CHILDREN'S THEORY OF MIND, JOINT ATTENTION, AND VIDEO CHAT

by

RYAN CURRY

Submitted in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Cognitive Science

CASE WESTERN RESERVE UNIVERSITY

May, 2021

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis of

Ryan Curry

Candidate for the degree of **Master of Arts***

Committee Chair

Fey Parrill

Committee Member

Mark Turner

Committee Member

Elizabeth Short

Date of Defense

March 9th, 2021

*We also certify that written approval has been obtained

for any proprietary material contained therein.

Table of Contents

List of Tables

List of Figures

Children's Theory of Mind, Joint Attention, and Video Chat

Abstract

by

Ryan Curry

Video chat offers a unique communication experience some children may find confusing and challenging to navigate, due to its inherent limitations, such as restricted visual access. In this study, we examine whether children's theory of mind (ToM) impacts how well they are able to accommodate or overcome these limitations in order to effectively communicate and engage in joint attention over video chat. We evaluated 31 children, between the ages of 2 and 4.5 years on their ToM, and then tested them in six joint attention trials testing their skill at initiating joint attention (IJA) or responding to joint attention (RJA). Our results show that ToM is significantly related to children's IJA, but not RJA, on video chat, including situations where the item of reference was outside the view of the child's webcam. These results suggest that unlike IJA, factors other than ToM likely impact a child's RJA on video chat.

Introduction

Video chatting has grown in prevalence over the past several years—one estimate states that from 2007 to 2011, there was a 900% increase in video chat usage (Scelfo, 2011). Earlier in the century, the American Academy of Pediatrics had advised against media exposure of any type for children under the age of 2 (American Academy of Pediatrics, 2011). But as of 2016, the AAP now recommends allowing moderate video chat usage, seeing it as an exception that merits forgoing strict limitation (American Academy of Pediatrics, 2016). Even before, the AAP made this recommendation, McClure, Chentsova-Dutton, Barr, Holochwost, and Parrott (2015) found that video chat was widely used by children under the age of 2, most often to connect with distant family members. They also found that parents view video chatting more positively than other sources of screen media, and often consider it an exception to parentally imposed media limitations. It is for good reason that video chatting is seen as an exception among other forms of media because from a developmental perspective, it is very much unlike other forms of media. Additionally, video chatting is unlike other forms of communication, given the complex visuospatial elements at play when trying to communicate with someone on the other end of the video chat. To support these two points, we will first look at the relationship children have with pre-recorded video, and then compare that relationship to the relationship children have with video chat. After that, we will move from talking about technology to discussing the development of children's theory of mind (ToM), and the role joint attentional abilities, perceptual perspective-taking, gesture usage, and language usage play in supporting that development. And lastly, we will return to talking about technology to mention some of the difficulties video chat imposes

on joint attention, before concluding our argument and proposing the purpose of this study.

Video Deficit

With the continued growth of media exposure in children, one key question is: how well can children consume the media they are viewing? The answer is that it depends. When watching a video, children under the age of 2 experience a 'video deficit' where they retain less information learned from pre-recorded video than from an equivalent live experience (Anderson & Pempek, 2005). However, that's not to say children gain nothing from watching videos of any sort. Multitudes of studies have shown there are educational benefits to watching television programs, such as *Sesame Street*, at a young age. For instance, Anderson et al. (2001) conducted a longitudinal study observing the relationship between preschool television viewing and later educational performance. They found that children who viewed informative children's programs—*Sesame Street* in particular—were seen to perform at higher levels in English, math, and science in school and do more leisure reading later in adolescence. Additionally, Linebarger and Walker (2005) found that children who watched shows such as *Dora the Explorer, Blue's Clues, Arthur, Clifford,* and *Dragon Tales,* starting at 6 months of age, showed larger vocabularies and superior expressive language usage at 30 months of age. However, children who watched *Teletubbies* showed the opposite results, with smaller vocabularies and less frequent expressive language usage. These studies show that children can learn from informationally rich programs, but not all programs are the same. There are instead certain elements of the learning context that dictate how well children can absorb the information they encounter.

