shot transition. The bin variable for the first (phase 1) and second (phase 2) piece of curve was denoted as bin1 and bin2, respectively. γ_{10q} represents the fixed effect of variable q (q = 1, 2 for age and condition). u_{0ik} and v_{00k} add random effects, at the shot level and bin level respectively, to the intercept. The residual in AUC at the group level is denoted by r_{ijk}.

Results

Preliminary Analyses

As a randomization check, we tested for experimental condition differences with respect to participant gender and exact age (i.e., age in years within the child or adult group). No significant differences were found in the child group [exact age: $t(16) = 1.66$, $p = 0.116$; gender: $\chi^2 (1, N = 20) = 0.88$, $p = 0.348$] or the adult group [exact age: $t(17) = -0.35$, $p = 0.730$; gender: $\chi^2 (1, N = 20) = < 0.001$, $p = 0.100$].

To identify potential participant-level covariates, we calculated bivariate correlations between the dependent variables and exact age, gender, and children's typical television exposure at home. See Table 1 for descriptive statistics, including bivariate correlations. Given that participant-level characteristics did not differ significantly by condition, nor were they correlated with the dependent variables, they were not considered further.

Table 1. Descriptive statistics and zero-order correlations for participant-level variables

	Descriptives			*Correlations*	
	Freq.	Mean	*S.D.*	AUC	DTC
Adult group					
AUC	--	0.21	0.06		
DTC	--	104.51	48.38		
Exact age	--	20.46	1.14	-0.22	-.14
Female	15%	--	--	0.33	0.17
Child group					
AUC	--	0.15	0.03		
DTC	--	123.85	45.49		
Exact age	--	4.51	0.10	0.03	0.29
Female	65%	--	--	-0.12	0.07
BTV	--	0.81	1.25	-0.37	-0.29
FTV	--	2.11	1.64	-0.19	0.04
Tot TV	--	2.92	2.78	-0.28	-0.16

Note: AUC was aggregated across all shots for each participant for bivariate correlations. Exact age was measured in years. The TV variables represented the hours of exposure per day averaged across a typical week, separately for background TV (BTV) and foreground TV (FTV) as well as total TV exposure (Tot TV).
N = 20 per age group. ***$p < .001$

Analysis 1. Salience-Based Gaze Prediction across the Entire Vignette

The distribution of SBGP is illustrated in Figure 2 by age and condition. The three fixed effects from the final model are reported in Table 2. Because the dependent variable was centered at chance, the intercept effect compares SBGP to chance in the reference group (adult-normal). As shown in *Model 1*, SBGP for adults in the normal condition was significantly above the chance level, $\gamma_{00} = 0.19$, $SE = 0.01$, $t(43) = 13.54$, $p < .001$. The age effect tests the difference between children and adults in the reference condition (normal). In the normal video condition, adults ($M = 0.69$, $SE = 0.04$) had higher SBGP than children ($M = 0.63$, $SE = 0.04$), $\gamma_{01} = -0.05$, $SE = 0.01$, $t(1050) = -6.01$, $p < .001$. The condition effect tests the difference between random and normal video in the reference age group (adults). Adults' SBGP was higher in the random-

edit video condition ($M = 0.23$, $S.E. = 0.03$) than the normal condition ($M = 0.19$, $SE = 0.04$), $\gamma_{02} = 0.04$, $SE = 0.01$, $t(1050) = 4.27$, $p < .001$.

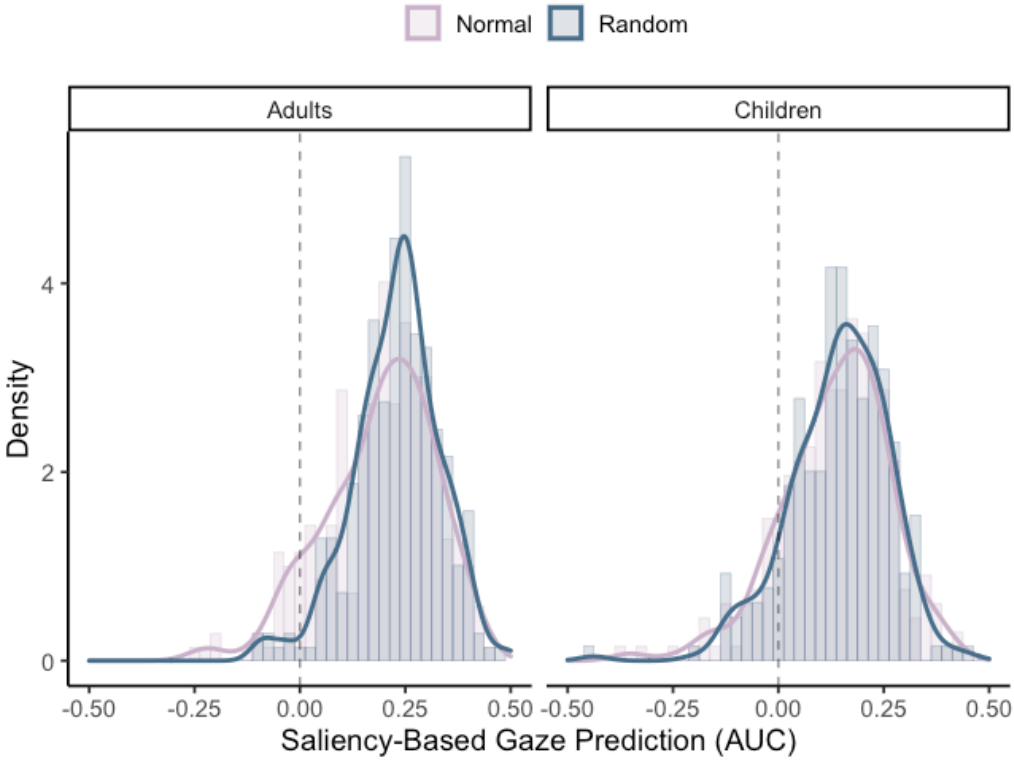

Figure 2. Histogram and density plot of SBGP, measured by AUC (centered at the chance level), by age group and video condition. The dashed line represents the chance level, so the area to the right of the line corresponds to positive SBGP.

Table 2. Fixed effects from the final mixed-effects model predicting SBGP for all shots in the vignette

Model 1				*Model 2*			
Predictor	β	*SE*	t-ratio	Predictor	β	*SE*	t-ratio
Intercept (γ_{00})	0.19	0.01	13.54***	Intercept (γ_{00})	0.14	0.01	10.05***
Age: Child (γ_{01})	-0.05	0.01	-6.01***	Age: Adult (γ_{01})	0.09	0.01	9.38***
Condition: Random (γ_{02})	0.04	0.01	4.27***	Condition: Normal (γ_{02})	-0.01	0.01	-0.76
Age × Condition (γ_{03})	-0.03	0.01	-2.43*	Age × Condition (γ_{03})	-0.03	0.01	-2.43*

Note: The dependent variable, SBGP, was centered at the chance level. Age was a binary variable with adult group coded as the reference group and child group as the contrast group in *Model 1*, and vice versa in *Model 2*. Condition was a binary variable with normal condition coded as the reference and random-edit video condition as the contrast group in *Model 1*, and vice versa in *Model 2*.
**p < .05. ***p < .001*

The final row of Table 2 (*Model 1*) depicts the interaction effect between the age group and video condition. Including this interaction effect significantly improved the fit of this model, $\chi^2(1) = 5.89$, $p < 0.05$. The interaction effect was significant, indicating that the condition effect was moderated by age, $\gamma_{03} = -0.03$, $SE = 0.01$, $t(1050) = -2.43$, $p < .05$. In *Model 2*, we rotated the reference groups to explore this interaction effect. As shown in Table 2 (*Model 2*), in the random-edit video condition, SBGP was again significantly higher in adults ($M = 0.23$, *S.E.* = 0.03) than children ($M = 0.14$, *S.E.* = 0.04), $\gamma_{01} = 0.09$, $SE = 0.01$, $t(1050) = 9.38$, $p < .001$. Moreover, the condition effect was not significant in *Model 2* where children were the reference group, $\gamma_{02} = -0.01$, $SE = 0.01$, $t(1050) = -0.76$, $p > .400$. That is, the condition effect was not significant for children. Although SBGP was lower in children than adults, children's SBGP was still significantly greater than chance, as indicated by a significant intercept effect with children as the reference group, $\gamma_{00} = 0.14$, $SE = 0.01$, $t(43) = 10.05$, $p < .001$.

Analysis 2. Salience-Based Gaze Prediction and Distance to Center Following Shot Transitions

Salience-Based Gaze Prediction. The temporal evolution of SBGP following a shot transition is plotted in Figure 3 as a function of age group and video condition. The plot depicts a decrease in SBGP at the transition in all four groups, followed by a recovery around 300 ms into the shots.

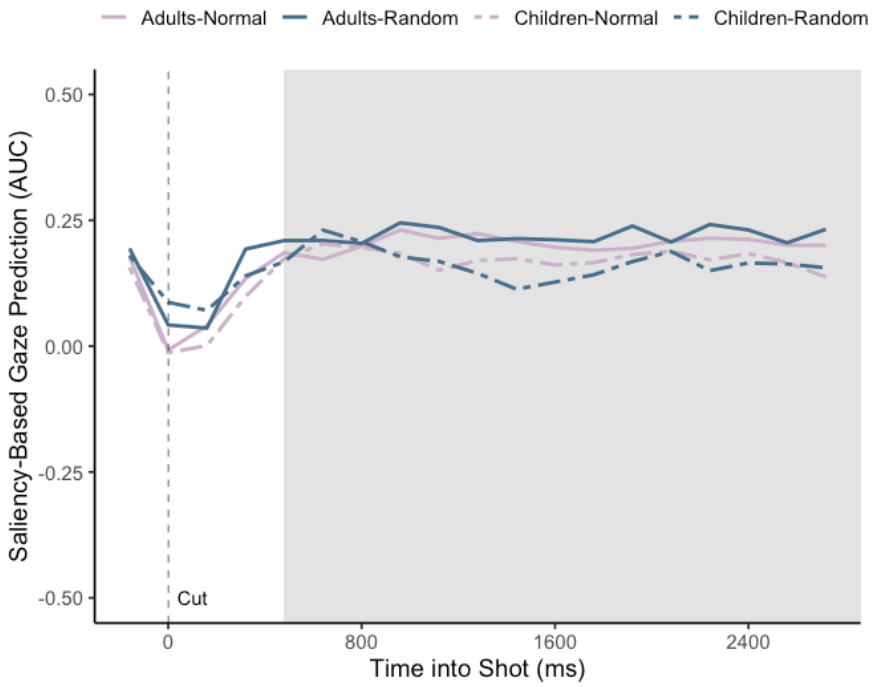

Figure 3. Temporal evolution of SBGP, measured by centered AUC (centered at the chance level), by age group and video condition beginning 160 ms before the transition to a new shot. Values are averaged over all shots. Analysis 2 focused on the window of time immediately surrounding the cut, as indicated by the unshaded area from -160 ms to 480 ms.

The mixed-effects model examined SBGP as a function of age group, condition, and time measured as bin1 (-160 ms to 160 ms) and bin2 (0 ms to 480 ms). A full report of fixed effects

can be found in Table 3. Consistent with the patterns shown in Figure 3, there was a significant negative effect of bin1, δ_{100}= -0.17, *SE* = 0.03, *t*(4) = -6.06, *p* < .010, and a significant positive effect of bin2, δ_{200}= 0.22, *SE* = 0.04, *t*(4) = 6.04, *p* < .010, suggesting a U-shaped pattern such that SBGP decreased as the cut occurred and then recovered following the cut. A main condition effect was found such that SBGP was higher for the random than the normal sequence, δ_{20k} = 0.04, *SE* = 0.01, *t*(416) = 3.26, *p* < .010. However, there was no difference in SBGP between the age groups, δ_{10k} = -0.01, *SE* = 0.01, *t*(416) = -0.80, *p* > .05. The age-by-condition interaction did not significantly improve the fit of the model, χ^2 (1) = 1.08, *p* = 0.298, and was therefore not included. Together, the model results suggest that SBGP was higher for the random than the normal sequence during this narrow window following cuts to new shots, regardless of age group.

Table 3. Fixed effects from the final mixed-effects model predicting SGBP after shot transitions

Predictor	β	*SE*	t-ratio
Intercept (δ_{000})	0	0.02	-0.23
Bin1 (δ_{100})	-0.17	0.03	-6.06**
Bin2 (δ_{200})	0.22	0.04	6.04**
Age: Child (γ_{10k})	-0.01	0.01	-0.80
Condition: Random (γ_{20k})	0.04	0.01	3.26**

Note: The dependent variable, SBCP, measured by AUC (centered at the chance level). Bin number denoted the time into a shot, with bin1 for phase 1 (160 ms before to 160 ms after the cut) and bin2 for phase 2 (0 ms to 480 ms after the cut); it was centered at the shot transition (i.e., the bin capturing 0 to 160 ms after the cut). Age was a binary variable with adult group coded as the reference group and child group as the contrast group. Condition was a binary variable with normal condition coded as the reference and random-edit video condition as the contrast group.
**p < .05. **p < .01*

Distance to Center. Prior research demonstrates that viewers, whether adults or young children, tend to look at the center of the screen following transitions to new video shots when

watching video in normal sequences (Kirkorian et al., 2012; LeMeur et al., 2007; Mital et al., 2010; Tseng et al., 2009; Tosi et al., 1997). Such centering of gaze could explain a drop in SBGP after a shot transition. Thus, we examined DTC immediately following shot transitions in the current study. Figure 4 plots the temporal evolution of DTC averaged over all shots, as a function of age group and video condition. Similar to SBGP, the plot depicts a decrease in DTC immediately after transitions to new shots in all four groups.

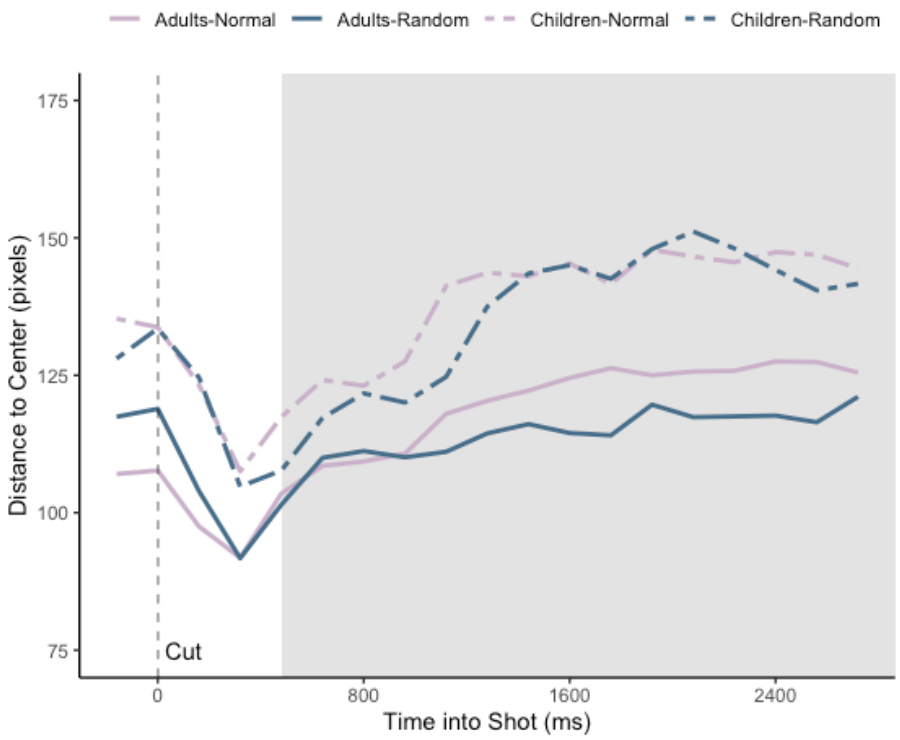

Figure 4. Temporal evolution of DTC by age group and video condition beginning 160 ms before the transition to a new shot. Values are averaged over all shots. Analysis 2 focused on the window of time immediately surrounding the cut, as indicated by the unshaded area from -160 ms to 480 ms.

Multilevel modeling was used to test the effect of age group and video condition on DTC following a shot transition, paralleling the model examining the temporal evolution of SBGP after shot transitions. The results are reported in Table 4. There was a non-significant effect of bin number at phase 1, δ_{100}= 1.88, *SE* = 5.00, *t*(420) = 0.38, *p* = .707, and a significant negative effect of bin number at phase 2, δ_{200}= -12.25, *SE* = 2.58, *t*(420) = -4.76, *p* < .001, suggesting a decrease of DTC after a shot transition. An age effect was found such that children displayed larger DTC (*M* = 123.37, *SE.* = 14.62) than adults (*M* = 106.14, *SE* = 14.65) immediately following a shot transition, δ_{10k} = 19.35, *SE* = 3.69, *t*(420) = 5.25, *p* < .001, while the condition effect was found not significant, δ_{20k}= 2.38, *SE* = 3.69, *t*(420) = 0.52, *p* = .518. The age-by-condition interaction did not significantly improve the fit of the model, χ^2 (1) = 1.57, *p* = 0.210, and was therefore not included. Thus, the distance to center was higher in children than adults, regardless of condition, during this narrow window immediately following shot transitions.

Table 4. Fixed effects from the final mixed-effects model predicting distance to center after shot transitions

Predictor	β	*SE*	t-ratio
Intercept (δ_{000})	112.97	6.22	15.15***
Bin1 (δ_{100})	1.88	5.00	0.38
Bin2 (δ_{200})	-12.25	2.58	-4.76***
Age: Child (γ_{10k})	19.35	3.69	5.25***
Condition: Random (γ_{20k})	2.38	3.69	0.52

Note: Bin number denoted the time into a shot, with bin1 for phase 1 (160 ms before to 160 ms after the cut) and bin2 for phase 2 (0 ms to 480 ms after the cut); it was centered at the shot transition (i.e., the bin capturing 0 to 160 ms after the cut). Age was a binary variable with adult group coded as the reference group and child group as the contrast group. Condition was a binary variable with normal condition coded as the reference and random-edit video condition as the contrast group
****p < .001.*

Discussion

The main purpose of this study was to determine (1) whether video comprehensibility impacted the degree to which gaze was predicted by low-level visual features during naturalistic video viewing, (2) whether such salience-based gaze prediction differed for adults and young children, and (3) how salience-based gaze prediction changed immediately following cuts to new video shots. We found that both 4-year-olds' and adults' eye movements were predicted by visual salience, as viewers in both groups were more likely to look at the high-salience regions than randomly selected regions. Overall, across the entire video, perceptual salience was a stronger predictor in adults than in 4-year-olds. However, while visual salience predicted 4-year-olds' eye movements equally in both video conditions, adults were less likely to direct their gaze toward visually salient regions while watching the shots in their original, comprehensible sequence than in a random sequence. This finding suggests that salience-based gaze prediction increased when video was rendered less comprehensible for adults but was not impacted in 4-year-olds. Implications for age-related changes in visual attention and comprehension to video are discussed.