Comparing children's ability to learn through various mediums, Barr (2010) found that infants in an experimental group who witnessed actions performed in books or videos produced significantly more of those target actions than infants in a control group who did not witness any material, suggesting that infants can, indeed, learn directly from television, as well as books and touchscreens. However, like Anderson and Pempek (2005), Barr (2010) found that infants imitated fewer target actions after viewing a 2D demonstration, compared to a live demonstration, showing a diminished return on time investment. Additionally, this 'video deficit' impacts the length of time for which a child remembers an action he or she has learned from a video. Brito, Barr, McIntyre, and Simcock (2012) found that 18-month-olds were only able to retain knowledge of how to perform an action sequence learned from a pre-recorded video for half as long as children who learned from a live performance. An equivalent result to children who learned from pre-recorded video was seen in children who learned the action sequence from a book.

Barr (2013) argues that this 'transfer deficit' of information learned from media results from disparities in the way children perceive 3D contexts within 2D contexts that lead to difficulty transferring information between these dimensions. This disparity leads to attributes represented in the original coding of a memory that do not match up with attributes perceived at the time of retrieval (Tulving, 1983), thus leading young children to have difficulty in recalling information. Barr (2013) goes on to outline three factors that lead to this mismatch. The first factor is the lack of perceptual stimulation of 2D images present in the 3D world. For instance, compared to experiencing objects in person, 2D objects are smaller and lack depth cues. The second factor is a contextual or source mismatch that is present in a 2D context. For instance, children may encode

physical properties of a 2D context, such as the edges of a screen or the pages of a book, that are not relevant to the information provided in that context, thus binding the two attributes into one memory, and therefore limiting their ability to retrieve the memory. The third factor is the symbolic nature of books and screens, which makes it harder for children to utilize information learned from them. Books and screens are simultaneously objects as well as the story or game or images displayed inside them. It may be challenging for children to hold this dual representation in their minds simultaneously. These three factors affect all forms of media, including video chat. Fortunately, some of this is ameliorated by other aspects of video chat.

Video Chat and Social Contingency

A major element that pre-recorded video lacks is the presence of social cues, which has been shown to contribute to children's learning. To start, a 1996 study by Hains and Muir found that 5-month-old infants visually attended to both live video chat and face-to-face interaction longer than a non-contingent, pre-recorded video replay. However, despite displaying equal levels of attention in both conditions, infants were seen to smile earlier and more often in the face-to-face condition compared to the video chat condition. This suggests that infants, despite their young age, can detect differences between the two types of interactions and thus display at least some variation in behavior. The reason for this lower level of attention to the non-contingent video is explained by Kuhl, Tsao, and Liu (2003), whose research suggests interpersonal social cues of face-to-face interactions, such as eye gaze and body language, help to attract an infant's attention to a speaker and allow for sharing of information referentially, which both help to facilitate learning. This is the process of joint attention, which, as Tomasello (2003)

argues, plays a crucial role in children's language acquisition and development in general. Joint engagement of a particular item or activity acts as a means to provide a common ground for parent and child to communicate, and this is most typically done through pointing gestures made either by the parent or by the child (Bangerter, 2005). In pre-recorded video, these elements of engagement are limited or absent altogether, therefore, making learning from non-contingent video less engaging for children, and much more challenging. Importantly, Tarasuik, Galligan, and Kaufman (2011) found that, despite its similarity to non-contingent video, video chatting is seen by children to be an acceptable supplement to physical presence. In this study, children who were left alone in a strange room without their parents physically there, but present via video chat, were content to stay in the room longer than children left alone entirely. The children in the video-chat condition additionally had the same level of interactivity with their parents via the video chat as they did during physically present free play.

These studies, therefore, point to the possibility of ameliorating a 'video deficit' in video chats through social contingency. Troseth, Saylor, and Archer (2006) was one of the first studies to test this by looking at infants' ability to learn from face-to-face interaction, contingent closed-circuit video, and pre-recorded video. In this study, 2-year-old children who were told the location of a hidden toy in a face-to-face interaction were much more likely to find the toy than children given the same information via a pre-recorded video. Likewise, children who were given the information via a closed-circuit video with an experimenter who explicitly provided socially contingent elements to the interaction, such as saying the child's name, responding to the child's words and actions, and providing information about the child's environment, were also more successful than

the children in the pre-recorded video condition. Similar findings were made in Roseberry, Hirsh-Pasek, and Golinkoff (2014), where children aged 24-30 months successfully learned novel words from both live interactions, as well as video chat (social contingent conditions), but not from watching a non-contingent video. In both contingent conditions, this learning was sufficiently robust to extend usage of the novel verb to new instances. Additionally, these findings were later supported and expanded upon by Myers, LeWitt, Gallo, and Maselli (2017), who found that children ages 17-25 months, divided into either a video chat or non-contingent video condition, were all attentive and responsive during the study, but only children in the video chat condition gave temporally synced responses; these children formed social connections with their partners, recognizing them after one week, and learning content more readily from their partners than children in the non-contingent video condition. Roseberry, Hirsh-Pasek, and Golinkoff's (2014) results suggest that children are very adept at distinguishing pseudo-socially-contingent interactions involving one-sided, "ask and wait" models—such as the approach they used in their non-contingent video and the approach commonly used in television programs such as *Blue's Clues* and *Dora the Explorer*— from actual socially-contingent interactions. When young children have live interactions with an adult, social contingency helps them to establish that adult as a reliable and accurate source through correct and temporally synced responses; in a pre-recorded video, however, the lack of contingency produces the opposite effort whereby reliability is undermined by inaccurate content (Roseberry et al., 2014). For this reason, video chat has a resistance to the 'video deficit', which makes it a comparable substitute to face-to-face interactions, although not a perfect substitute given the challenges resulting from the inherent limitations of the

medium. We will cover these limitations, but for now it is important to look more deeply into children's use of joint attention and how it relates to their development of ToM.

Theory of Mind and Joint Attention

Theory of mind is the understanding that everyone, including oneself, has unique mental states, such as differing desires, beliefs, intentions, and emotions, that influence one's actions. This understanding requires realizing that though mental states may reflect reality and lead to overt actions, they are nonetheless distinct from real-world events as they are entirely internal. Therefore, it is possible to hold false-beliefs that do not reflect reality (Wellman, Cross, &Watson, 2001). Tomasello, Kruger, and Ratner (1993) argue that children's understanding of others as intentional agents, whose attention and behavior can be followed, or guided to an object in their physical environment, begins around their first birthday, and develops into an understanding of others as unique mental agents, capable of holding false beliefs; around the age of 4, this understanding is referred to as obtaining a 'representational theory of mind'. The evaluation of a child's false-belief understanding has become a common milestone used to evaluate a child's overall ToM understanding. Wimmer and Perner (1983) was the first study to introduce the following story to use in assessing a child's false-belief understanding: a child named Maxi puts his chocolate bar in a cupboard, then leaves the room. While he's gone, his mother moves the chocolate bar to a different cupboard. The children being evaluated are then asked, "When Maxi returns, where will he look for the chocolate bar?" Younger children will typically fail to account for Maxi's false belief and choose the second drawer, where the chocolate bar actually is. Whereas older children typically choose the first drawer, where Maxi had put the chocolate bar.

In their meta-analysis of the literature on false-belief research, Wellman, Cross, and Watson (2001) found support for an existing view that children's ToM undergoes a conceptual change throughout childhood, such that as they develop, so too does their understanding of false-belief and ToM as a whole. Wellman and Lui (2004) expands on this argument by going beyond focusing on false-belief tasks and proposing a developmental progression that young children consistently follow as they gradually develop ToM. They argue that progress of a child's individual ToM development can be evaluated through the use of a scale they created from their observations. This scale involves seven tasks of increasing difficulty, such that children who pass later tasks, should also be able to pass all earlier tasks on the scale. For instance, understanding another person's desires— which can be seen in children as young as 18-months-olds (Repacholi & Gopnik, 1997)—is considered the most basic level of ToM understanding because it simply involves realizing that the other person has internal urgings as a result of external stimuli. Understanding another person's beliefs, however, is more challenging because it requires understanding that the other person holds a mental representation of the world that may be both different from one's own and contradict reality (i.e. a false belief) (Wellman et al., 2001). Thus, as children's perception of other's mental states develop, they progress along Wellman and Lui's scale, and gain a more complete understanding of ToM.