Effects of Comprehensibility on Adults' Salience-Based Gaze Prediction

Our findings replicate prior research in demonstrating that salience-based gaze prediction increases with age, with more predictable fixations among adults than infants and young children (Franchak et al., 2016; Frank et al., 2009; Rider et al., 2018). Based on correlational designs, previous research, however, failed to attribute adults' higher salience-based gaze prediction to their better comprehension-driven voluntary control as compared to children. This is because, in addition to improved video comprehension, many age-related differences exist that could increase the predictability of adults' eye gaze, such as temporal and spatial intersubject

consistency in eye movements (Frank et al., 2011; Franchak et al., 2016; Kirkorian & Anderson, 2018).

The current study was intended to more directly test the *comprehension-based hypothesis* by examining the extent to which age-related increases in salience-based gaze prediction can be explained by age-related increases in comprehension of the content. To that end, we compared salience-based gaze prediction in adults watching a normal video sequence versus those watching the same video shots in a random sequence. We found that salience-based gaze prediction in adults increased, rather than decreased, when watching the random video sequence. This result suggests that comprehension-related, top-down control of eye movements cannot account for the age-related increase in salience-based gaze prediction, eliminating the *comprehension-based hypothesis.*

The findings from this experimental study provide more direct evidence that better comprehension of the video does not necessarily lead to higher salience-based gaze prediction in adults. Why might salience-based gaze prediction increase when adults watched the less comprehensible video sequences? Presumably, in the absence of comprehension-driven process, adults who might otherwise comprehend the content well may rely more heavily on perceptual features to guide their attention. For example, prior research on eye gaze patterns before and after cuts demonstrated that adults sometimes anticipate the reappearance of an object after a cut (Kirkorian & Anderson, 2017). This top-down control over eye movements could lead viewers to look at parts of the screen with low salience, such as an empty part of the screen where the object is to reappear. By contrast, without a meaningful plot and coherent action sequences, adults watching a random shot sequence may be more likely to be driven by bottom-up, reactive processes than to search for meaningful information or predict upcoming content.

Our findings are consistent with previous research demonstrating higher salience-based gaze prediction during free viewing as compared to during task-orientated viewing in adults (Smith & Mital, 2013). Smith and Mital suggested that providing an explicit task goal (i.e., identifying the locations depicted in the video clips) could direct viewers' attention away from the visually salient features. Together, these findings indicate that decreasing top-down influences in turn increases salience-based gaze prediction in adult viewers. Given that perceptually-salient regions may overlap with meaningful ones (Wass & Smith, 2014), future work could examine the extent to which the degree of overlap between visually salient features and semantic features impacts the magnitude, even the direction, of top-down influences on salience-based gaze prediction.

Age Differences in Salience-Based Gaze Prediction (During Normal Video Viewing)

What drives the age-related increase in salience-based gaze prediction during normal video viewing, if it cannot be explained by a viewer's ongoing comprehension of a story as it unfolds in a logical sequence (i.e., the *comprehension-based hypothesis*)? Given the developing dominance of bottom-up and top-down influences on visual attention during video viewing, we are less convinced by the *salience-based hypothesis* that older viewers' eye movements are simply controlled more by visually-salient features as compared to younger viewers'. Theories of children's attention development in general and overt attention to television in particular suggest an increasingly dominant role of content comprehension versus formal features in visual attention during video viewing (e.g., Huston & Wright, 1983; Ruff & Rothbart, 1996; Gola & Calvert, 2011). Specifically, such theories posit that children's attention is initially driven by salient features in a reflexive manner due to a lack of efficient attention and executive control. With increased maturation and video experience, a higher-level attention system is engaged, and

meaningful information, such as narratives or informational content, increasingly directs visual attention.

A more plausible explanation might be the *overall predictability hypothesis.* Specifically, it is possible that the higher salience-based gaze prediction in adults was due to a global tendency that adults' gaze behavior is generally more predictable than children's during free viewing of edited video. Adults' eye movements are likely controlled by a combination of low-level attention strategies and high-level attention strategies. For example, viewers may attend to salient regions because these regions are more likely to be meaningful and informative (e.g., Wass & Smith, 2014; Taya et al., 2011). Simultaneously, they also look at other non-salient regions that they understand as semantically relevant or where they anticipate something interesting to occur (e.g., Henderson et al., 1999; Land et al., 1999; Morgante, Haddad, & Keen, 2008; Kirkorian & Anderson, 2017). Both viewing experience and cognitive maturation are needed to develop these sophisticated strategies. Thus, as compared to children, adults' gaze could be more systematically attending to both visually salient and semantically meaningful regions, resulting in better salience-based gaze prediction as well as comprehension-based gaze prediction. As evidence, previous studies comparing eye movements to socially-meaningful features (i.e., faces) versus visually-salient features show age-related increases in the performance of both face-based predictive models and salience-based predictive models (Franchak et al., 2016; Rider et al., 2018), which the authors explained as adults looking more at both faces and high-salience regions (Franchak et al., 2018). However, since faces could be both semantically-relevant and visually-salient, future research is needed to disentangle top-down and bottom-up influences on eye gaze.

In addition, it is important to note that the age-related increase in salience-based gaze prediction is not universal. For instance, Kadooka & Franchak (2020) showed seven 2-min video clips to children of a wide age range (6-months- to 10-years-old) and adults. They found that, except of two videos (e.g., *Sesame Street*, and a music video), there was no age effect in looking at relatively more salient regions. A potential reason for the conflicting findings between our study and theirs is the particular video content, as Kadooka and Franchak suggest. This leaves ample space for future studies to investigate the exact mechanism underlying the developmental changes in visual attention to perceptually-salient features.

Effects of Comprehensibility on Children's Salience-Based Gaze Prediction

Like adults, 4-year-olds showed slightly higher salience-based gaze prediction when watching the randomly-edited video. However, unlike adults, the condition effect was not significant in children when considering the overall main effect of condition across the entire video (Analysis 1). On the surface, this might be due to 4-year-olds' insufficient comprehension to the normal video sequence in the first place. Prior research demonstrates that children at this age have limited understanding of video content, with adult-like comprehension emerging around 12 years of age (Anderson & Hanson, 2010; Collins & Wellman, 1982). Thus, the degree to which their visual attention is controlled by content comprehension may be relatively low even when watching the comprehensible normal version. This could explain why children's salience-based gaze prediction remained at the similar level, given that randomizing the shot sequences should not change the overall salience of the video.

Although video comprehension develops gradually across childhood, accumulated evidence demonstrates that young children do have sensitivity to the comprehensibility of video sequences, reflecting some – albeit limited — comprehension of video (e.g., Smith et al., 1985;

Pempek et al., 2010; Richards & Cronise, 2000). Indeed, our Analysis 2 shows some evidence for a condition effect in children when examining the short window of time surrounding transitions to new video shots. Similar to adults, 4-year-olds' salience-based gaze prediction dropped immediately following cuts and was higher for the random video sequence than the normal video sequence. This adult-like eye movement pattern suggests that differences existed in processing shot transitions while the children were viewing the normal versus random-edit video. Moreover, a prior study, using the same random-sequence manipulation as the current study, found that children as young as 18 months old made longer looks toward normal sequences as compared to randomized sequences of the same shots (Pempek et al., 2010). Such an obvious looking preference for coherent video sequences indicates a perception of shot relations emerges even in infants.

Given that children are sensitive to random-edit video manipulations by 18 months old, why did 4-year-olds' salience-based gaze prediction show relatively little change while viewing the incomprehensible video in the current study? Firstly, it is possible that the earliest beginnings of video comprehension are evident in overt looks toward the screen as in Pempek et al. (2010) but not in their eye movement patterns as observed in the current study. In this sense, differences in degree of salience-based gaze prediction might be more subtle than differences in overt gaze toward the screen, requiring a more sophisticated understanding of the narrative. Secondly, given that eye movement patterns tend to be more idsiosyncratic in children than in adults (Frank et al., 2011; Franchak et al., 2016; Kirkorian & Anderson, 2018), perhaps a general effect of condition across the video was less detectable in children. This is consistent with findings from Analysis 2 regarding the age group difference in distance to the center. That is, immediately following a cut

to a new shot, eye gaze was more consistently drawn to the center of the screen in adults than in children.

Limitations and Future Directions

The current study has some limitations that should be considered in interpreting the results and considering future research directions. First, as a secondary data analysis, we are limited by the quantity and quality of the existing data set. Eye-tracking technologies and computational methods have improved since these data were collected. The findings should be replicated in a larger, more diverse sample of participants with less data loss. Second, the present study did not have a direct measure of comprehension, although we believe that viewer's comprehension was successfully manipulated by rearranging the video shots in a random order. Future research should directly examine the relation between salience-based gaze prediction and viewers' comprehension of the video. Third, future research could consider individual differences in children's prior experience with TV and other edited video as a moderator of eye gaze (e.g., salience-based gaze prediction, anticipatory eye movements). We did not observe a correlation between salience-based gaze prediction and naturalistic TV viewing in the child sample. However, our sample was small, and the media exposure measure did not capture the content of video exposure. Future work could examine the impact of video viewing experience, with a focus on video content, on the effects observed in the current study. Additionally, while the current study focused on visual salience, the video stimuli used was accompanied by the soundtrack. It is possible that experimentally randomizing the shot at the audio level added noise to the video condition effect at salience-based gaze prediction at the visual level. Future work could considerate to control for this influence by only manipulating visual features rather than the audio.

Conclusion

In summary, the present study replicated the finding that perceptual salience is a stronger predictor of eye movements in adults than in 4-year-olds, bridging prior research on free viewing in infants and older children. Moreover, we found that disrupting video comprehensibility increases overall salience-based gaze prediction in adults but not 4-year-old children. Our findings, from the perspective of gaze prediction, extended previous work and suggested that adults' gaze behavior is more predictable than children's, perhaps due to more strategic and systematic use of informative visual features. Among other findings, this study also demonstrated some sensitivity to edited video in 4-year-olds that was limited to the window of time surrounding transitions to new video shots. Together, the present findings underscore the complex relation between visual attention and video comprehension, and between bottom-up and top-down processes. These relations are not straightforward, and future research should continue to use experimental methods to tease apart the effects of comprehension and other top-down processes on attention.

References

Alwitt, L. F., Anderson, D. R., Lorch, E. P., & Levin, S. R. (1980). Preschool children's visual attention to attributes of television. *Human Communication Research*, *7*(1), 52-67.

Amso, D., Haas, S., & Markant, J. (2014). An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PloS one*, *9*(1), e85701.

Anderson, D. R., Fite, K. V., Petrovich, N., & Hirsch, J. (2006). Cortical activation while watching video montage: An fMRI study. *Media Psychology*, *8*(1), 7-24.

Anderson, D. R., & Hanson, K. G. (2010). From blooming, buzzing confusion to media literacy: The early development of television viewing. *Developmental Review*, *30*(2), 239-255.

Anderson, D. R., Lorch, E. P., Field, D. E., & Sanders, J. (1981). The effects of TV program compre- hensibility on preschool children's visual attention to television. *Child Development*, 52, 151–157.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67,1–48. doi:10.18637/jss.v067.i01

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Social attention and real-world scenes: The roles of action, competition and social content. *The Quarterly Journal of Experimental Psychology*, *61*(7), 986-998.

Calvert, S. L. (1988). Television production feature effects on children's comprehension of time. *Journal of Applied Developmental Psychology*, *9*(3), 263-273.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of vision*, *6*(9), 4-4.

Castelhano, M. S., Wieth, M., & Henderson, J. M. (2007, January). I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In *International workshop on attention in cognitive systems* (pp. 251-262). Springer, Berlin, Heidelberg.

Collins, W. A., & Wellman, H. M. (1982). Social scripts and developmental patterns in comprehension of televised narratives. *Communication Research*, *9*(3), 380-398.

Cutting, J. E., Brunick, K. L., DeLong, J. E., Iricinschi, C., & Candan, A. (2011). Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception*, *2*(6), 569-576.

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy*, 21, 262–287.

Gola, A. A. H., & Calvert, S. L. (2011). Infants' visual attention to baby DVDs as a function of program pacing. *Infancy*, *16*(3), 295-305.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of the 19th International Confer- ence on Neural Information Processing Systems* (pp. 545– 552). Cambridge, MA: MIT Press.

Hawkins, R. P., Tapper, J., Bruce, L., & Pingree, S. (1995). Strategic and non-strategic explanations for attentional inertia. *Communication Research*, 22,188–206.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., &Mack, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain*. (pp. 537?562) Oxford: Elsevier.

Henderson, J. M., Weeks Jr, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of experimental psychology: Human perception and performance*, *25*(1), 210.

Huston, A. C., & Wright, J. C. (1983). Children's processing of television: The informative functions of formal features. In J. Bryant & D. R. Anderson (Eds.), *Children's understanding of television: Research on attention and comprehension* (pp. 35–68). New York, NY: Academic.

't Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., Konig, P., & Ein- h€auser, W. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head- fixed viewing conditions. *Visual Cognition*, 17, 1132–1158.

Ildirar, S., & Schwan, S. (2015). First-time viewers' comprehension of films: Bridging shot transitions. *British Journal of Psychology*, *106*(1), 133-151.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 631–637). IEEE.

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic imaging*, *10*(1), 161-169.

Itti, L. (2000). *Models of bottom-up and top-down visual attention* (Doctoral dissertation, California Institute of Technology).

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12, 1093–1123.

Kadooka, K., & Franchak, J. M. (2020). Developmental changes in infants' and children's attention to faces and salient regions vary across and within video stimuli. *Developmental Psychology*.

Kirkorian, H. L., & Anderson, D. R. (2017). Anticipatory eye movements while watching continuous action across shots in video sequences: A developmental study. *Child development*, *88*(4), 1284-1301.

Kirkorian, H. L., & Anderson, D. R. (2018). Effect of sequential video shot comprehensibility on attentional synchrony: A comparison of children and adults. *Proceedings of the National Academy of Sciences*, 115, 9867–9874.

Kirkorian, H. L., Anderson, D. R., & Keen, R. (2012). Age dif- ferences in online processing of video: An eye movement study. Child Development, 83, 497–507.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311-1328.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision research*, *47*(19), 2483-2498.

Lorch E. P., Bellack D. R., & Augsbach L. H. (1987) Young children's memory for televised stories: Effects of importance. *Child Development,* 58(2):453–463.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3,5–24.

Morgante, J. D., Haddad, J. M., & Keen, R. (2008). Preschoolers' oculomotor behaviour during their observation of an action task. *Visual cognition*, *16*(4), 430-438.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*(2), 160-170.

Pempek, T. A., Kirkorian, H. L., Richards, J. E., Anderson, D. R., Lund, A. F., & Stevens, M. (2010). Video comprehensibility and attention in very young children. *Developmental Psychology*, *46*(5), 1283.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, *18*(4), 321-349.

Richards, J. E., & Cronise, K. (2000). Extended visual fixation in the early preschool years: Look duration, heart rate changes, and attentional inertia. *Child Development*, *71*(3), 602-620.

Rider, A. T., Coutrot, A., Pellicano, E., Dakin, S. C., & Mareschal, I. (2018). Semantic content outweighs low-level saliency in determining children's and adults' fixation of movies. *Journal of Experimental Child Psychology*, 166, 293–309.

Ruff, H. A., & Rothbart, M. K. (1996). *Attention in early development: Themes and variations*. Oxford University Press.

Schmitt, K. L., Anderson, D. R., & Collins, P. A. (1999). Form and content: Looking at visual features of television. *Developmental psychology*, *35*(4), 1156.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology*, *20*(7), 649-656.

Smith, R., Anderson, D. R., & Fischer, C. (1985). Young children's comprehension of montage. *Child development*, 962-971.

Smith, T. J., Levin, D., & Cutting, J. E. (2012). A window on reality: Perceiving edited moving images. *Current Directions in Psychological Science*, *21*(2), 107-113.

Smith, T., & Henderson, J. (2008). Attentional synchrony in static and dynamic scenes. *Journal of Vision*, *8*(6), 773-773.

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of vision*, *13*(8), 16-16.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11, 1–23.

Taya, S., Windridge, D., & Osman, M. (2012). Looking to score: The dissociation of goal influence on eye movement and meta-attentional allocation in a complex dynamic natural scene. *PLoS One*, *7*(6), e39060.

Tosi, V., Mecacci, L., & Pasquali, E. (1997). Scanning eye movements made when viewing film: Preliminary observations. *International Journal of Neuroscience*, *92*(1-2), 47-52.

Tseng, P. H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, *9*(7), 4-4.

Valkenburg, P. M., & Vroone, M. (2004). Developmental changes in infants' and toddlers' attention to television entertainment. *Communication Research*, *31*(3), 288-311.

Wass, S. V., & Smith, T. J. (2014). Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy*, *19*(4), 352-384.

Wang, H. X., Freeman, J., Merriam, E. P., Hasson, U., & Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision*, 12, 1–27.

CHAPTER 3

Abstract

The experiment reported here was designed to examine the effect of interactivity on toddlers' video-based symbolic transfer. Eighty children, ages 23 to 37 months, participated in an object-retrieval task in one of four conditions: 1) watching through a window as an experimenter hid a toy, 2) watching a video recording of an experimenter hiding a toy, 3) interacting by pointing toward the toy (through the window) to see an experimenter hiding the toy, and 4) interacting by tapping the toy (on the touchscreen) to see an experimenter hiding the toy. Across all trials, interactivity was found to enhance children's overall errorless performance when and only when they watched the hiding event in video; however, it did not disproportionately affect the rate of perseverative errors, nor did it affect the latency of children's searches. Together, these findings have implications for the particular mechanisms by which interactivity affects toddlers' symbolic transfer.

Keywords: symbolic retrieval, transfer, interactivity, video-mediated learning, toddlers

The Effect of Interactivity on Toddlers' Video-based Transfer in the Symbolic Retrieval Task

During the first few years of life, children gain the ability to learn from not only direct observations of the physical environment, but also from indirect experience, such as reading a picture book or watching television (DeLoache & Chiong, 2010; Waxman & Medin, 2007). The cognitive ability of representing something with something else enables children to acquire information from symbols, such as replica toys in symbolic play, graphs, and letters and numerals. As a defining characteristic of human cognition (DeLoache et al., 1999; Myers & Liben, 2012), this symbolic capacity plays an important role in children's acquisition of knowledge (Hatano & Inagaki, 2002; Tarlowski, 2006). Despite the rapid change in symbolic functioning of very young children (DeLoache, 1987), there is large variation across symbolic task constraints and individual differences (Myers & Liben, 2012). The present study aims at understanding symbolic transfer from videos to real life in toddlers and how could that be influenced by characteristics at the content, child, and socio-contextual level.

Development of Symbolic Capacity

Children in any culture are surrounded by numerous symbolic artifacts, such as pictures, calendars, maps, graphs, and television. However, it is challenging for young children to benefit from the wealth of knowledge provided by these artifacts, given research showing that the informative function of symbolic artifacts is far from transparent to young children (Callaghan, 1999, 2003; DeLoache, 2002; DeLoache & Burns, 1994). To transfer information contained in a symbol to its referent, a proficient symbol user must have the insight that the symbol has a referential function (DeLoache, 1995). Such a referential understanding has been found to develop rapidly in the first years of life (e.g., Bloom, 2000; DeLoache, et al., 1998; Liben, 1999;

Newcombe & Huttenlocher, 2000), even though its development could persist into adulthood (e.g., Callaghan et al., 2003, 2004; Liben, 1999; Klima & Bellugi, 1979; Namy, 2008).

A series of studies (e.g., DeLoache, 1991; DeLoache & Burns, 1994) have charted the development of symbolic understanding and use of symbolic artifacts employing the object retrieval paradigm. In the experiments, children have to use the symbolic artifact (e.g., a small-scale model of a real room) as the source of information to find a hidden object. For example, in the scale-model version of the task, children observe where a miniature toy (e.g., a small-scale toy of a real toy) was hidden in the artifact and use that information to find a larger toy (e.g., the real toy) hidden in the corresponding location in the referent space (e.g., the real room). As symbolic artifacts that are ubiquitous in many children's daily lives, pictures (e.g., photographs, line drawing) and replica objects (e.g., scale models) have been substantially studied in research on early development of symbolic functioning. Such an analogical transfer from symbolic artifacts to their referents was found to be quickly achieved by children from age 24 to 30 months when pictures were the symbols and from age 30 to 36 months when scale models were the symbols (DeLoache, 1987, 1991; DeLoache et al., 1991). The discrepancy in the developmental progress between the two artifacts has been attributed to artifact features that may affect young children's symbolic insight (for a review, see Troseth et al., 2019). Specifically, scale models are three-dimensional (3D) objects on their own right that afford physical exploration catering to children at this age, whereas pictures may be relatively less interesting as physically present objects so that children are more likely to attend to their representational role. That is, children may perceive a model room as an object in its own right, such as a dollhouse they can play with, whereas photographs are less likely to offer such affordances for physical play or exploration.

Beyond Symbolic Insight: Videos are More Complicated Symbols

Recent decades have seen an explosion of digital media in children's everyday lives (Rideout, 2016). Video-based digital media have many distinct affordances and features from other symbolic artifacts. For example, whereas videos provide information based on two-dimensional (2D) images that appears similar to picture symbols, videos also create highly realistic visual experience that seems less distinguishable from real objects. Video that features live actions with an accompanying soundtrack can realistically show people's behavior (Troseth et al., 2018). Video symbols, therefore, might have some unique symbolic characteristics (Sheehan & Uttal, 2016; Troseth et al., 2003a) and trigger different symbolic development progress in children. The complicated symbolic nature of videos is also reflected by conflicting empirical evidence regarding children's symbolic understanding of videos. For example, a study using the object-retrieval task found that, similar to findings with pictures, 2.5-year-olds can use video as symbols to retrieve hidden toys, suggesting that video's symbolic function is easier to be appreciated as a 2D medium (Troseth & DeLoache, 1998). However, other studies found that a large proportion of 3-year-olds interpreted objects in video images as real objects and attempted to manipulate them (Flavell et al., 1990; Suddendorf, 1999). Thus, the developmental course of understanding the symbolic nature of videos remains unclear.

As with other symbolic media (e.g., scale model), what is depicted on video represents real objects and events, thus requiring the insight to link the video symbol and its referent (Troseth & DeLoache, 1998). While this is not transparent to young children (Troseth, 2003a), research has documented the challenge by which children are less likely to succeed when using video to search a hidden toy in the real room as compared to non-symbolic, direct experience of seeing the hiding event, more likely to *perseverate* to search the correct hiding location in the

previous trial rather than the current trial (i.e., perseverative error), and even take longer to search (i.e., search latency) (Troseth & DeLoache, 1998; Schmitt & Anderson, 2002; Schmidt et al., 2007).

In addition to symbolic understanding, decades of research have indicated multiple factors hindering video-based transfer, which may be less of a concern for other symbolic media in daily life and thus have not been addressed in the literature on other symbolic artifacts such as pictures and scale models. Firstly, video lacks social relevance and contingence that benefit learning. Evidence has indicated that the absence of social relevance and contingency would weaken mental representations derived from video symbols through decreased arousal and engagement (Gergely, et al., 2007; Krcmar, 2010; Kuhl, 2007; Troseth et al., 2003, 2010). While this also applies to other symbolic media such as photographs and scale models, the lack of social relevance and contingence may have a more adverse effect for video symbols. One reason is that, unlike screen media, learning materials based on photographs (e.g., picture books) are more likely to involve parents, and parental engagement has been well documented to facilitate young children's learning from symbols (e.g., Strouse et al., 2013). As for scale models, since children use them in daily life more for entertaining purposes (e.g., doll house) than learning purposes, the need is minimized to appreciate their informative functions and learn from them, at least when children are young.

Another factor that limits transfer from videos – and other symbolic artifacts -- is the perceptual differences between the symbol and its referent. Even though *iconic* symbols that physically resemble their referents in one or more ways are designed to help users recognize or remember the referents (e.g., Carlson et al., 2005; Luk & Bialystok, 2005; Thompson et al., 2009), perceptual differences exist between the symbol and referent context and are particularly

salient to naïve users (e.g., Klima & Bellugi, 1979; Namy, 2008). Symbolic artifacts, particularly such 2D ones as pictures and videos, are distinguishable from the reality (Schmitt & Anderson, 2002). Such a difference would impact information encoding from the symbol and result in less detailed memories and take more cognitive resources (Carver et al., 2006; Kirkorian et al., 2016; Schmitt & Anderson, 2002). Taken together, symbolic transfer from videos is hindered by not only representational insight at the conceptual level, but may also involve other cognitive competences, such as mapping 2D images onto real 3D spaces. Thus, it is important to acknowledge and understand the unique, complicated characteristics of video symbols in order to facilitate young children's transfer of learning from screen media in the digital world.

Interactivity Effects on Video-based Symbolic Retrieval

Despite the surging interest in incorporating interactive features to foster children's learning from videos, it remains unclear about how those interactive features affect children's ability to transfer knowledge from symbols to the real world (Sheehan & Uttal, 2016). Empirical findings have been mixed on the effectiveness of interactivity in promoting the video-based transfer of information. For example, in an object-retrieval experiment, 2.5- to 3-year-olds who interacted (i.e., pressing buttons) with a computer game to reveal the hiding location of an object performed better than children who merely observed the video of the hiding event (Lauricella et al., 2010). However, more recent studies using other tasks found a better transfer of math knowledge when 3- to 5-year-olds watched a video of someone else playing the game as compared to when the children played the interactive game themselves (Alade et al., 2016; Schroeder & Kirkorian, 2016). Moreover, the inconsistent interactivity effects have been found even within studies. For example, research on 2D object retrieval (Choi & Kirkorian, 2016) and word learning (Kirkorian et al., 2016) found a positive impact of interactivity on

transfer from videos in younger 2-year-olds but a negative impact in older 2-year-olds. These findings suggest that the role of interactivity in transfer of information from video symbols to their referents in real life may vary considerably across development, individuals, and contexts. However, it remains to be seen whether video interactivity can complete ameliorate the challenge of symbol transfer (i.e., increase learning to the level of non-symbolic, live demonstrations).

Theoretical Accounts Related to Interactivity Effects. At least three mechanisms, which are likely complementary rather than mutually exclusive, could underlie interactivity effects on video-based symbolic retrieval. First, from a perspective of symbolic insight, interactivity could influence transfer from video by either impeding symbolic representation of video symbols or fostering the direct mapping between video symbols and their referents in reality (Sheehan & Uttal, 2016; Troseth et al., 2019). Specifically, interacting with video screens may lead young children to regard the video symbol as an object for them to manipulate and thereby distract their attention from its symbolic function of providing information about its referent. In this sense, interactive video would make it more difficult for young children to transfer information to real life (e.g., refer to memory encoded from the room rather than those encoded from video), as compared to non-interactive video. Evidence from scale model research shows that playing with a scale model for merely 10 minutes reduced 3-year-olds' object retrieval performance significantly (DeLoache, 2000). Whether this is the case for video symbols has not been tested.

On the other hand, interactivity may not hinder symbolic insight but rather enhance it. Interactivity could contribute to iconicity by clarifying the temporal relevance of events on video, thereby increasing children's perception of similarity between screen depictions and what they stand (Sheehan & Uttal, 2016; Troseth et al., 2019). This increased perception of similarly

could lead to two consequences: 1) facilitate symbolic insight, or/and 2) lead to a direct mapping between the video symbol and its referent in reality while bypassing symbolic insight (e.g., a cat tires to catch virtual fish on a touchscreen that dart away when touched). Both will result in a positive effect of interactivity on the transfer.

Secondly, interactivity provides contingencies between children and video, which could increase children's arousal and overall engagement by promoting motivation and a sense of agency (Kuhl, 2003, 2007; Beihler & Snowman, 1997) and increasing available cognitive resources (Kirkorian et al., 2016). All of these have been believed to be key elements in learning in both animals and humans (Kuhl et al., 2003; Doupe & Kuhl, 1999; Merzenich & Jenkins, 1995). Indeed, the important role of contingency has been well-documented in learning generally. In the real-life situation, social contingency based on interpersonal, two-way exchange has been found to enhance learning in young children (Bloom et al., 1987; Goldstein et al., 2003, 2009). Nonetheless, in the video context, evidence from studies on video contingency shows inconsistent findings. While some research demonstrates that toddlers retrieved objects and learned words better from interactive video than from non-interactive video, even in the absence of true interactions with contingently responsive social partners (Lauricella et al., 2010; Kirkorian et al., 2016), others show the opposite (e.g., Troseth et al., 2018; Tsuji et al., 2021).

Moreover, there are sparse and mixed evidence for whether this arousal/engagement mechanism is the primary or universal account for the effect of interactivity. In a 3D symbolic retrieval experiments, Lauricella et al. (2010) found that 30- to 36-month-olds transferred better when asked to press a key in the keyboard to see the hiding event. However, the 2D symbolic retrieval study by Choi and Kirkorian (2016) found that contingency benefited 24- to 30-months-olds' transfer only when the it was designed to emphasize specific information on the screen ,and

thus suggested that contingent interaction may support transfer mainly through other mechanisms rather than increased arousal and engagement. Given the many differences between the two studies (e.g., number of trial and hiding locations, child age, search space dimension, video device, interactivity interface), a study that uses 3D search space, like Lauricella et al. (2010), while replicating Choi and Kirkorian (2016) in other aspects would clarify the inconsistency in their findings as an effort to shed light on the mechanism underlying the interactivity effect.

Lastly, interactivity also could influence transfer from video through cognitive load (for a review, see Kirkorian, 2018). On the one hand, interactivity may support children in encoding information from video when it directs their visual attention to target information on the screen. Research demonstrates that toddlers performed better in word learning and 2D object retrieval when interacting with the screen by touching a specific region relevant to target information, compared to simply touching anywhere on the screen (Kirkorian et al., 2016; Choi & Kirkorian, 2016). For example, Choi and Kirkorian (2016) showed 24- to 36-months-olds videos of a 2D animated hiding event and then asked children to find the 2D object (i.e., a sticker paper) on the corresponding 2D felt board. During the search task, children who were asked to touch the specific location on the screen relevant to the hiding location performed better than children who touched anywhere on the screen across all four trials, suggesting that the contingency afforded by interactivity facilitates the symbolic transfer, perhaps through selective attention to and encoding of target information on the screen. On the other hand, interactivity might influence cognitive load during encoding information from video in an alternative manner. Specifically, interactivity could also create extra cognitive demands and hence distract children from the target information. Evidence is from research on transfer of math concepts with 3- to 5-year-olds

in which watching a game was more beneficial than playing a game, particularly for a task with higher demands (i.e., relatively far transfer between perceptually different contexts) (Alade et al., 2016; Schroeder & Kirkorian, 2016).

Individual Differences and Social Context

Individual differences and social-context factors have been well documented in research of media effects on young children, and learning is no exception (for a review, see Valkenburg & Peter, 2013). The literatures on both symbolic understanding and video-based learning emphasize the importance of taking into account the individual children and the social context when video symbols are being used. Given the varying effects of video interactivity, it is particularly essential to consider individual differences and social context factors that could moderate interactivity effects.

Regarding individual differences at the development level, age is the foremost factor according to decades of research on video-based transfer (Strouse & Samson, 2021) and symbolic development (DeLoache et al., 1999). Many aspects of cognitive competence increase with age, such as working memory, cognitive inhibitory control, and cognitive flexibility (Strouse & Samson, 2021). All of these skills have been found to impact children's performance with symbolic artifacts (Hartstein & Berthier, 2018; Jenkins & Berthier, 2014; Barr, 2010; Barr et al., 2016). As such, age should predict children's video-based symbolic transfer.

Additionally, at the socio-contextual level, children's daily media use and household media environment could also play a role. According to DeLoache et al. (1999), efficient symbol use needs substantial social support. For example, in the family environment, parental co-use of symbolic media (e.g., active mediation) scaffolds the transfer of learning by connecting information on screen to real life experience, thereby linking entities represented on the screen to

those in the real world (Strouse et al., 2018; Myers et al., 2018). Research demonstrates that young children's naturalistic experience with interactive media, but not noninteractive, is related to their video-based symbolic retrieval in the laboratory (Kirkorian & Choi, 2017; Troseth et al., 2007). As reviewed earlier, children's prior experience with symbolic artifacts could influence their symbolic insight for the artifacts (DeLoache, 2000). Considering the pervasiveness of screen media today (Rideout & Saphir, 2013), individual differences in naturalistic experience with videos could impact their performance on tasks in laboratory experiments. Indeed, naturalistic experience with live, not pre-recorded, video positively was found to predict video-based symbolic retrieval in toddlers (Troseth et al., 2007).

Overview of the Current Study

The purpose of this study was threefold: 1) to determine whether interactivity could improve toddlers' video-mediated symbolic retrieval to the level of performance achieved in the non-symbolic situation, 2) to determine whether interactivity influences toddlers' symbolic and non-symbolic retrieval differently, and 3) to identify potential relations between symbolic retrieval and child characteristics. We used the symbolic object-retrieval retrieval paradigm widely adopted by symbolic understanding research (DeLoache, 1987 Troseth, 1998) and video-mediated learning research (Schmidt et al., 2007; Lauricella et al., 2010). As suggested by prior studies (e.g., Choi & Kirkorian, 2016; Deocampo & Hudson, 2005), this task enabled a microgenetic approach to explore the mechanism of mental representation by examining transfer across multiple trials, with a particular focus on perseverative errors. With a Modality (2: video vs window) × Interactivity (2: point vs watch) between-subjects design, there were four experimental conditions. The window-watch group watched through a window as an experimenter hid a toy, the video-watch group watched a video recording of an experimenter

hiding a toy, the window-point group pointed toward the toy (through the window) to see an experimenter hiding the toy, and the video-point group tapped the toy (on the touchscreen) to see an experimenter hiding the toy.

In addition to errorless retrieval, we also examined perseveration errors and search latency as supplementary metrics of retrieval performance. Preservation has been examined in relevant studies as an important measure to understand young children's symbolic retrieval (e.g., DeLoache & Burns, 1994; Troseth, 2010; Schmidt & Anderson., 2007), because it reflects the competition between mental representations of multiple trials, which helps elucidate underlying mechanisms of error retrievals (e.g., random guessing versus proactive interference from prior trials). Similarly, search latency can reveal group differences in children's search performance even when differences do not appear in the rate of errorless retrievals (Schmitt & Anderson, 2002).

Based on literature on children's video-mediated transfer (e.g., Schmitt & Anderson; Troseth & DeLoache, 1998; Troseth et al., 2007), we predicted that children would have better retrieval performance (e.g., more errorless retrievals, fewer perseveration errors, shorter search latency) when watching the window demonstration through a window as compared to watching the video demonstration. Regarding the impact of interactivity, there is mixed theoretical and empirical evidence. Although interactivity is proposed to promote young children's transfer by increasing arousal/engagement through contingency and fostering a direct mapping between video symbols and the reality and/or enhancing symbolic insight through perception of similarity, it is also believed to distract young children's attention from the representational function of video symbols and thereby hinder symbolic insight (for reviews, see Sheehan & Uttal, 2016; Troseth et al., 2019). Empirical research also suggests that interactivity could either

facilitate or impede symbolic transfer through cognitive load (for a review, see Kirkorian, 2018). Thus, the effect of interactivity on vide-mediated symbolic retrieval remained an open research question.

Of particular interest in this study was the potential interaction between modality and interactivity, as it would help disentangle multiple mechanisms underlying the interactivity effect. To the extent that interactivity benefits video-mediated retrieval performance, we sought to determine whether such a benefit improves toddlers' video-mediated symbolic retrieval to the level of performance achieved in the non-symbolic situation. Prior research has demonstrated a positive effect of interactivity in younger preschoolers merely through contingent responding in sync with the child's behavior (e.g., Kirkorian et al., 2016; Choi & Kirkorian, 2016). The current study extended this research on touchscreens by comparing a similar interactive video experience to a live condition. Such a comparison enabled us to understand whether this interactivity interface (i.e., contingency based on synced actions in the absence of general interpersonal social cues) was sufficient to ameliorate the difficulty with symbolic transfer and improve young children's performance to the level achieved in non-symbolic situation. We treated this as an exploratory question, due to the mixed findings on the difference in children's symbolic retrieval performance between an interactive-video condition and live conditions (e.g., Lauricella et al., 2010; Troseth et al., 2018).

Furthermore, by including a live-interactive (henceforth, window-point) condition, this study allows us to compare interactivity effects between video and live situations. Different theoretical accounts for interactivity effects would make different predictions for video versus window conditions. First, interactivity would impact, in either a positive or negative direction, retrieval performance by influencing children's *symbolic insight*. On one hand, interacting with a

symbolic artifact (e.g., physical manipulation) could increase the symbol's salience as an object and distract children attention from appreciate its representational function, thereby hindering symbolic insight (Tare et al., 2010; Troseth et al., 2019;). If so, we expected a negative effect of interactivity only in the video conditions rather than window conditions in which no symbolic relation was involved. This effect would be evidenced by higher errorless retrieval, lower perseveration, and shorter latency. On the other hand, interactivity could enhance symbolic performance by facilitating symbolic insight or the direct mapping between video symbols and their referents in reality by highlighting the similarity between them (Troseth et al., 2019; Sheehan & Uttal, 2016). If this is the case, we expected totally reversed results. Importantly, if interactivity influences, regardless of whether positively or negatively, transfer through impacting children's symbolic insight, we expected it influences the perseverative errors at the information-representing stage. Specifically, a better (or worse) symbolic insight would lead children to be better (or worse) at linking the video and the room, and their mental representation of video in the current trial would thereby be less (or more) likely to be interfered by that of the live retrieving experience in the previous trial.