One of the most crucial activities in which children must participate to improve their understanding of other people's mental states— and ToM by extension— is joint attention engagement with adults and other children. Joint attention, as briefly mentioned earlier, is the act of establishing a common ground, using an item or activity, from which

a child and parent can communicate. But, importantly, joint attention requires active knowledge that the other person is focusing on the same item or activity as one's self (Tomasello, 1995). Joint attention behaviors can be broken down into two major components. The first is 'referential looking' which takes place either when someone looks towards the object of another person's visual focus, or when someone tries to get someone else to look at something by directing them with their own gaze (Baron-Cohen, 1991). The second component is gesture, specifically deictic gestures, such as pointing, giving, or showing, as opposed to conventional gestures, such as head nods, which have failed to show any relation to joint attention (Salo, Rowe, & Reeb-Sutherland, 2018). We will talk more about deictic gestures, in particular, pointing, later. An extension of referential looking is gaze shifting, which is highly common in children, first starting at around 9 months of age, in that they will continue to look back and forth between an object and an adult's face, while also making a communicative gesture, until the adult takes notice of the child's referent (Bates, 1976). Prior to the age of 12 months, children's engagement in joint attention is typically limited to their visual field at that moment, but between the ages of 12 and 18 months, children gain the ability to search outside their visual field, such as to either side or behind themselves, to locate a target of reference (Butterworth & Jarrett, 1991). This development helps to better accommodate responding to joint attention (RJA).

The first hints of an understanding of general attention can be seen in children as young as seven- to nine-months-old and is considered by some to be the origin of ToM (Stern, 1985; Baron-Cohen, 1991). But a full understanding of attention goes beyond simply determining someone's line of sight, as it requires also understanding that vision

and other perceptions can be selectively directed based on the level of interest someone has in an object or event in a certain location (Baron-Cohen, 1991), and that different aspects of the object or event can be the focus of someone's attention (Tomasello, 1995). Therefore, simultaneous orientation of two people's attention to the same location does not guarantee joint attention, for joint attention requires that both parties hold active knowledge that the other person is focusing on the same object or event that they are, hence, the importance of utilizing gaze shifting to maintain that knowledge. For this reason, Baron-Cohen (1991) argues that children must first understand perception before they can understand attention. This is not to say that understanding perception in its entirety is necessary prior to any understanding of attention, but instead that progress in understanding the former precedes progress in understanding the latter.

A related component of these progressions, and another step along the path to attaining ToM, is learning to understand knowledge access, or the absence of knowledge in the form of ignorance. Understanding ignorance is a prerequisite to understanding false belief because for a false belief to occur, the individual who holds that belief must be ignorant to true reality. This logic was tested and affirmed by Hogrefe, Wimmer, and Perner (1986). The development of knowledge access understanding in children reflects the development of visual perception understanding, and can be described using a model first proposed by Flavell (1978, 1992), consisting of two distinct levels of understanding. Level 1 perceptual perspective-taking (L1P) develops early in life and allows children to realize that what another person can and cannot see may be different from what they, themselves, can and cannot see. Level 2 perceptual perspective-taking (L2P) develops later and gives children the understanding that even when an object is visible to another

person, they may have a different visual experience of that object as a result of their differing perspective. Flavell, Shipstead, and Croft (1978) found that L1P understanding is obtained around the age of 2.5, a finding Moll and Tomasello (2006) expanded by observing the beginnings of L1P ability in children as young 2-years-old, which they assume then undergoes further development over the next several months. L2P understanding starts to surface around the age of 4, at the same time false-belief understanding begins to take hold, which Perner, Stummer, Sprung, and Doherty (2002) argues is because understanding false-beliefs requires that children be able to solidly grasp different perspectives such as one holding a false belief vs. the perspective of reality.