Second, interactivity may improve retrieval performance by increasing *selective attention and encoding* of target information (Choi & Kirkorian, 2016). If so, we expected greater improvement in video conditions than in window conditions, as evidenced by higher errorless retrieval. This is because toddlers typically attend to and encode information inefficiently from video that is perceptually impoverished relative to real-life experience (Kirkorian et al., 2015; Troseth et al., 2003, 2010). Because this is not the case for in-person events, interactivity was expected to not have an impact on children in window conditions if it supports learning from video by increasing selective attention and encoding. However, unlike the influence on

information-representing stage as in the *symbolic insight* hypothesis, influences on the attending and encoding stage are less likely to affect perseveration (Choi & Kirkorian, 2016). Thus, we expected no effect on perseverative errors under this *selective attention and encoding hypothesis.*

Finally, if interactivity benefits retrieval performance by increasing *arousal and engagement*, we predicted a generalized benefit regardless of the modality. The contingency afforded by interactivity might increase a sense of agency and better engage children, thus increasing available resources and learning generally (Beihler & Snowman, 1997; Kuhl, 2003; Kirkorian et al., 2017), in both video and window conditions, particularly given the minimized interpersonal social cues involved window conditions of the present study. In addition, we also predicted that such a contingency would reduce preservative errors, if the arousal and engagement account holds. This is because increased mental resources, among other benefits, could strengthen mental representation such that the mental representation of the hiding event in current trial to be more competitive over that in the outdated trials (Choi & Kirkorian, 2016). Even though Choi and Kirkorian (2016) found no effect of interactivity on perseverative errors and thereby suggested that increased arousal and engagement is less likely to account for the positive interactivity effect in their study, the present study would test the generalizability of this effect by comparing the effects of a similar interactive video experience across different situations through a direct test of interactivity-by-modality interaction.

Regarding individual and socio-contextual factors, we hypothesized, consistent with previous findings (Kirkorian, 2018; Strouse & Troseth, 2014), an association between child age and search performance, including errorless retrieval, perseveration error, and search latency. Given the association found by some studies between naturalistic media experience and children's symbolic transfer (Kirkorian & Choi, 2017; DeLoache, 2000; Troseth 2003, 2010), we

predicted children's search performance would correlate with their touchscreen exposure and use at home. More importantly, despite the well-documented scaffolding effect of parent co-use practices in children's symbolic thinking and connecting video and reality (Strouse & Troseth, 2014), no research has examined whether this scaffolding could transfer to other tasks where parents are not involved. Thus, we examined the correlation between parental mediation at home and children's search performance in the lab as an open research question.

Method

Participants and Design

We selected 24- to 36-month-old children because that is when children are most likely to exhibit the difficulty transferring information from symbolic artifacts (e.g., Schmitt & Anderson, 2002; Troseth & DeLoach, 1998) and the age range that has been typically studied by relevant studies on video-mediated symbolic retrieval (e.g., Kirkorian et al., 2016; Choi & Kirkorian, 2016; Troseth et al., 2018). Considering previously reported age effects in young children's symbolic and video-mediated performance, we used a stratified randomization approach to achieve the covariate balance among conditions with respect to age (Suresh, 2011).

Eighty toddlers between 23 and 37 months of age (44 females, 36 males) were randomly assigned to one of the four conditions: video-watch (n = 19, $Mean_{age}$ = 29.2 months, 11 female), video-point (n = 20, M_{age} = 31.3 months, 11 female), window-watch (n = 21, M_{age} = 29.0 months, 12 female), and window-point (n = 20, M_{age} = 29.7 months, 10 female). An additional 11 children were dropped from the sample due to their refusal to participate (n = 9), technical problems with the camera (n = 1), or experimenter error (n = 1). Ninety percent of the children in sample were White; 3.8% African American, 2.5% Asian American, and 3.7% of mixed races. Regarding parent education, approximately 28% of the parents had a graduate degree, 36% had a

bachelor's degree, 15% had a high school degree, and the rest had a degree lower than high school. Data were collected from January 2016 through February 2020 in a small college town in Virginia, USA.

Parent Survey

The parent completed a brief survey that consisted of 13 questions about demographic information, such as the child's race or ethnicity and parents' educational background, and information about the child's media exposure and use at home. For touchscreen exposure, the parent rated how often the child used a touchscreen device on average using an 8-point Likert scale, where response options ranging from "never" to "more than 3 times a day". The parent also reported the age (in months) when the child began to regularly use a touchscreen device, how many minutes the child used any type of touchscreen device the previous day, and whether that use yesterday was typical. Additionally, parental mediation was asked by the question, "When your child uses a touchscreen, how often are you or another caregiver helping?" with a 5-point Likert scale, where response options ranging from "Does not use touchscreen" to "Always".

Setting and Materials

Children were tested individually by an experimenter and an assistant. Sessions took place at a university laboratory and lasted approximately 20-minutes. Two adjacent rooms (see Figure 1), a "hiding" room and an "observation" room, were used for the object-retrieval task. The hiding room was comfortably furnished with a carpet, a couch, and a TV set on a stand. As shown in Figure 1 and Figure 2, there were four additional items used as hiding places – a pillow on the couch, a basket on the floor, a carboard box on the floor, and an artificial potted plant on the floor. The target object that was hidden in this room was a small stuffed toy, "Sammy the

Turtle" (about 20 cm in length). Four stationary cameras, set in each corner of the room, recorded the whole session from four different angles to best capture the child's searching behaviors.

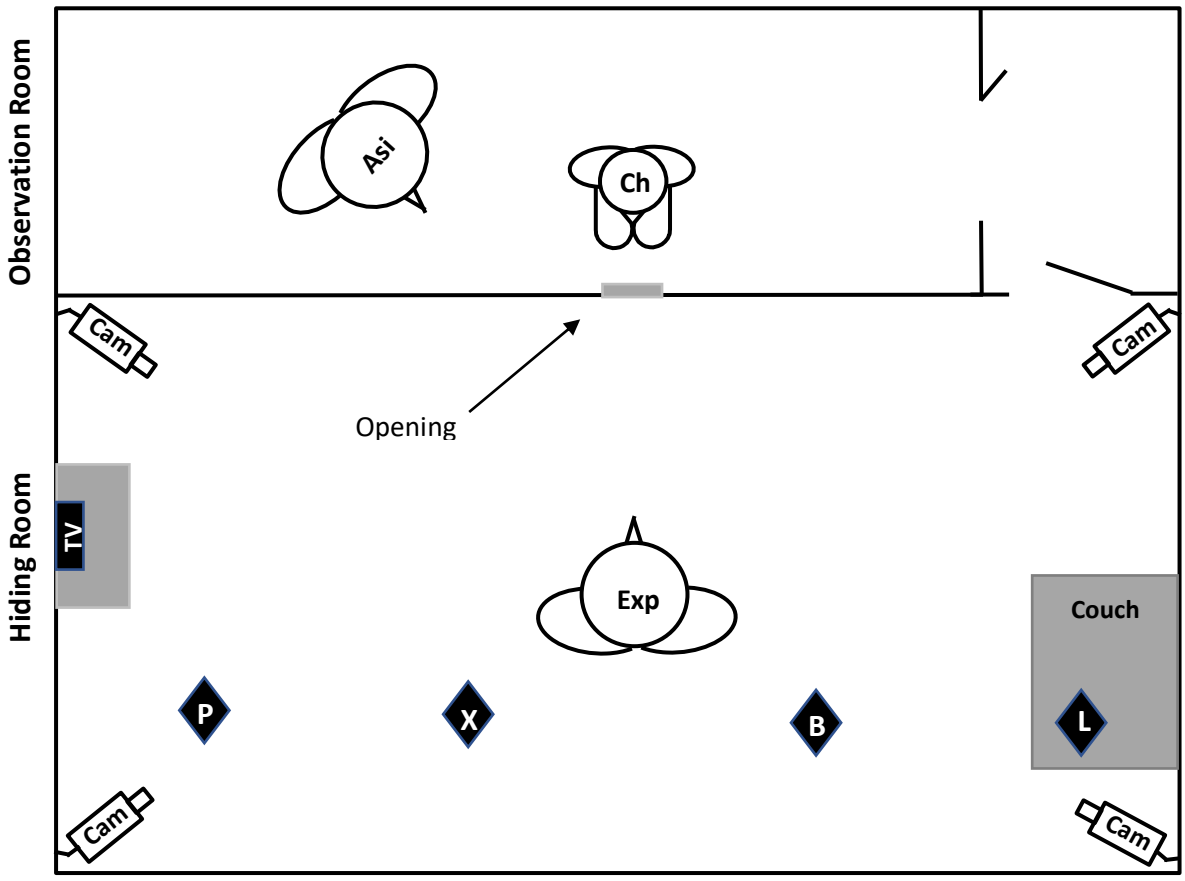

Figure 1. Spatial layout and setup of the hiding room and observation room. The opening was the "window" through which children looked into the hiding room under the window conditions and was covered by a same-size tablet under the video conditions. P stands for the plant, X for box, B for basket, and L for pillow.

The child watched the hiding event from the observation room, which was separated from the hiding room by a one-way mirror (transparent from the observation room to the hiding room). The mirror was covered with black poster board to prevent the child from looking into the

hiding room except through a rectangular opening the same size as our tablet computer (24.5cm × 14.9cm) such that the children in the window condition could see, through the "window", as the experimenter hid the toy in the hiding room. Note that, with the one-way mirror, the experimenter in the hiding room was unable to see the children in the observation room during the demonstration. The view the child saw from the observation room is shown in Figure 2.

Figure 2. Experimental setting in the hiding room from the child's point of view from the observation room, with four hiding locations (from the left to right: pillow, basket, box, plant) and the experimenter standing in the middle.

In the video conditions, video stimuli were used to demonstrate the hiding event on a touch-screen tablet computer (10.1-in. Galaxy Tab; Samsung America, San Jose, CA) using a mobile application that was developed for this project. The tablet computer was on a stand in front of the mirror, covering the opening "window" that was used in the window conditions, so that the perceptual differences between the window and video hiding events were minimized. The video stimulus was a recording of live actions of the hiding event as in the window conditions. In the video-interactive condition (henceforth, video-point condition), the video

recording was programmed to pause once the assistant instructed the child to interact (i.e., "Point to Sammy to see Sammy hide.") and resumed once the child finished the action. Children had to touch the screen on that particular location of the hiding location. Touching any other location did not resume the video. In the video-watch condition, the child merely watched the recording without any pause or any input from the child. The window conditions used the same script and analogous actions as the video conditions (see more details in Procedure section).

Procedure

Two researchers conducted sessions: the experimenter hid the toy on each trial without interacting with the child, while the assistant provided instructions and interacted with the child. There were three phases involved in the experimental session: orientation, correspondence training, and test. During the orientation phase, the assistant welcomed and interacted with the child to familiarize the child with the laboratory environment, while the experimenter interacted with the parent and provided the consent form and parent survey. After becoming acquainted with the researchers, the child was introduced to Sammy the Turtle and the hiding spaces. The experimenter showed four spaces in the hiding room, one by one in a fixed order by hiding and removing the toy in each space while labeling the hiding space (e.g., "Sometimes she hides Sammy behind the pillow"). The experimenter stood between the second and third hiding location (see Figure 2) and always came back to this point after she hid the toy. The parent was present throughout the experiment but was instructed not to interfere with the child's searching.

During the correspondence training phase, the child and the parent were escorted to the observation room, while the door of the hiding room remained closed. The purpose of this phase was to emphasize the correspondence between what was seen in the hiding event and that in the hiding room. For the video conditions, the child viewed the correspondence events on the tablet.

The assistant pointed out the relation between this correspondence event (displayed on screen) and the hiding room: “We have a special camera so we can watch where Sammy hides in the room. Look! It’s the room we were just in! Let’s look at all the places where Sammy can hide.” The assistant then labeled a hiding location and prompted the children to either *look at* or *point to* the location, depending on whether the child was assigned to the watch or point condition (e.g., “Sometimes she hides Sammy behind the pillow! Look at / point to the pillow to see Sammy hide.”). Then, in the video-watch condition, the video proceeded without any input from the child to show the experimenter walking to hide and remove the toy in that location. In the video-point condition, the video paused after the instruction to point to the location and only proceed once the child tapped the correct location on the screen (e.g., the pillow). The same correspondence procedure was repeated for each of the other three hiding locations.

The correspondence training phase for window conditions was identical to the video conditions with the following exceptions: 1) the instruction referred to a “special window” instead of a “special camera”, 2) children observed the live orientation through the opening in the one-way mirror rather than a video on the tablet computer, and 3) in the window-point condition, children pointed toward the hiding location through the window rather than touching it on the tablet computer. Specifically, in the window-point condition, the experimenter paused all actions and words once the assistant instructed the child to interact using the same script as in the video-point condition and resumed to hide the toy once the child finished the action. This was analogous to the video-point condition except that children pointed through the window into the hiding room rather than on the tablet computer. In the window-watch condition, the child merely watched the experimenter hiding the toy without any pause or any input from the child, which was analogous to the video-watch condition.

During the testing phase, there were four test trials in which the child first viewed the hiding event in the same manner as in the orientation, except that the child was instructed to look at / point to the toy rather than the hiding location. Next, the child was led to the hiding room and prompted to search for the toy. If the child failed to search or searched in incorrect locations, increasingly explicit prompts were given until the toy was retrieved. In this way, the child always experienced uncovering the toy in the correct location. However, only the child's initial, unprompted search attempt was coded for analysis. Each of four hiding locations was used once per trial for four trials total. The trials were used in one of two predetermined orders, counterbalanced across conditions.

Data Scoring Procedure

All the video recordings were double-coded independently by trained research assistants independently. For each trial, coders first determined if the child ever attempted to search for the toy. If the child did search, a series of variables were coded to capture the child's searching performance. An errorless retrieval was scored for each trial in which the child successfully retrieved the toy on the first attempt without any prompts from the experimenters. A perseverative retrieval was coded for trials 2 to 4 in which the child's first search in the current trial was at the correct location of the previous trial. For perseverative retrievals, we also coded self-correction retrieval, defined as when the child made a perseverative error on the first attempt but subsequently searched the correct location on the second attempt without any prompting. Additionally, search latency was coded for all retrievals as the time lag between the moment when the child stepped into the room and the moment when the child touched on the space in the first attempt, regardless of correctness. Interobserver reliabilities were high on errorless retrieval (97%), perseverative retrieval (100%), self-correction retrieval (100%), and search latency (ICC

= .993). Disagreements were resolved through consensus, once both coders viewed the videotape of the child again.

Analytical Approach

The primary analysis examined errorless retrievals as a function of experimental condition. To be consistent with prior research, we first calculated the proportion of errorless retrievals that combined all four trials for each child, a widely-examined outcome variable of errorless retrieval performance. We then conducted a two-way analysis of variance (ANOVA) to test the modality and interactivity effect on this variable. However, given a large body of research suggesting that symbolic retrieval performance in toddlers differs between trials (Kirkorian et al., 2016; Schmitt et al., 2002; Sharon & DeLoache, 2003), our main analysis was based on a mixed-effects logistic model with individual trial as the unit of analysis and the binary retrieval outcome – errorless (coded as 1) or error (coded as 0) – as the dependent variable. To account for the potential clustered standard errors at the child level (i.e., an individual child may display similar retrieval performance across the trials), a mixed-effects model was fitted with children as the random effect. Specifically, a two-level hierarchical random intercept model with trials (Level 1) nested within children (Level 2) was estimated to test the fixed effect of condition and trial on errorless search.

The proportion of variance in the dependent variable explained by the child-level clustering was up to 59% of the total variance, as indicated by the intraclass correlation. Despite a difference in overall errorless retrieval performance among children, we expected the changing pattern of errorless retrieval across trials to be consistent for all children. To test this, we examined a random slope model by adding a random effect to allow the slope of trial to vary

across children. However, this model showed no improvement in the model fit, compared to the random intercept only model and thereby was not considered further.

Given that many children perseverate on trial 2, many studies report a V-shape pattern in search performance, such that errorless retrieval drops from trial 1 to trial 2 and then increases from trials 2-4 (Sharon & DeLoache, 2003). To account for this "second trial dip" as observed in our data as well, two splines were used to model a piecewise linear regression with a knot fixed at the second trial to index the change in retrieval performance before (trial phase1: trial 1-2) and after the second trial (trial phase 2: trial 2-4). We used the function glmer from the package lme4 (Version 1.1-12; Bates et al., 2015) in the R software environment (Version 3.3.0; R Core Team, 2016) to estimate the model. The model specification was as follows:

Level 1 model (trial level):

$$\text{Probability } (ER_{ti} = 1) = p_{ti}$$

$$\log[p_{ti} / (1 - p_{ti})] = Y_{ti}$$

$$Y_{ti} = \beta_{0i} + \beta_{1i}(trial\_phase1_{ti}) + \beta_{2i}(trial\_phase2_{ti}) + Rti$$

Level 2 model (child level):

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(modality_i) + \gamma_{02}(interactivity_i) + \gamma_{03}(modality_i \times interactivity_i) + U_{0i}$$

$$\beta_{1i} = \gamma_{10}$$

$$\beta_{2i} = \gamma_{20}$$

Combined model:

$$Y_{ti} = \gamma_{00} + \gamma_{01}(modalityi) + \gamma_{02}(interactivity_i) + \gamma_{03}(modality_i \times interactivity_i) +$$

$$\gamma_{10}(trial\_phase1_{ti}) + \gamma_{20}(trial\_phase2_{ti}) + U_{0i} + R_{ti}$$

In these models, p_{ti} and Y_{ti} denoted the probability of errorless retrieval (ER in the equations above) and the log-odds of ER (i.e., the logit function) at trial *t* for participant *i*. The variable for trial, which was indexed by the time order of trials, was treated as a continuous predictor centered at the first trial. As such, the intercept, β_{0i}, reflects log-odds of ER at the first trial. The trial variable for the first (Trials 1-2) and second (Trials 2-4) piece of the curve was denoted as *trial_phase1* and *trial_phase2*, respectively. β_{0i} represents the random intercept for children, and β_{1i} and β_{2i} represent the fixed effects, at the child level, of trial during the first and second trial phase. γ_{00} represents the overall intercept, or the log-odds of ER for the reference group (video-watch). γ_{01} and γ_{02} represent the fixed effect, at the trial level, of modality and the fixed effect of interactivity, respectively, on β_{0i}. U_{0i} added the random effect at the child level to β_{0i}. γ_{10} and γ_{20} denoted the average linear regression slope for phase1 trials and phase2 trials, respectively. The residual in the log-odds of ER at the trial level is denoted by R_{ti}.

To more carefully examine potential underlying mechanisms of the interactivity effect, we examined the types of errors children made as well as the latency of their searches using the same two-level hierarchical random intercept model with trials (Level 1) nested within children (Level 2) due to the same nested data structure. Since preservation errors could only occur in trial 2-4, we did not expect a V-shape pattern across trials. Thus, the model specification was the same with the model for errorless retrieval except that trial was treated as a simple linear predictor for Trials 2-4 instead of two piecewise predictors. Note that we calculated two metrics for perseveration: 1) perseverative errors across all trials for the likelihood of a perseverative error for any trial, and 2) perseverative errors across error trials for the likelihood of an error being perseverative in nature (i.e., given that an error was made, what is the likelihood that the error was perseverative in nature?). In addition to perseveration, we also explored children's

error retrievals by examining self-correction retrievals across conditions. However, the analysis was based on descriptive statistics due to an insufficient amount of self-correction trials for significance testing. Regarding search latency, a linear mixed-effects model, instead of a logistic model, was fit for the continuous dependent variable (time in seconds). Although no prior research has examined the trial effect on search latency, trial was fit as a simple linear predictor based on the visual inspection of a linear trend in the data from our sample.

Lastly, we conducted a series of analyses to examine the associations between children's symbolic retrieval performance in the laboratory, age, and their naturalistic touchscreen media use (e.g., touchscreen exposure, touchscreen use, parental mediation). The unit of analysis was individual children, and the dependent variable was the proportion of errorless retrieval for each child. We created two variables for touchscreen use time, given the positively skewed distributions and the large number of children with no touchscreen use on the previous day. First, we dummy-coded children who had zero touchscreen use and those who had some. Second, for children with at least some touchscreen use, we performed a log transformation to approximate a normal distribution. Correlational and additional linear regression analyses were conducted to test the relations between the proportion of errorless retrieval and the individual and socio-contextual variables.

Results

Preliminary Analyses

Of the 320 trials in the study, there were 304 valid trials and 16 trials in which the children never attempted to retrieve. All the parents completed the survey, revealing that 85% of the children had some exposure to touchscreen devices and 75% had regular exposure. As a randomization check, one-way multivariate analysis of variance (MANOVA) was applied to test

for condition differences with respect to child gender (binary), age (continuous in months), race and ethnicity (categorical), test order (binary), parent education (continuous in years), children's touchscreen exposure and use, and parental mediation. Bonferroni correction was applied to the significance levels to adjust for the likelihood of false positive results. The adjusted significance levels were: $p < .016$; $p < .003$; $p < .0003$, resulted from diving the standard cut-off points for a significant p-value (i.e.., $p < .05$, $p < .01$, $p < .001$) by the number of tests performed (Dunn, 1961). Using Pillai's trace, there was no significant condition difference on gender [$F(3, 76) = 0.10$, $p = .961$], age [$F(3, 74) = 2.27$, $p = 0.088$], race and ethnicity [$F(3, 74) = 0.16$, $p = .920$], test order [$F(3, 74) = 0.03$, $p = .992$], parent education [$F(3, 74) = 1.05$, $p = .374$], touchscreen use time [$F(3, 74) = 1.27$, $p = .289$], and parental mediation [$F(3, 74) = 0.76$, $p = .523$]. That is, child gender, age, race and ethnicity, test order, parent education, touchscreen use time, and parental mediation were evenly distributed across the four conditions. Thus, covariate balance was maintained across the conditions. Although children's age at first touchscreen exposure [$F(3, 74) = 2.97$, $p = .037$] and touchscreen use frequency [$F(3, 74) = 3.88$, $p = .012$] were found to differ across conditions, they were not correlated with children's errorless retrieval performance (see Table 5 in the Individual and Socio-contextual Factors section) and were not considered further.

Errorless retrieval

Overall Proportion of Errorless Retrievals. The proportion of errorless retrieval was calculated for each child combing all four trials. The average proportion of errorless retrievals was greater than would be expected by chance alone in all four conditions (all ps $< .05$). The 2 (interactivity condition: watch vs. point) x 2 (modality condition: window vs. video) two-way analysis of variance (ANOVA) revealed a main effect of modality, $F(1, 76) = 13.51$, $p < .001$,

and a significant interaction between modality and interactivity, $F(1, 76) = 4.00$, $p = .049$. For illustrative purpose, a multiple regression was fitted with interactivity, modality, and their interaction term as predictors to test the marginal effects. As shown in Figure 3, children who watched the video demonstration ($M = .38$, *Standard Deviation* = .24) had significantly fewer errorless retrievals than children who watched the live demonstration through the window ($M = .76$, $SD = .30$), $B = -.38$, $SE = .09$, $p < .001$; however, no significant difference was found between children in the video-point condition ($M = .60$, $SD = .33$) and children in the window-point condition ($M = .71$, $SD = .32$), $B = -0.11$, $SD = .09$, $p = .239$. The significant interaction effect, $B = .27$, $SE = .13$, $p = .049$, suggested differential effects of interactivity as varied by modality (and vice versa). Specifically, the difference between children's window-based and video-based performance under watch conditions ($B = .38$, $SE = .09$, $p < .001$) was different from that under point conditions ($B = .11$, $SE = .09$, $p = .239$), and the difference between children's point-based and watch-based performance under video conditions ($B = .22$, $SE = .10$, $p < .05$) was different from that under window conditions ($B = -.05$, $SE = .10$, $p = .599$).

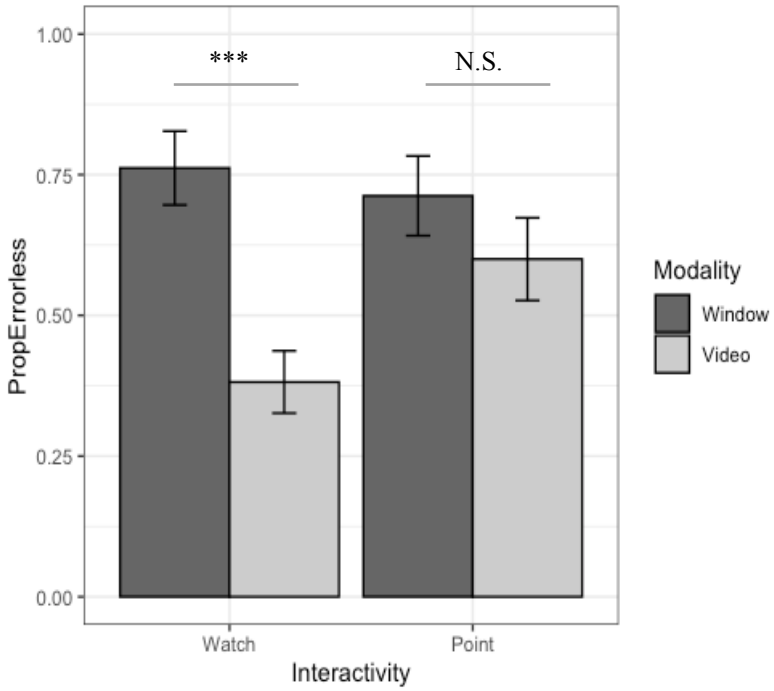

Figure 3. Average proportion of errorless retrievals per child (across all trials) by condition. Significance tests indicated a difference ($p < .001$) between window and video under watch conditions and no difference ($p = .239$) under point conditions.

Probability of Errorless Retrieval by Trial and Condition. The change in errorless retrieval and condition effects across trials was examined more fully in a logistic mixed effects model. A full report of fixed effects from the final model are presented in Table 1, and the predicted probability of errorless retrieval as modeled by this mixed-effects logistic regression is plotted in Figure 4. Modality was a binary variable with video condition coded as the reference group and window condition as the contrast group. Interactivity was also a binary variable with watch condition coded as the reference and point condition as the contrast group. Trial was modeled as two piecewise linear predictors.

Table 1. Fixed effects in the mixed-effects logit model examining condition and trial effect on errorless retrieval

Predictor	*B*	*SE*	Wald z	OR	95% CI (OR)
Intercept (γ_{00})	-0.52	0.46	-1.13	0.59	[0.24, 1.50]
Modality (γ_{01})	2.19	0.60	3.67***	8.90	[2.70, 29.35]
Interactivity (γ_{02})	1.17	0.56	2.08*	3.21	[1.05, 9.86]
Modality × Interactivity (γ_{03})	-1.50	0.81	-1.86†	0.22	[0.04, 1.13]
Trial Phase1 (γ_{10})	-0.70	0.37	-1.88†	0.50	[0.24, 1.05]
Trial Phase2 (γ_{10})	1.27	0.51	2.49*	3.54	[1.28, 9.79]

Note: Modality was a binary variable with video condition coded as the reference group and window condition as the contrast group. Interactivity was a binary variable with watch condition coded as the reference and point condition as the contrast group. The trial variable for the first (trial 1-2) and second (trial 2-4) piece of the curve was denoted as *trial_phase1* and *trial_phase2*, respectively.
$^{\dagger}p < .06$. *$p < .05$. ***$p < .001$

The results were consistent with the ANOVA model of the proportion of errorless retrieval across all trials. There was a significant effect of modality under watch conditions, a significant effect of interactivity under video conditions, and a marginally significant interaction between modality and interactivity. Specifically, children who watched the window demonstration had significantly higher probability of errorless retrieval than children who watched the video demonstration, γ_{01}= 2.19, *SE* = 0.60, Wald *z* = 3.67, $p < .001$, Odd Ratio = 8.90, and children who interacted to see the video demonstration had significantly higher probability of errorless retrieval than children who noninteractively observed the video demonstration, γ_{02} = 1.17, *SE* = 0.56, Wald *z* = 2.08, $p < .05$, OR = 3.21. The marginally significant interaction effect, γ_{03} = -1.50, *SE* = 0.81, Wald *z* = -1.86, $p = .006$, OR = 0.22, suggested differential effects of interactivity as varied by modality (and vice versa). Specifically, the difference between children's window-based and video-based performance under watch conditions (*B* = 2.19, *SE* = 0.60, Wald *z* = 3.67, $p < .001$, OR = 8.90) was different from that under point conditions (*B* = 0.68, *SE* = 0.57, Wald *z* = 1.21, $p = .228$, OR = 1.98); and the

difference between children's point-based and watch-based performance under video conditions (B = 1.17, SE = 0.56, Wald z = 2.08, p < .05, OR = 3.21) was different from that under window conditions (B = -0.34, SE = 0.58, Wald z = 0.58, p = 0.563, OR = 0.71).

As predicted, there was a marginally significant negative trial effect during trial phase1, γ_{10} = -0.70, SE = 0.37, Wald z = -1.88, p = .060, OR = 0.50, and a significant positive trial effect during trial phase2, γ_{20} = 1.27, SE = 0.51, Wald z = 2.49, p < .50, OR = 3.54, suggesting a V-shaped pattern such that the probability of errorless retrieval dropped at the second trial.

An additional model was estimated to examine the potential interaction between trial and condition. However, model comparisons suggested no improvement in model fitness from the original model without the trial-by-condition interaction, LLR $\chi^2(6)$ = 6.25, p = .396, and none of the interactions involving either trial phase was significant. As such, neither of the condition effects changed across trials, nor was the trial effect moderated by experimental conditions.

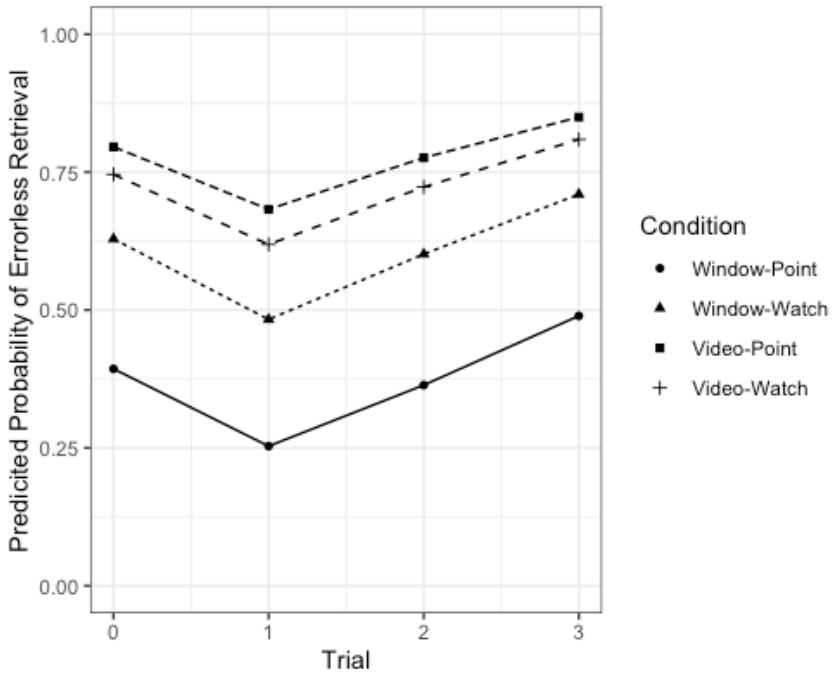

Figure 4. Predicted probability of errorless retrieval as a function of trial and condition

To summarize, children were more successful on the search task (i.e., had more errorless retrievals) after interactive hiding events than non-interactive hiding events. However, this interactivity effect was limited to the video demonstrations. The video-point group scored higher than the video-watch group but did not differ significantly from the window-point group. As predicted, we also observed a drop in errorless retrieval in Trial 2, regardless of condition.

Analyses of Errors

Among the total of 320 trials, there were 194 errorless retrievals and 110 error retrievals, including 16 trials in which the children never attempted to retrieve. A total of 45 error retrievals were due to perseveration (14% of all trials, 41% of error trials), among which over half (i.e., 25 out of 45) occurred on the second trial as shown in Figure 5. Next, a mixed-effects logistic model was fit to examine the likelihood of a perseverative error as a function of trial and condition.

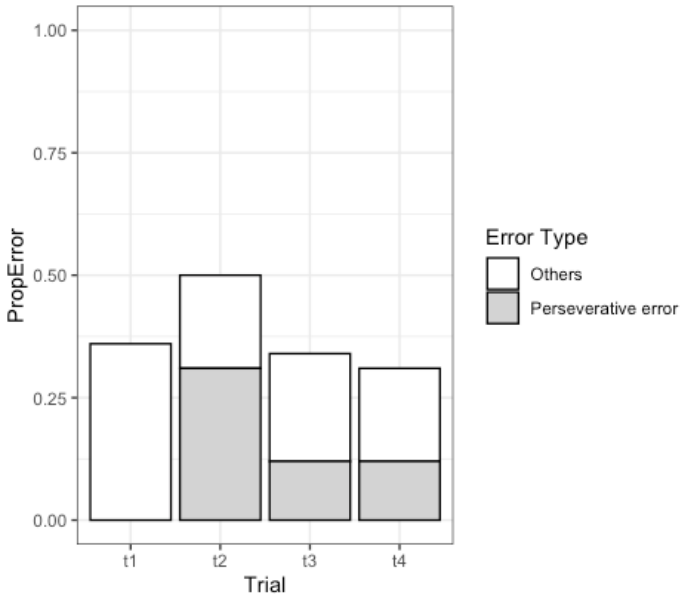

Figure 5. Average proportion of error retrievals per child on each trial by error type. Numbers in the bars represent the proportion for each error type across all children and conditions.

Perseverative Errors across All Trials. The condition effects were largely mirror those for errorless retrieval, but there were some differences. The overall proportions of perseverative errors by condition were graphed in Figure 6. Results from the mixed-effects logistic model were presented in Table 2. We found a significant effect of modality under the watch conditions, a significant effect of interactivity under the window conditions, and a marginally significant interaction between modality and interactivity. Specifically, children who watched the video demonstration had a significantly higher probability of perseveration retrieval than children who watched the window demonstration, γ_{01}= 1.84, *SE* = 0.63, Wald *z* = 2.93, *p* < .01, OR = 6.30, and children who interacted to see the window demonstration had a marginally higher probability of errorless retrieval than children who noninteractively observed the window demonstration, γ_{02} = 1.26, *SE* = 0.65, Wald *z* = 1.94, *p* = .052, OR = 3.52. The marginally significant interaction effect, γ_{03} = 1.43, *SE* = 0.79, Wald *z* = 1.81, *p* = .070, OR = 4.18, suggested differential effects of interactivity as varied by modality (and vice versa). Specifically, the difference between children's video-based and window-based performance under watch conditions (*B* = 1.84, *SE* = 0.63, Wald *z* = 2.93, *p* < .01, OR = 6.30) was different from that under point conditions (*B* = 0.41, *SE* = 0.49, Wald *z* = 0.84, *p* = .401, OR = 1.51), and the difference between children's point-based and watch-based performance under video conditions (*B* = -0.17, *SE* = 0.45, Wald *z* = -0.38, *p* = .703, OR = 0.84) was different from that under window conditions (*B* = 1.26, *SE* = 0.65, Wald *z* = 1.94, *p* = .052, OR = 3.52.).

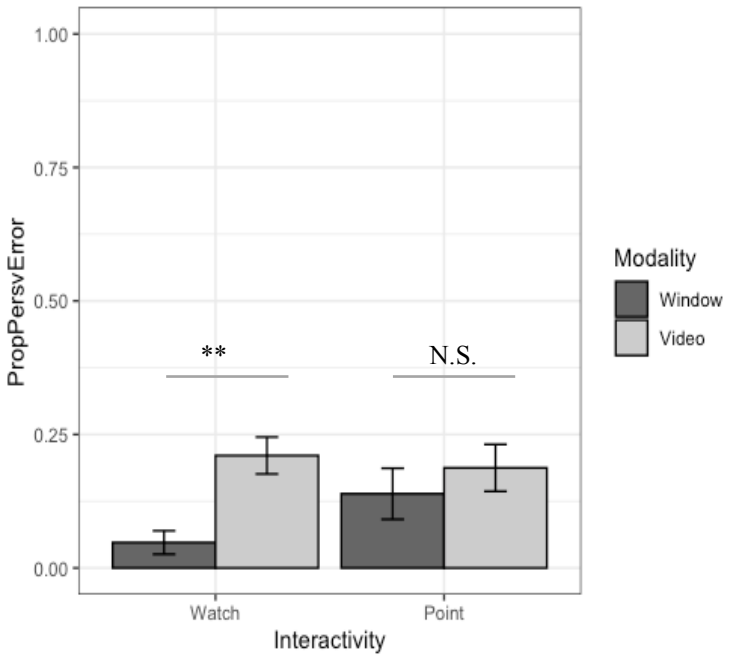

Figure 6. Average proportion of perseveration retrievals per child (across all trials) by condition. Significance tests indicated a difference ($p < .01$) between window and video under watch conditions and no difference ($p = .401$) under point conditions.

Table 2. Fixed effects in the mixed-effects logit model examining condition and trial effect on perseverative retrieval

Predictor	*B*	*SE*	Wald z	OR	95% CI (OR)
Intercept (γ_{00})	-0.34	0.37	-0.92	0.71	[0.34, 1.49]
Modality (γ_{01})	-1.84	0.63	-2.93**	0.16	[0.05, 0.56]
Interactivity (γ_{02})	-0.17	0.45	-0.38	0.84	[0.34, 2.07]
Modality x Interactivity (γ_{03})	1.43	0.79	1.81†	4.18	[0.86, 20.31]
Trial (γ_{10})	-0.69	0.23	-2.99**	0.50	[0.31, 0.80]

Note: Trial included trials 2 to 4 and was centered at Trial 2. Modality was a binary variable with video condition coded as the reference group and window condition as the contrast group. Interactivity was a binary variable with watch condition coded as the reference and point condition as the contrast group.
$^{\dagger}p < .07$. *$p < .05$. **$p < .01$

Perseverative Errors across Error Trials. Since the more error retrievals were made on a trial, the more possibility that children would make perseverative errors (and other errors) on

this trial. To take this into consideration, we analyzed the perseverative errors across error (not all) trials to examine, given that an error was made, the likelihood that the error was perseverative. As illustrated in Figure 7, this likelihood was 63% in the video-point condition, 45% in the video-watch condition, 64% in the window-point condition, and 27% in the window-watch condition. Thus, except of the window-watch condition, the errors made by the children were well above the chance level to be perseverative errors.