This L1P distinction of being, or not being, able to see something is closely related to the distinction between knowing, or not knowing, something (knowledge access) because of the role sight plays as a source of knowledge (Hogrefe et al., 1986). To this end, Taylor (1988) proposed a hypothesis parallel to that of Flavell's, concerning children's conceptual perspective-taking. This hypothesis consists of two levels, as well. Level 1 conceptual perspective-taking (L1C) marks the beginning of basic knowledge access understanding. At this level, children struggle with differentiating their own knowledge from what is available in the environment, but they do understand that if someone does not see an object they might not know about it. At Level 2 (L2C), children understand that the same visual information may not lead to the same understanding or knowledge in two people. Taylor claims that L1C emerges around the age of 4, and L2C around the age of 6. However, Pillow (1989) found that children as early as age 3 show some L1C ability, performing equally well at reporting their own knowledge vs.

assessing another person's knowledge. So as we can see from these ages of emergence, perceptual understanding proceeds conceptual understanding for both levels. Returning to Baron-Cohen's (1991) argument of perception leading to attention, Baron-Cohen builds on this by arguing that belief arises from attention, a conclusion he formed by studying autistic children who, as a whole, struggle with establishing joint attention, which then appears to inhibit their ability to understand others' belief mental states. However, these children do not struggle with understanding others' simple desires or the emotions related to those desires, which do not require attention to interpret. He postulates that this is because beliefs are always abstract, while emotions and desires need not be, which makes desires easier for children to understand at a young age.

From these findings, we see that prior to their second birthday, most children have some understanding of desires (Repacholi & Gopnik, 1997). Then starting around the age of 2, children gain a L1P ability (Moll & Tomasello, 2006), which then develops into a L1C ability over about the next year (Pillow, 1989). These developments reflect an improved ability to initiate joint attention (IJA; for a full list of terms mentioned in this section, see Table 1). Then as these abilities develop further, right around the age of 4, children gain a L2P ability which enables their understanding, first, of other's beliefs, and then, of other's false beliefs (Perner et al., 2002). This progression of an understanding of desires proceeding that of beliefs, which relies first on an understanding of perception and then knowledge access, which support understanding of attention, all fits within the progression outlined by Wellman and Lui (2004) and paints the complicated picture that is the development of ToM. But the important take away for our purposes is

Table 1

List of acronyms and their definitions

Acronym	Term
L1P	Level 1 Perceptual Perspective-Taking
L2P	Level 2 Perceptual Perspective-Taking
L1C	Level 1 Conceptual Perspective-Taking
L2C	Level 2 Conceptual Perspective-Taking
IJA	Initiating Joint Attention
RJA	Responding to Joint Attention

understanding the relationship between perception, knowledge access, and joint attention, and the role they play in helping children ultimately attain ToM.

Gestures and Theory of Mind

Returning to the second of the two components of joint attention: gesture, it's important to review the role this component, specifically pointing gestures, plays in helping children establish ToM. As mentioned, both gestures and referential looking are often done in combination (Baron-Cohen, 1991), and pointing is frequently used by children for IJA (Bangerter, 2004). However, there are two major types of pointing: imperative and declarative. Imperative pointing is used by children to request help from an adult in attaining a goal, for instance, in requesting assistance with getting something that is out of reach (Tomasello, Carpenter, & Liszkowski, 2007). Vygotsky (1978) proposed that imperative pointing results from failed action, in that when infants first attempt to reach for an object that's too far away, and their mother then brings the object to them, they learn to point in order to obtain objects in the future. In other words, imperative points lead to behavioral effects in others by causing them to *do* something.

Declarative points, on the other hand, cause epistemic effects in others by causing them to *know* something. This difference between the two type of pointing gestures is reflected in their relation to referents. In imperative bids, the referent of a child's gesture is often clear and easily attended to because it is typically located in the child and adult's shared visual field, and is possibly already an object of focus. But in declarative bids, the referent is more ambiguous as it is often outside either the adult's or the child's visual field, and therefore, requires an alteration in perception in order to attend (Kristen et al., 2011).