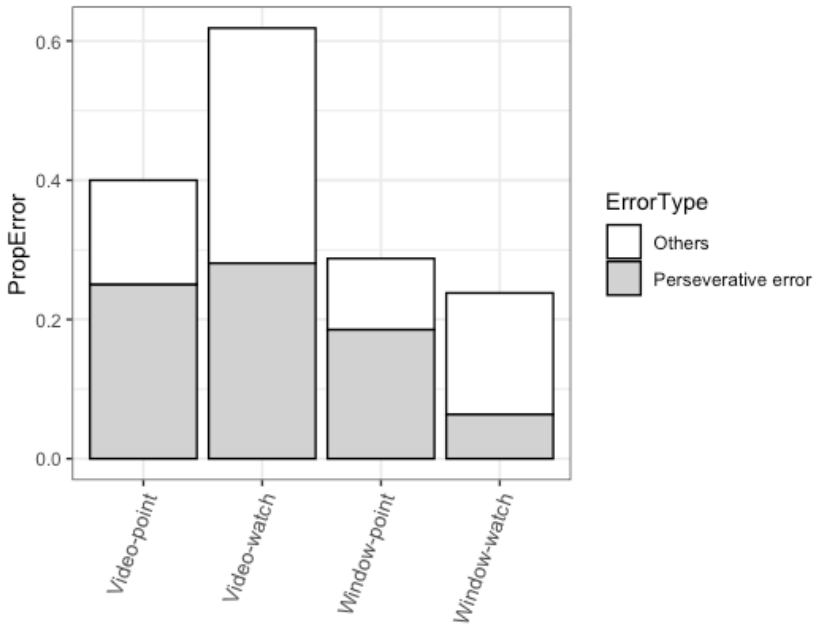

Figure 7. Proportion of perseveration retrievals (across all error trials) by condition.

As Table 3 shows, we found a significant main effect, $\gamma_{02} = 0.91$, $SE = 0.45$, Wald $z = 2.03$, $p < .05$, OR = 2.49. Specifically, for children who interacted to see the demonstration, once an error was made, it was more likely to be a perseverative error as compared to children in the watch conditions. Consistent with the results for perseverative errors across all trials, there was a negative trial effect by which the likelihood of an error being perseverative decreased through trial 2 to 4, $\gamma_{10} = -0.54$, $SE = 0.27$, Wald $z = -2.99$, $p < .05$, OR = 0.58.

Table 3. Fixed effects in the mixed-effects logit model examining condition and trial effect on perseverative retrieval

Predictor	*B*	*SE*	Wald z	OR	95% CI (OR)
Intercept (γ_{00})	0.19	0.38	0.50	1.21	[0.56, 2.61]
Modality (γ_{01})	-0.54	0.46	-1.16	0.58	[0.23, 1.48]
Interactivity (γ_{02})	0.91	0.45	2.03*	2.49	[1.01, 6.12]
Trial (γ_{10})	-0.54	0.27	-2.99**	0.58	[0.34, 1.01]

Note: Trial included trials 2 to 4 and was centered at Trial 2. Modality was a binary variable with video condition coded as the reference group and window condition as the contrast group. Interactivity was a binary variable with watch condition coded as the reference and point condition as the contrast group.
*$p < .05$.

Self-correction of Perseverative Errors. In order to better understand the perseverative errors, we also analyzed the frequency of self-correction for perseverative retrievals, in which children spontaneously corrected themselves on the second attempt after a perseverative retrieval on the first attempt. There was a total of 16 self-correction retrievals among the 45 perseverative trials (36%): 19% of these self-correction retrievals were in window-watch condition, 38% were in the video-watch condition, 25% were in the window-point condition, and 19% were in the video-point condition.

To summarize, across all retrievals, children made more perseverative errors after watched the video demonstration than the window demonstration, but no difference between watching and touching the video demonstration. However, once an error was made, children who interacted to see the demonstration were more likely to make a perseverative error as compared to children who watched to see the demonstration. This is because the point group made fewer errors in total and similar amount of perseverative errors as compared to the watch group.

Search Latency

The average search latency of all retrievals was 13.36 seconds (*SD* = 17.85, range = 1 to 130). Figure 8 shows latency (log-transformed seconds) across conditions for each type of retrievals. The linear mixed-effects model showed a significant effect of trial was found by which children were faster to retrieve the toy on later trials, $\gamma_{10} = -0.29$, $SE = 0.03$, $p < .001$. There was a marginally significant main effect of interactivity, with longer latency under watch conditions compared to the point conditions, $\gamma_{01} = 0.28$, $SE = 0.16$, $p = .080$. However, we did not find an effect for modality, $\gamma_{02} = -0.18$, $SE = 0.16$, $p = .268$.

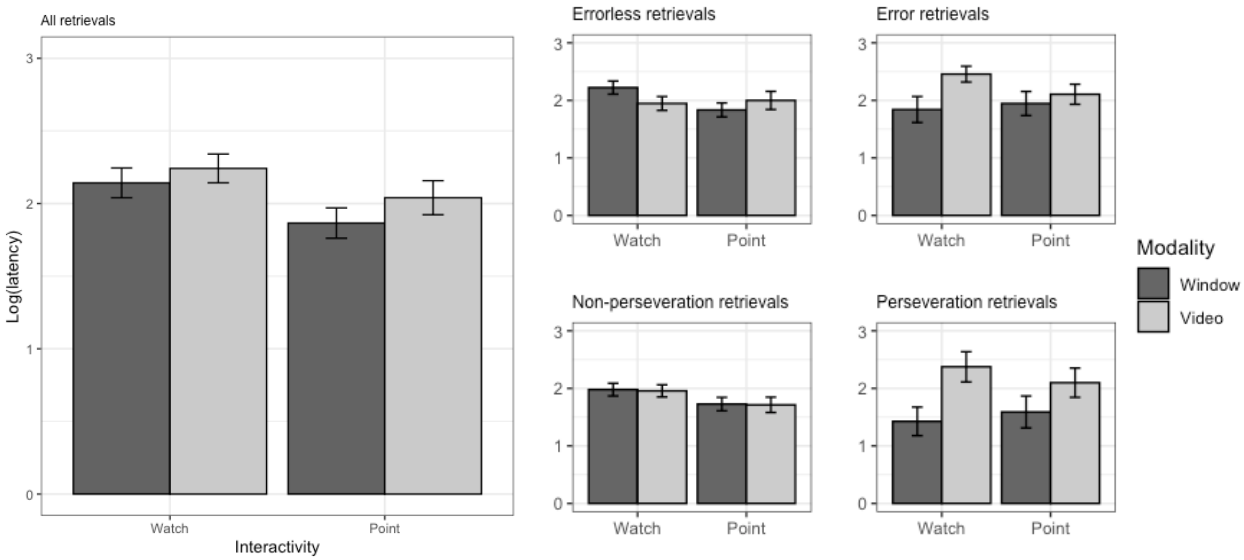

Figure 8. Latency for different types of retrievals: all retrievals, errorless retrievals, error retrievals, non-perseveration retrievals, and preservation retrievals, across different conditions. Latency was the log-transformed seconds.

Sub-sample analyses were conducted to examine the latency in errorless (Model 1) and error (Model 2) retrievals. For trials in which children successfully retrieved the toy on the first attempt, the average search latency was 12.60 seconds (*SD* = 17.86, range = 1 to 130). For trials on which children searched an incorrect location on the first attempt, the average search latency

was 14.71 seconds (*SD* = 17.84, range = 2 to 96.5). As shown in Table 4, consistent results were found for a negative trial effect in both errorless and error retrievals, whereas no effect was found for modality or interactivity.

Table 4. Fixed effects in the mixed-effects logit model examining condition and trial effect on search latency for errorless retrievals and error retrievals

Predictor	***Model 1***			***Model 2***		
	B	*SE*	t	*B*	*SE*	t
Intercept (γ_{00})	2.85	0.21	13.82***	3.07	0.24	12.90***
Modality (γ_{01})	0	0.18	0.01	-0.24	0.26	-0.95
Interactivity (γ_{02})	-0.17	0.17	-0.99	-0.13	0.25	-0.54
Trial (γ_{10})	-0.27	0.05	-5.90***	-0.29	0.06	-4.75***

Note: Trial was centered at Trial 1. Modality was a binary variable with video condition coded as the reference group and window condition as the contrast group. Interactivity was a binary variable with watch condition coded as the reference and point condition as the contrast group.
****p* < .01*

To summarize, children across all conditions show decreasing latency in retrieving the hiding object. While generally children who watched to see the demonstration spent longer time than those who touched to see, no significant difference across conditions was detected for any particular types of retrievals.

Individual and Socio-contextual Factors

The parent survey included information about the child's touchscreen exposure and use at home. Of the 75% of children who had been regularly exposed to touchscreen devices, 40% started using a touchscreen device at 18 months or younger, 48% began between 19 and 29 months, and 12% began after 30 months. The average age when these children began to have regular touchscreen exposure was 15 months (*SD* = 10.96, range = 0 to 34). Eighty-five percent of children had at least some exposure to touchscreen devices, among which 84% used a

touchscreen device at least once a week, and 44% used one at least once a day. On a typical day, the children were exposed to an average of 23 min (*SD* = 39.22, range = 0 to 240) of touchscreens. Among children who had at least some exposure to touchscreen, 44% of the parents never helped their children with touchscreen use, 30% often helped, and 22% always helped.

A series of analyses were conducted to examine the associations among child age, children's symbolic retrieval performance in the laboratory, and their touchscreen experience at home. The unit of analysis was individual children, and the dependent variable was the proportion of errorless retrieval for each child. Correlations between the key variables are presented in Table 5. Subsequent linear regression indicated that only age significantly predicted children's retrieval performance, $B = 0.04$, $SE = 0.02$, $p < .05$. For children who had been exposed to at least some touchscreens, the age of initial use was positively correlated with children's errorless retrieval, $B = 0.02$, $SE = 0.01$, $p < .05$; however, it became nonsignificant when current age was controlled, $B = 0.01$, $SE = 0.01$, $p = .108$. For the dummy variable (i.e., touchscreen use) not presented in Table 5, no difference was found between children who zero touchscreen use and those who had some, $B = -0.01$, $SE = 0.07$, $p = .856$. Note that this study was not designed to test age as a moderating factor, and it is not adequately powered to do so. However, exploratory analyses examining potential age moderation are presented in Appendix.

Table 5. Bivariate correlation between symbolic retrieval and individual and social-contextual factors

	Mean	*SD*	1	2	3	4	5	6
1. Errorless retrieval	0.61	0.32	--					
2. Child age (months)	29.81	3.93	0.32**	--				
3. Age_TS (months)	15.82	10.96	0.10	0.30**	--			
4. TS use frequency	3.59	2.07	0.16	0.10	0.61***	--		
5. TS use time (minutes)	3.23	0.98	-0.02	-0.03	0.18	0.46**	--	
6. Parental mediation	2.33	1.23	-0.10	-0.11	0.30**	0.33**	-0.30*	--

Note. Age_TS = the age when children began to regularly use touchscreen devices. TS use frequency refers to how often the child uses a touchscreen device on average. TS use time was the log-transformed minutes for the minutes of touchscreen use on the previous day.

Discussion

Decades of research have established that it is cognitively challenging for young children to use information from symbolic artifacts and media to solve problems in real life, as replicated by our findings based on video symbols. The present study sought to understand the role of interactivity in toddlers' symbolic transfer from video symbols, with a particular interest in the underlying mechanisms and comparison to live, unmediated conditions. A secondary goal considered the relation between symbolic transfer and child characteristics (e.g., child age and naturalistic interactive media experience). Findings from the current study replicate and extend those from past research. In this discussion, we consider our findings in the context of prior research on toddlers' use of video in object-retrieval tasks as well as implications for future research.

Can Interactivity Simulate Real-life Experience?

Our first purpose was to determine the degree to which interactivity (i.e., contingency of responding to the child) with video improved toddlers' object retrieval. We replicated the previous finding that interactivity facilitated 30- to 36-month-olds' use of information from video to retrieve a toy hidden in an adjacent room (Lauricella et al., 2010), and we extended this

finding to younger children under 30 months. Notably, in Lauricella et al. (2010) children viewed the hiding event six times in the video conditions; however, we merely showed the demonstration once. Moreover, as the first study examining the impact of interactivity on search latency, we found that children spent slightly less time under the interactive conditions across all retrievals, suggesting a potential facilitative effect of interactivity on not only the performance accuracy but also the speed. Note that, since this difference between point and watch group is marginally significant, it needs to be examined more thoroughly in future research.

Importantly, both the current study and the Lauricella et al. (2010) study found no significant difference in search performance between the video-interactive condition and window conditions, even though quantitively the video-point group had fewer errorless retrievals, more perseveration errors, and longer search latency as compared to the window groups. These results indicate the potential of contingency-based interactivity in improving symbolic transfer to the level (albeit not perfectly) of performance achieved in the non-symbolic situation, even in the absence of interpersonal interactions (e.g., true social interactions with a live social partner in person or via video chat).

While our findings suggest the possibility of overcoming (not just alleviating) the learning disparities between in-person and video-mediated learning, there is emerging evidence that contingency may not be sufficient to support toddlers' learning from video. For example, 24- and 30-month-olds were found to successfully learn a novel word from live in-person demonstration, but not live video chat interaction in which the speaker engaged the child with direct eye gaze and actions contingent on the child's behavior (e.g., pausing if the child became distracted) (Troseth et al., 2018). Similarly, Tsuji and colleagues (2021) showed that, while 16-month-olds can learn a novel word-object association from the in-person demonstration, they did

not learn from the speaker through video chat or from the virtual agent, both of which provided contingent reactivity to the children's eye gaze. Besides the many differences between object-retrieval and word-learning tasks (Strouse & Samson, 2021), one possible explanation for the discrepant findings is that contingency in some paradigms (i.e., joint engagement through followed gazed or reactive actions of the speaker) might not be as efficient as contingency that seeks physical interactivity from the children, such as touching the screen in our study or pressing a key on the keyboard in Lauricella et al. (2010). Perhaps young children need the extra prompt of executing some action to better direct their attention or engage. In addition to a comprehensive investigation of interactivity effects across different task domains, future research could use eye-tracking methods to compare the influence of the two contingency features on children's selective attention while watching the on-screen demonstrations.

Furthermore, our findings extended the positive interactivity effect in 30- to 36-month-olds, as found in Lauricella et al. (2016), to younger toddlers aged 24 to 30 months. However, this is inconsistent with prior research using the same interactivity interface (i.e., tapping the toy on the screen to watch the hiding event). It is worthwhile to note that, while Choi and Kirkorian (2016) reported an effect of interactivity decreasing with age (from 24 to 34 months) by which its benefit was greatest among youngest children and disappeared among older children, our findings did not reveal a moderating effect of age (e.g., exploratory analysis of age in Appendix). According to Choi and Kirkorian, their task might be particularly challenging to the youngest children and interactivity thus provided efficient support, whereas older children were capable of this task even without the interactive features. Given the crucial role of task difficulty (Kirkorian, 2018; Strouse & Samson, 2021) and child age and working memory (Choi et al., 2018, 2021) in video-mediated performance, the conflicting findings in older children may be due to the object-

retrieval tasks used in the two studies. Specifically, the 2D task in Choi and Kirkorian (2016) using a felt board as the searching space and a paper sticker as the target object were of the identical size and dimension of the tested space and object; whereas our 3D task using a real room and real toy could be much more difficult, because research has demonstrated that perceptual mismatch in size and dimension hinders toddlers' transfer in object-retrieval tasks (Choi et al., 2016; DeLoache et al., 1991; Schmidt & Anderson, 2002). This difficulty could also be evident by the overall lower errorless retrieval proportion in our study as compared to Choi and Kirkorian (2016). In this sense, the support of interactivity perhaps protracted into older children due to the relatively difficult task in the current study. Our findings, together with those from preschool-aged children that interacting with video might be less beneficial than merely watching when the task is cognitively demanding (Alade et al., 2016; Schroeder & Kirkorian, 2016), suggest that interactivity is likely to be most useful when a task is just above children's capability and is ineffective or even harmful if the task is too easy or too difficult. There are similar hypotheses derived from more general learning theories and specific empirical research. For example, optimal learning occurs within the zone of proximal development (Vygotshky,1978), and children learn best when the Amount of Mental Effort (Salomon, 1983) invested to comprehend the media is just right (Tiwari, 2020). In this sense, it is essential for media producers to design the interactive interface while considering the difficulty of media content as well as child age and abilities.

Mechanism(s) underlying Interactivity Effect

The second purpose of this study was to examine potential mechanisms underlying the interactivity effect, particularly as they relate to symbolic insight, attention/encoding, and engagement/arousal. We found that, while interactivity improved video-mediated retrieval

performance, it had little impact on in-person retrieval. Children who observed the live experimenter hiding the toy searched as well as, even quantitatively better than, those who interacted to see the live hiding event.

First of all, our finding of the positive interactivity effect under video conditions did not support the negative hypothesis under the *symbolic insight* mechanism that interactivity hinders toddler's understanding of the symbolic relation between video symbols and their referents. Theory on symbolic development (Troseth et al., 2019) and empirical evidence from research using scale rooms and picture books suggest that increasing the symbol's salience as an object might make it more difficult to appreciate the symbolic relation (DeLoache, 2000; Sheehan & Uttal, 2016; Tare et al., 2010). However, our results, and those of others (Lauricella et al., 2010; Kirkorian et al., 2016), indicate this may not be the case for digital symbols, at least for the interactivity of touching the screen during the activity as in the present study. This might be related to children's perceptions of and beliefs about digital devices. Unlike scale models, video-mediated symbols have a variety of forms and functions, ranging from informative media to entertainment tools (Troseth et al., 2019). Children's perception of a video device might have developed from their daily media exposure and use and is thus less likely to be influenced by the interactive interface at the moment during a laboratory activity. It is possible that the interactive feature in the current study was not sufficient physical manipulation to change children's symbolic insight, as compared to a separate session of longer exposure before the task (e.g., DeLoache, 2000) or more complex and intensive interactivity with multiple steps and actions (e.g., Alade et al., 2016; Schroeder & Kirkorian, 2016). This might explain the negative interactivity effects reported by previous studies (e.g., DeLoache, 2000; Alade et al., 2016;

Schroeder & Kirkorian, 2016), but future research is needed to experimentally test the impact of exposure to interactivity on children's symbolic insight.

Secondly, other scholars have suggested that interactivity may increase transfer by increasing children arousal and engagement (Kuhl, 2003, 2007; Beihler & Snowman, 1997; Kirkorian et al., 2016). In this case, we would have expected a similar effect under window and video conditions. However, we found the facilitative effect of our interactive feature only in the video condition, suggesting that interactivity may not work by increasing children's arousal and engagement. There was minimized social contingency (e.g., general interpersonal social cues) in our window conditions. In other words, the window condition in the current study was semi-in-person, as it provided social relevance afforded by human presence but no social contingency afforded by two-way exchange between children and the experimenter as defined by Troseth et al. (2006) and seen in other studies (e.g., Troseth et al., 2018; Tsuji et al., 2021). Given the well-documented positive effect of contingency on learning even in the in-person situation (Bloom et al., 1987; Goldstein et al., 2003, 2009), interactivity should have benefited the window conditions if it effectively simulated the social contingency. However, this is not supported by our finding of no interactivity effect in window conditions, suggesting that the contingency created by our interactive feature may not increase children's arousal and engagement. It seems doubtful the the absence of an interactivity effect in the window condition is was due to a ceiling effect, because the symbolic retrieval task is believed to be a hard task for toddlers (Schmitt & Anderson, 2002), and the performance of children in our sample was far from perfect.

Moreover, our finding for perseveration provides further evidence that the contingency created by our interactive interface may not increase children's arousal and engagement. As reviewed earlier, increased arousal and engagement could strengthen mental representation of the

hiding event in current trial to make it more competitive over that of the outdated previous trial, thereby reducing perseveration errors (Choi & Kirkorian, 2016). However, we found that interactivity benefitted trial 1 equally well as subsequent trials and had no impact on perseverative errors. Taken together, our study, with others (Choi & Kirkorian, 2016; Troseth et al., 2018; Tsuji et al., 2021), provides evidence for the ineffectiveness of contingency that lacks interpersonal social cues. In other words, human presence might be necessary to provoke arousal and engagement and is unlikely to be substituted by on-screen interactivity. This might be due to the mechanism of relevance created by human presence, by which children dismiss on-screen information as being real and relevant to their immediate surroundings (Henderson et al., 2013; Wilson & Sperber, 2004). Future research could use physiological measures such as heart rate to directly assess arousal under different conditions.

How would interactivity enhance the symbolic retrieval if it did not provoke children's arousal and engagement? A possibility is that interactivity facilitated selective attention to and encoding of important information on the screen, as suggested by previous research (Choi & Kirkorian, 2016; Kuhl, 2007). Given that toddlers typically attend to and encode information inefficiently from video that is perceptually and socially impoverished relative to real-life experience (Carver et al., 2006; Kirkorian et al., 2016), interactivity would have greater benefit to video conditions as compared to window conditions. Consistently, the current study found the interaction between interactivity effect and modality by which interactivity was only effective under the video, but not the window, conditions. Moreover, recent eye-tracking evidence demonstrates that the interactive feature successfully directed toddlers' attention such that they looked more at the target locations (Kirkorian et al., in revision).

Another possibility concerns children's symbolic insight. As reviewed earlier, interactivity is theorized to enhance symbolic performance by facilitating symbolic insight or the direct mapping between video symbols and their referents in reality through increasing similarity between them (Troseth et al., 2019; Sheehan & Uttal, 2016). We doubt that children in this study were fooled by the contingency created by the interactive feature and believed what they saw on video was real even happing in front of them. Instead, it is possible that, rather than hindering symbolic insight, interactivity helped children linking video depiction to reality by clarifying the temporal relevance of the hiding event on video. This possibility is supported by our finding of the interaction between interactivity effect and modality, since effects of interactivity on symbolic insight should only apply to the symbolic, video conditions rather than the non-symbolic, window conditions. Nonetheless, our findings that interactivity did not alleviate perseveration provide no support to this symbolic insight hypothesis. This is because, if interactivity facilitated symbolic insight, children would be better at using video to retrieve in the room and their mental representation of video in the current trial would thereby be less likely to be interfered by that of the live retrieving experience in the previous trial. In this sense, the *attending and encoding hypothesis* is a more plausible explanation.

Child Characteristics at Socio-contextual Level

A final goal of the current study was to identify potential correlates of toddlers' symbolic transfer at the child and context level. We found that children's retrieval performance was predicted by their age as predicted but not associated with their naturalistic experience with touchscreens at the social-contextual level. While age has been consistently reported as a predictor of video-mediated learning, it is typically a proxy for many aspects of cognitive competence that are particularly important to video-mediated symbolic transfer. For instance,

working memory emerges before 2 years of age (Hughes, 1998; Carlson, 2005; Carlson et al., 2013) and develops with age until early adolescence (Carlson et al., 2013; Luna et al., 2004), and research has suggested that working memory resources could ameliorate the difficulty with video-mediated symbolic search by reducing the cognitive burden from perceptual mismatch (Barr, 2010; Troseth et al., 2010) and perseveration (Troseth, 2010). As discussed earlier, the symbolic retrieval task requires children to maintain dual representation and update outdated representation of previous trials, which taxes mental resources and is likely to be supported by working memory (Choi & Kirkorian, 2016; Hartstein & Berthier, 2018; Sheehan et al., 2020). Consistently, our finding shows that child age predicted a decrease in perseveration errors and shorter search latency.

Contrary to our prediction, no association was found between the children's retrieval performance and their naturalistic touchscreen exposure and use. Considering previous findings that toddlers' video-mediated symbolic retrieval was predicted by their naturalistic use of interactive media (Kirkorian & Choi, 2017) and live video (Troseth et al., 2010), it is possible that only symbolic use of touchscreens actually helps children experience and practice with the symbolic function of touchscreens, whereas the overall use time and exposure as measured in the current study are not representative of children's symbolic experience with touchscreens. Given the various functions afforded by touchscreen devices, children might differ markedly in the interactive (or noninteractive) experience gained from using a touchscreen device. Prior work suggests that toddlers' naturalistic experience with interactive (but not noninteractive) media predicted their video-mediate retrieval in the laboratory.

It is worthwhile to note that parental mediation was also found to not associate with children's symbolic retrieval performance. While the scaffolding role of parent engagement has

been well documented in the literature (Friedrich & Stein, 1975; Reiser et al., 1984; Strouse et al, 2013), no evidence suggests that such a scaffolding could transfer to other tasks and in the absence of parent engagement. Our finding indicated that toddlers' symbolic transfer in laboratory tasks might not be influenced by parental mediation at home. While there is no study directly testing a delayed effect of parental mediation on children's video-mediate learning in laboratory, intervention research suggests that the facilitative effects of parental mediation may help 3-year-olds' story comprehension and vocabulary after repeated exposures of well-designed active mediation (e.g., dialogical questioning) over 4 weeks (Strouse et al., 2013). We did not collect information about the type or quality of parents' active mediation, and future research should include more comprehensive measures of parents' active mediation to identify specific practices associated with more generalized improvements in children's learning from screen media.

Limitations and Future Directions

This study extends prior research on video-mediated symbolic transfer and interactivity effect using a carefully controlled experimental design. However, there are some limitations that inform our interpretations and suggestions for future research. A primary limitation to this research is the relatively small sample size, which limits the statistical power for more complicated analyses. We addressed this limitation to the extent possible by conducting additional analyses (e.g., model fitness comparison, randomization check) and using plots for visual evaluation (see Appendix for exploratory analyses). However, we were unable to include child age in the model for experimental condition effects and to statistically test the extent to which age moderates the interactivity effect, modality effects, or their interaction. Secondly, future research should address the homogeneity of this convenience sample to broaden

participation in research and establish the generalizability of the findings to other populations. Another limitation regards the window-point condition. In an attempt to match the procedure and script as much as possible to the video-point condition for higher internal validity, our window-point condition (i.e., pointing to an experimenter through a window) likely deviated from what children would have expected from real-life interactions, resulting in lower external validity. Future research should consider ways to study more naturalistic/active interactive learning from screens while maintaining experimental control. Lastly, regarding the parent survey for naturalistic media experience, future research would benefit from questions about specific media content and specific functions used by children for a certain device and specific parent media behaviors that are likely to affect transfer (i.e., to identify specific media experiences that are likely to foster symbolic insight).

Conclusion

The current study builds on the extant literature by replicating that interactivity enhances symbolic transfer from video to real life in toddlers and, more importantly, by demonstrating that the contingency provided by physical interactivity, in the absence of interpersonal social cues, might even aid younger toddlers to achieve the level of performance as an in-person experience. Moreover, our findings shed light on potential mechanisms underlying the interactivity effect by comparing the impacts of interactivity between symbolic and non-symbolic situations. They illuminate that, while physical interactivity based on contingent actions might facilitate toddlers' symbolic transfer through increased selective attention to and encoding of target information or direct mapping between video and reality, it may lack the social contingency that would otherwise provoke arousal and engagement as in in-person interactions.

It is important to note that, even though the current study aims at separating and disentangling potential mechanisms to identify the predominate one(s), they are less likely to be mutually exclusive than complementary to each other. And the effects of interactivity may vary across tasks and by child characteristics. The present study highlights the potential of interactive media in supporting young children's symbolic use of video and even to a degree that is close to real-life experience, but more research is necessary to understand the nuanced relation between interactive features, task characteristics (e.g., content domain and task difficulty), and child characteristics (e.g., age and working memory).

References

Aladé, F., Lauricella, A. R., Beaudoin-Ryan, L., & Wartella, E. (2016). Measuring with Murray: Touchscreen technology and preschoolers' STEM learning. *Computers in human behavior*, *62*, 433-441.

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review*, *30*(2), 128-154.

Barr, R., Moser, A., Rusnak, S., Zimmermann, L., Dickerson, K., Lee, H., & Gerhardstein, P. (2016). The impact of memory load and perceptual cues on puzzle learning by 24-month-olds. *Developmental psychobiology*, *58*(7), 817-828.

Beihler, R. F., & Snowman, J. (1997). Psychology applied to education. (8th ed). Boston: Houghton Mifflin Company.

Bloom, P. (2000). How children learn the meanings of words. Cambridge, MA: MIT Press. Borovsky.

Bloom, K., Russell, A., & Wassenberg, K. (1987). Turn taking affects the quality of infant vocalizations. *Journal of Child Language, 14*, 211–227. doi:10.1017/S030500 0900012897.

Callaghan, T. C. (1999). Early understanding and production of graphic symbols. *Child Development*, 70, 1314–1324. https://doi.org/10.1111/1467-8624.00096.

Carlson, S. M., Wong, A., Lemke, M., & Cosser, C. (2005). Gesture as a window on children's beginning understanding of false belief. *Child Development*, 76, 73-86

Carlson, S. M., Zelazo, P. D., & Faja, S. (2013). Executive function. In P. D. Zelazo (Ed.), The Oxford handbook of developmental psychology: Body and mind (Vol. 1, pp. 706–743). New York, NY: Oxford University Press.

Callaghan, T. C., Rochat, P., MacGillivray, T., & MacLellan, C. (2003). The social construction of pictorial symbols in 6-to 18-month-old infants. *Unpublished manuscript, St. Francis Xavier University*.

Callaghan, T., Rochat, P., MacGillivray, T., & MacLellan, C. (2004). Modeling referential actions in 6- to 18-month-old infants: A precursor to symbolic understanding. *Child Development*, 75(6), 1733–1744. https://doi.org/10.1111/j.1467-8624.2004.00813.x.

Carver, L. J., Meltzoff, A. N., & Dawson, G. (2006). Event-related potential (ERP) indices of infants' recognition of familiar and unfamiliar objects in two and three dimensions. *Developmental Science, 9*, 51–62.

Choi, K., & Kirkorian, H. L. (2016). Touch or watch to learn? Toddlers' object retrieval using contingent and noncontingent video. *Psychological science*, *27*(5), 726-736.

Choi, K., Kirkorian, H. L., & Pempek, T. A. (2018). Understanding the transfer deficit: Contextual mismatch, proactive interference, and working memory affect toddlers' video-based transfer. *Child development*, *89*(4), 1378-1393.

Choi, K., Kirkorian, H. L., & Pempek, T. A. (2021). Touchscreens for Whom? Working Memory and Age Moderate the Impact of Contingency on Toddlers' Transfer From Video. *Frontiers in Psychology*, *12*.

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science, 238*, 1556–1557. https://doi.org/10.1126/science.2446392.

DeLoache, J. S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development*, 62, 736–752. https://doi.org/10.2307/1131174.

DeLoache, J. S., & Burns, N. M. (1994). Early understanding of the representational function of pictures. Cognition, 52,83–110. https://doi.org/10.1016/0010-0277(94)90063-9.

DeLoache, J. S. (1995). Early symbol understanding and use. In D. L. Medin (Ed.), Vol. 33. The psychology of learning and motivation: Advances in research and theory (pp. 65–114). San Diego, CA: Academic Press.

DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengren, K. S., & Gottlieb, A. (1998). Grasping the nature of pictures. *Psychological Science*, 9, 205–210. https://doi.org/10.1111/1467-9280.00039

DeLoache, J. S., Peralta, O. A., & Anderson, K. (1999). Multiple factors in early symbol use: Instructions, similarity, and age in understanding a symbol–referent relation. *Cognitive Development, 14*, 299–312. https://doi.org/10.1016/S0885-2014(99)00006-4.

DeLoache, J. S. (2000). Dual representational and young children's use of scale models. *Child Development, 71*(2), 329–338. https://doi.org/10.1111/1467-8624.00148

DeLoache, J. S. (2002). The symbol-mindedness of young children. In W. W. Hartup & R. A. Weinberg (Eds.), Vol. 32. Child psychology in retrospect and prospect: The Minnesota symposia on child psychology (pp. 73–101): LEA Publishing.

DeLoache, J. S., & Chiong, C. (2010). Babies and baby media. *American Behavioral Scientist*, *52*(8), 1115-1135.

Deocampo, J. A., & Hudson, J. A. (2005). When seeing is not believing: Two-year-olds' use of video representations to find a hidden toy. *Journal of Cognition and Development*, *6*(2), 229-258.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience*, *22*(1), 567-631.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, *56*(293), 52-64.

Flavell, J. H., Flavell, E. R., Green, F. L., & Korfmacher, J. E. (1990). Do young children think of television images as pictures or real objects? Journal of Broadcasting and Electronic Media, 34, 399–419. https://doi.org/10.1080/08838159009386752.

Friedrich, L. K., & Stein, A. H. (1975). Prosocial television and young children: The effects of verbal labeling and role playing on learning and behavior. *Child Development*, 27-38.

Gergely, G., Egyed, K., and Király, I. (2007). On pedagogy. Dev. Sci. 10, 139–146. doi: 10.1111/j.1467-7687.2007.00576.x.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 8030– 8035. doi:10.1073/pnas.1332441100

Goldstein, M. H., Schwade, J. A., & Bornstein, M. H. (2009). The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child Development, 80*, 636–644. doi:10. 1111/j.1467-8624.2009.01287.x

Hatano, G., & Inagaki, K. (2002). Domain-specific constraints of conceptual development. In W. W. Hartup & R. K. Silbereisen (Eds.), Growing points in developmental science: An introduction (pp. 123–142). New York: Psychology Press.

Hartstein, L. E., & Berthier, N. E. (2018). Transition to success on the model room task: the importance of improvements in working memory. *Developmental science*, *21*(2), e12538.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements* (pp. 537-III). Elsevier.

Hughes, C., & Ensor, R. (2005). Executive function and theory of mind in 2 year olds: A family affair? Develop- mental Neuropsychology, 28, 645–668. doi:10.1207/ s15326942dn2802_5

Jenkins, I. L., & Berthier, N. E. (2014). Working memory and inhibitory control in visually guided manual search in toddlers. *Developmental psychobiology*, *56*(6), 1252-1262.

Klima, E. S., & Bellugi, U. (1979). The signs of language. Cambridge, MA: Harvard University Press.

Kirkorian, H. L. (2018). When and how do interactive digital media help children connect what they see on and off the screen?. *Child Development Perspectives*, *12*(3), 210-214.

Kirkorian, H. L., Anderson, D. R., & Keen, R. (2012). Age differences in online processing of video: An eye movement study. *Child development*, *83*(2), 497-507.

Kirkorian, H. L., Wartella, E. A., & Anderson, D. R. (2008). Media and young children's learning. *The Future of children*, 39-61.

Kirkorian, H. L., & Anderson, D. R. (2018). Effect of sequential video shot comprehensibility on attentional synchrony: A comparison of children and adults. Proceedings of the National Academy of Sciences, 115, 9867–9874.

Kirkorian, H. L., Choi, K., & Pempek, T. A. (2016). Toddlers' word learning from contingent and noncontingent video on touch screens. *Child development*, *87*(2), 405-413

Kirkorian, H. L., & Choi, K. (2017). Associations between toddlers' naturalistic media experience and observed learning from screens. *Infancy*, *22*(2), 271-277.

Kirkorian, H. L., Lavigne, H. J., Hanson, K. G., Troseth, G. L., Demers, L. B., & Anderson, D. R. (2016). Video deficit in toddlers' object retrieval: What eye movements reveal about online cognition. *Infancy*, *21*(1), 37-64.

Kirkorian, H., Pempek, T., & Choi, K. (2017). The role of online processing in young children's learning from interactive and noninteractive digital media. In *Media exposure during infancy and early childhood* (pp. 65-89). Springer, Cham.

Krcmar, M. (2010). Can social meaningfulness and repeat exposure help infants and toddlers overcome the video deficit? *Media Psychol. 13*, 31–53. doi: 10.1080/ 15213260903562917.

Krcmar, M., Grela, B. G., and Lin, Y.-J. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychol. 10*, 41–63. doi: 10.1080/15213260701300931

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, 10, 110–120. doi:10.1111/j.1467- 7687.2007.00572.x

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, USA, 100, 9096–9101. doi:10.1073/pnas.1532872100

Reiser, R. A., Tessmer, M. A., & Phelps, P. C. (1984). Adult–child interaction in children's learning from ''Sesame Street". *Educational Communication and Technology Journal, 32*, 217–223.

Schroeder, E. L., & Kirkorian, H. L. (2016). When seeing is better than doing: Preschoolers' transfer of STEM skills using touchscreen games. *Frontiers in Psychology, 7*, 1377. https://doi.org/10.3389/fp syg.2016.01377

Sheehan, K. J., & Uttal, D. H. (2016). Children's learning from touch screens: a dual representation perspective. *Frontiers in psychology*, *7*, 1220.

Strouse, G. A., O'Doherty, K., & Troseth, G. L. (2013). Effective coviewing: Preschoolers' learning from video after a dialogic questioning intervention. *Developmental psychology*, *49*(12), 2368.

Strouse, G. A., Nyhout, A., & Ganea, P. A. (2018). The role of book features in young children's transfer of information from picture books to real-world contexts. *Frontiers in psychology*, *9*, 50.

Suresh, K. (2011). An overview of randomization techniques: an unbiased assessment of outcome in clinical research. Journal of Human *Reproductive Sciences, 4*(1), 8–11.

Lauricella, A. R., Pempek, T. A., Barr, R., & Calvert, S. L. (2010). Contingent computer interactions for young children's object retrieval success. *Journal of Applied Developmental Psychology*, *31*(5), 362-369.

Liben, L. S. (1999). Developing an understanding of external spatial representations. In I. E. Sigel (Ed.), Development of mental representation: Theories and applications. Mahwah, NJ: Erlbaum.

Luk, G., & Bialystok, E. (2005). How iconic are Chinese characters? *Bilingualism: Language and Cognition, 8*, 79-83.

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., and Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Dev. 75*, 1357–1372. doi: 10.1111/j.1467-8624.2004.00745.x

Merzenich, M. M. & Jenkins, W. M. (1995) in SFI Studies in the Sciences of Compliexity, eds. Julesz, B. & Kovacs, I.(Addison–Wesley, Reading, MA), Vol. XXIII, pp. 247–272.

Myers, L. J., Crawford, E., Murphy, C., Aka-Ezoua, E., & Felix, C. (2018). Eyes in the room trump eyes on the screen: effects of a responsive co-viewer on toddlers' responses to and learning from video chat. *Journal of Children and Media*, *12*(3), 275-294.

Myers, L. J., & Liben, L. S. (2012). Graphic symbols as "the mind on paper": Links between children's interpretive theory of mind and symbol understanding. *Child Development*, *83*(1), 186-202.

Namy, L. (2008). Recognition of iconicity doesn't come for free., *Developmental Science, 11*, 841-846.

Newcombe, N., & Huttenlocher, J. (2000). *Making space: The development of spatial representation and reasoning*. MIT Press.

R Development Core Team. (2016). R: A language and environment for statistical computing (Version 3.3.0) [Computer software]. Retrieved from https://www.r-project.org/index.html

Rideout, V. (2016). Measuring time spent with media: The Common Sense census of media use by US 8-to 18-year-olds. *Journal of Children and Media*, *10*(1), 138-144.

Rideout, V., Saphir, M., Tsang, V., & Bozdech, B. (2013). Zero to eight: Children's media use in America. Common Sense Media.

Salomon, G. (1983). Television watching and mental effort: A social psychological view. In J. Bryant & D. R. Anderson (Eds.), Children's understanding of television: Research on attention and comprehension (pp. 181–198). New York: Academic.

Schmidt, M. E., Crawley-Davis, A. M., & Anderson, D. R. (2007). Two-year-olds' object retrieval based on television: Testing a perceptual account. *Media Psychology*, *9*(2), 389-409.

Sharon, T., & DeLoache, J. S. (2003). The role of perseveration in children's symbolic understanding and skill. *Developmental Science, 6* (3), 289–296. doi: 10.1111/1467-7687.00285 .

Schmitt, K. L., & Anderson, D. R. (2002). Television and reality: Toddlers' use of visual information from video to guide behavior. *Media Psychology*, *4*(1), 51-76.

Sheehan, K. J., & Uttal, D. H. (2016). Children's learning from touch screens: a dual representation perspective. *Frontiers in psychology*, *7*, 1220.

Sheehan, K. J., Ferguson, B., Msall, C., and Uttal, D. H. (2020). Forgetting and symbolic insight: delay improves children's use of a novel symbol. *J. Exp. Child Psychol. 192*:104744. doi: 10.1016/j.jecp.2019.104744.

Suddendorf, T. (1999). Children's understanding of the relation between delayed video representation and current reality: a test for self-awareness?. *Journal of Experimental Child Psychology*, *72*(3), 157-176.

Strouse, G. A., & Samson, J. E. (2021). Learning From Video: A Meta-Analysis of the Video Deficit in Children Ages 0 to 6 Years. *Child development*, *92*(1), e20-e38.

Strouse, G. A., & Troseth, G. L. (2014). Supporting toddlers' transfer of word learning from video. *Cognitive Development*, *30*, 47-64.

Tiwari, S. (2020). Understanding the 3Cs: Child, Content, and Context in Children's Educational Media. *TechTrends*, 1-3.

Tarlowski, A. (2006). If it's an animal it has axons: Experi- ence and culture in preschool children's reasoning about animates. *Cognitive Development, 21*, 249–265.

Tare, M., Chiong, C., Ganea, P., & DeLoache, J. (2010). Less is more: How manipulative features affect children's learning from picture books. *Journal of Applied Developmental Psychology, 31*, 395–400. https://doi.org/10.1016/j.appdev.2010.06.005.

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Development, 69*, 950–965. https://doi.org/10.1111/j.14678624.1998.tb06153.x.

Troseth, G. L. (2003a). TV guide: Two-year-old children learn to use video as a source of information. *Developmental Psychology*, 39(1), 140–150. https://doi.org/10.1037/0012-1649.39.1.140.

Troseth, G. L. (2010). Is it life or is it Memorex? Video as a representation of reality. *Developmental Review*, *30*(2), 155-175.

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development, 77*, 786–799. http://doi.org10.1111/j.1467-8624.2006.00903.x.

Troseth, G. L., Bloom Picard, M. E., & DeLoache, J. S. (2007). Young children's use ofscale models: Testing an alternative to representational insight. *Developmental Science, 10*(6), 763–769. https://doi.org/10.1111/j.1467-7687.2007.00625.x.

Troseth, G. L., Strouse, G. A., Verdine, B. N., & Saylor, M. M. (2018). Let's chat: On-screen social responsiveness is not sufficient to support toddlers' word learning from video. *Frontiers in psychology*, *9*, 2195.

Troseth, G. L., Flores, I., & Stuckelman, Z. D. (2019). When representation becomes reality: Interactive digital media and symbolic development. In *Advances in child development and behavior* (Vol. 56, pp. 65-108). JAI.

Tsuji, S., Fievet, A. C., & Cristia, A. (2021). Toddler word learning from contingent screens with and without human presence. *Infant Behavior and Development*, *63*, 101553.

Valkenburg, P. M., & Peter, J. (2013). The differential susceptibility to media effects model. *Journal of communication*, *63*(2), 221-243.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children*, *23*(3), 34-41.

Waxman, S., & Medin, D. (2007). Experience and cultural models matter: Placing firm limits on childhood anthropocentrism. *Human Development*, *50*(1), 23-30.

Wilson, D., and Sperber, D. (2004). "Relevance theory," in The Handbook of Pragmatics, eds L. R. Horn and G. Ward (Oxford: Blackwell), 607–632.

Appendix

Exploratory Analyses on the Moderating Effect of Child Age

Due to the relatively small sample size in current study, the statistical power was limited for more complicated analyses on the potential interaction between child age experimental condition. We addressed this limitation to the extent possible by conducting additional analyses (e.g., model fitness comparison, randomization check) to ensure that our results reported in the Results section were robust to the inclusion of age. However, in order to acknowledge the important role of age in children's video-mediate learning, we used plots to visually inspect the potential moderating role of age in interactivity effects on children's symbolic transfer.

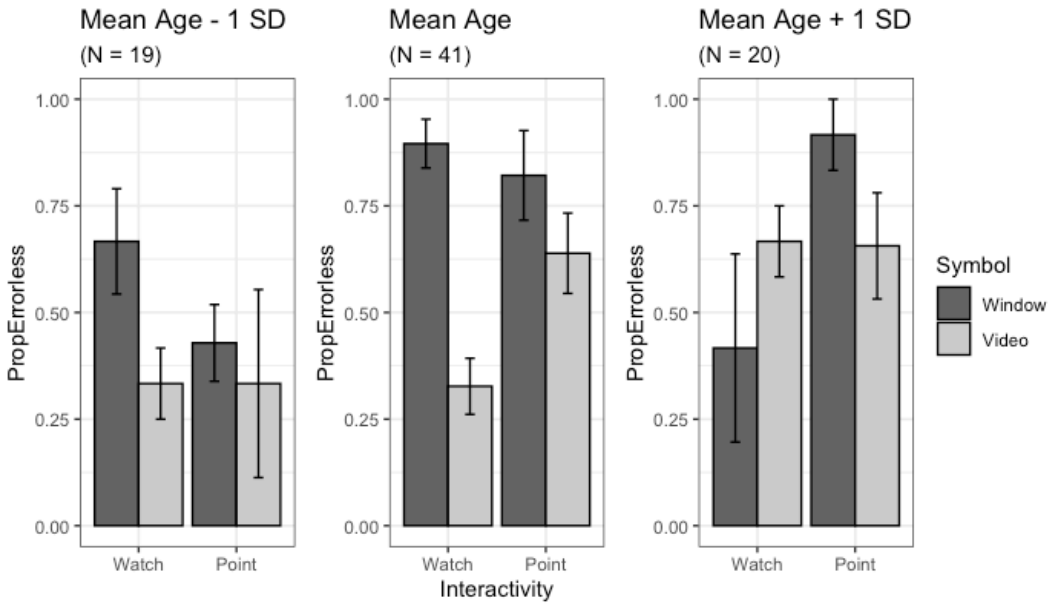

Figure 9. Average proportion of errorless retrievals per child (across all trials) by condition and by age group: 1 *SD* below the mean age (i.e., 26 months), mean age (i.e., 30 months), and 1 *SD* above the mean (i.e., 34 months).

As shown in Figure 9, based on visual inspection, the video-point condition outperformed the video-watch condition in children at the mean age, while there was little difference between these two conditions in the younger (i.e., 1 *SD* below the mean) and older children (i.e., 1 *SD* above the mean). This tendency suggests that age could moderate the effect of video interactivity, by which the interactive features is most beneficial to children of certain age while less effective in children who are too young or too old. While future research is needed to test whether this tendency is meaningful, it is in line with our discussion on the role of task difficulty that the effectiveness of interactivity may be maximal when the task is just above children's capacity. In the case of age, while the older children in our sample found the task too easy and younger children found it too difficult, children in the middle age benefitted most from the interactive features.

CHAPTER 4

General Discussion

Since the very beginning of life, children are exposed to substantial information and gradually acquire the knowledge about the world from all possible sources of information around them. Starting from observing the physical environment and the behavior of people around them, they quickly expand their learning opportunities by gaining information from indirect experience. Learning from symbolic media is critical to children, as it helps to solve a wide range of problems, acquire abstract concepts, and be more efficient and sophisticated learners (DeLoache et al., 2004; Uttal et al., 2009). Screen media are ubiquitous in today's life and becoming increasingly important learning tools for children at increasingly younger ages (Rideout, 2016). Therefore, extending basic research on how children process information from digital media is important to building the literature on learning as well as developing materials that support early learning.

To better understand children's processing of digital media information and further facilitate video-mediated learning, this dissertation investigated children's behaviors at two stages of information processing, including the initial stage of attending to information and the later stage of retrieving information. Influencing factors at the content, developmental, and contextual level were examined to explore potential mechanisms underlying video-mediated learning at each information-processing stage.

Despite the many barriers that children could encounter in perceiving and learning from digital media, this dissertation replicated prior research in demonstrating that children's digital media activities are cognitively active, rather than devoid of understanding or learning (e.g., Anderson & Lorch, 1983; Huston & Wright, 1983; Salomon, 1983). Specifically, the visual

attention of children in Study 1, as signaled by their eye-movement patterns, was found to be impacted by the shot transitions (i.e., gaze moving toward the center of the screen). This tendency of reorienting eye fixation to the center of the screen has been well documented in adults during video watching as a sophisticated strategy of processing the video content more efficiently (Kirkorian et al., 2012). Four-year-olds in our sample started to show such an adult-like viewing, consistent with prior findings on other gaze behavior (Franchak et al., 2016). Shot transitions have been studied extensively as an important context to learn about children's video comprehension, because this editing technique of conveying continuity of the scenes and objects through a sequence of discrete shots is critical to comprehending video content (Smith et al., 1985). Thus, our finding on the shot transition suggests that, rather than being reflexively attending to the video due to dynamic production features (Lesser, 1977; Singer, 1980), children can and do cognitively process the on-screen information while watching video. Moreover, as a more direct evidence, Study 2 found that toddlers used the information on video to retrieve the hidden object at above chance level, particularly when the interactive feature was added to support the process (e.g., directing attention to target information).

Nonetheless, there was apparent deficit involved in children's screen-mediated behaviors, when compared to adults (Study 1) or compared to children's own behaviors during in-person experience (Study 2). Our finding from Study 2 replicated prior research and reveals a video deficit whereby children transferred information better from in-person events as compare to video. Furthermore, while adults in Study 1 showed significantly different gaze patters when watching an incomprehensible (vs. comprehensible) video, children's eye movement, as measured by salience-based gaze prediction, remained similar regardless of the content feature – video comprehensibility. As discussed in the Study 1 paper, this finding may reveal relatively

weaker comprehension-related top-down control in children's visual attention to on-screen information, compared to adults. Indeed, prior research using other visual attention metrics, such as overt looking at the screen or attentional synchrony, has similarly demonstrated age-related changes in attention to and visual processing of video (e.g., Pempek et al., 2010; Franchak et al., 2016; Frank et al., 2009). Consistently, Study 2 examined the age effect at the information-retrieval stage, and found that children's object-retrieval performance increased with age. These findings, and those from developmental research (e.g., Barr, 2010), indicate a rapid development in early processing of screen media in spite of the video deficit.

While acknowledging both the limitation and potential of early learning from screen media, this dissertation investigated how media characteristics could influence the information processing of screen media content. The results presented in this dissertation suggests that content comprehensibility may have a marginal impact on visual attention in 4-year-olds, but an interactive interface could benefit toddlers' information retrieval to a significant degree. Focusing on the initial stage of attending to information, Study 1 found no impact of video comprehensibility on children's eye movements and an interaction between the comprehensibility effect and age. By examining the age-related change in the effect of video content on salience-based gaze, this study enabled a nuanced discussion on the developmental mechanisms involve in children's visual attention to video. The findings suggest that 1) the coherence of video content, or video comprehensibility, may not affect the degree to which children's gaze is directed by low-level visual features during naturalistic view viewing and 2) the increased visual attention to salience-related bottom-up features observed in adults might be due to something other than their better comprehension ability. An *overall predictability*

hypothesis was proposed to explain the age-related difference in visual attention and processing of video content.

Focusing on the retrieval outcome, Study 2 built on prior work to disentangle potential mechanisms underlying the interactivity effect by incorporating the interactive interface, equivalent to that in video context, to the live scenario. The findings supplement previous research on a positive effect of interactivity on symbolic transfer and provide further evidence to the *attending and encoding hypothesis* to account for the role of interactivity at the information-encoding stage.

Taken together, the two studies, by focusing on different information-processing stages, highlight the active engagement of multiple cognitive processes in children's interaction with screen media. In particular, they collectively underscore that visual attention allocation is a key process for young children's information processing of on-screen information. According to findings from Study 1, children's visual attention was predicted by visual salience on the screen, in spite of little influence from the content comprehensibility of video. This research indicates that media creators should use visual salience strategically, drawing attention to important information on the screen and avoiding irrelevant visual noise. While it remains to be seen whether these effects on attention lead to differences in actual comprehension, future research should include a comprehension assessment to directly measure children's video comprehension. Meanwhile, Study 2 suggests that the facilitative effect of interactive features is more likely through enhancing children's selective attention to and encoding of relevant information rather than other mechanisms. It suggests the potential that visual features are sufficient to enhance video-mediated transfer. Future work could test this by 1) replacing the interactive feature based on physical action (as used in Study 2) with interactive interface that merely use visually-salient

features to direct children's attention and 2) including direct measures of visual attention (e.g., eye tracking).

Importantly, neither study found an association between the children's screen-mediated behaviors (e.g., visual attention, information retrieval) and their characteristics at the socio-contextual level. Prior research has demonstrated that learning from screen media is situated in and conditional upon the context where the learning happens (Guernsey, 2007; Valkenburg & Peter, 2013). Considering the measures for naturalistic media experience used as discussed in the limitations in Chapter 3, future research might benefit from focusing on particular contextual factors using experimental manipulation or more specific survey questions (e.g., addressing specific media content or functions) to examine how these factors could moderate children's learning from screen media.

In addition, despite substantial research and the results from this dissertation show age as a key factor in the influence of media characteristics on screen-mediated learning, the limited sample size lacks sufficient power to analyze the moderating role of age in the effect of interactivity. While I did an exploratory analysis based on visual inspection (see Appendix of Chapter 3), more work is needed to examine the specific moderating role of age and, particularly, seek to identify specific correlates of age that might underlie age effects (e.g., working memory, inhibitory control, media experience).

In sum, this dissertation sought to provide a comprehensive view of children's information processing of screen media at the basic cognitive level while taking into consideration the 3 C's (Guernsey, 2007): content, context, and child factors. The findings provide direct evidence for the effectiveness of two important media characteristics on screen-mediated learning, and add to the literature about how the learning is supported by fundamental

cognitive systems, such as selective attention and encoding and symbolic representation. Findings from this type of research can be applied to help create scalable and cost-effective media that facilitate early learning and development.

References

Anderson, D. R., Lorch, E. P., Field, D. E., Collins, P. A., & Nathan, J. G. (1986). Television viewing at home: Age trends in visual attention and time with TV. *Child development*, 1024-1033.

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review*, *30*(2), 128-154

DeLoache, J. S., Uttal, D. H., & Rosengren, K. S. (2004). Scale errors offer evidence for a perception-action dissociation early in life. *Science*, *304*(5673), 1027-1029.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110, 160–170.

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy*, *21*(3), 262-287.

Guernsey, L. (2007). *Into the minds of babes: How screen time affects children from birth to age five*. Basic Books.

Kirkorian, H. L., Anderson, D. R., & Keen, R. (2012). Age differences in online processing of video: An eye movement study. *Child development*, *83*(2), 497-507.

Lesser, H. Television and the preschool child. New York Academic Press, 1977.

Pempek, T. A., Kirkorian, H. L., Richards, J. E., Anderson, D. R., Lund, A. F., & Stevens, M. (2010). Video comprehensibility and attention in very young children. *Developmental psychology*, *46*(5), 1283.

Rideout, V. (2016). Measuring time spent with media: The Common Sense census of media use by US 8-to 18-year-olds. *Journal of Children and Media*, *10*(1), 138-144.

Salomon, G. (1983). Television watching and mental effort: A social psychological view. In J. Bryant & D. R. Anderson (Eds.), Children's understanding of television: Research on attention and comprehension (pp. 181–198). New York: Academic.

Singer, J. L. (1980). The power and limitations of television: A cognitive– affective analysis. In P. H. Tannenbaum & R. Abeles (Eds.), The entertainment functions of television (pp. 31–65). Hillsdale, NJ: Erlbaum.

Smith, R., Anderson, D. R., & Fischer, C. (1985). Young children's comprehension of montage. *Child development*, 962-971.

Uttal, D. H., O'Doherty, K., Newland, R., Hand, L. L., & DeLoache, J. (2009). Dual representation and the linking of concrete and symbolic representations. *Child Development Perspectives*, *3*(3), 156-159.

Wright, J. C., & Huston, A. C. (1983). A matter of form: Potentials of television for young viewers. *American Psychologist*, *38*(7), 835.

Valkenburg, P. M., & Peter, J. (2013). The differential susceptibility to media effects model. *Journal of communication*, *63*(2), 221-243.