Another distinction between imperative and declarative point is that declarative pointing can be broken down into either expressive declaratives: sharing with a companion one's emotional state regarding an object, event, location, etc., or informative declaratives: sharing with a companion information that they need or desire (Tomasello et al., 2007). A significant difference between these two types of declarative points is that informative declarative pointing, but not expressive declarative pointing, is strongly tied to ToM ability, such that children with greater ToM ability produced more informative declarative points (Cochet et al., 2016). This is likely the case as a result of the difference in mental state awareness needed for each type of point. As Baron-Cohen (1991) argues, imperative pointing does not rely on a person's mental state, as it simply instigates physical, rather than mental, interaction, such as getting or giving an object. Whereas, declarative pointing requires mental state acknowledgment, as the intention is to get another person to take notice of or comment on an object, rather than physically interact with it. We will amend this statement by adding that this is even more so the case for

informatives, as they require interpreting the other person's desires, as opposed to expressives, which simply share one's own mental state.

Following its use to establish joint attention, gesture can then act as precursor for language that might not yet be within a child's capability. In fact, the skill with which a child uses gesture to communicate has been argued to predict later language comprehension and production (Colonnesi, Stams, Koster, & Noom, 2010). Support of this claim can be seen in Iverson and Goldin-Meadow (2005), which found that the objects at which children first point were then referenced verbally about 3 months later. Additionally, Iverson and Goldin-Meadow found that once children find the ability to meaningfully combine speech and gesture, gesture-plus-word usages (i.e. pointing at a mug and saying "daddy" to mean "daddy's mug") strongly predict the future two-word combination. Thus, in both the scenarios, children are first using gesture in place of an utterance they could not yet produce. This happens because shared attention on an item that is perceptually salient to both a child and an adult provides the child with a nonlinguistic scaffolding, from which they can assume the topic of discussion and produce utterances of relevance to the interaction, as well as combinations of gestures and words (Tomasello, 1988). What's important about this role pointing plays in supporting language development—with regards to our discussion thus far—is that language then in turn plays an important role in ToM development.

Joint Attention and Language

Language development is vital to attaining certain aspects of ToM understanding. For instance, based on correlational data, de Villiers and de Villiers (2000) proposed that a prerequisite of false-belief understanding is having sufficient language capability to

comprehend the use of complements in complex speech constructions, such as "he thought he had left the door open (when in reality he had shut it)." This sentence proposes a mental representation of the world that is false: a false belief. Therefore, children need sufficient linguistic competency to make sense of this representation, because that competency gives them a structure with which to conceptualize and talk about this false belief. Lohmann and Tomasello (2003) verified this hypothesis through experimental data, concluding that sentential complements, as well as perspective-shifting discourse and linguistic experience as whole, do in fact play a central role in facilitating the development of false-belief understanding. Milligan, Astington, and Dack (2007) further supported this argument through its meta-analysis of numerous studies by finding a significant relationship between false-belief understanding and children's language ability. Researchers broke down the relationship by various facets of language: complements, semantics, receptive vocabulary, syntax, and general language. Of significance to our purposes here, they found that complements and general language had an effect size accounting for 44% and 27% of variance in false-belief understanding, respectively, in all the studies' participants. On top of this, successful connected communication, such as that afforded by the social contingency video chat, concurrently predicts false-belief capability (Dunn & Cutting, 1999).

Beyond seeing how language development fits in to false-belief understanding, we can also see how it relates to our previous discussion on children's development of visual perception: we can see this development reflected in their speech capability. As children get older, their ability to account for the perspective of a companion in order to use spatial deictic terms, such as "here", "there", "in front of", or "behind", gradually

improves. Around the age of 2.5, children are better at using simple deictic terms that don't require a perspective shift, such as "this", "that", "here", or "there", but by around age 4, they are adept at translating speech into their companion's perspective in order to use more complex terms, such as "in front of" or "behind" (de Villiers & de Villiers, 1974). Early use of even simple deictic terms by young children shows at least a minimal understanding of another person's mind as use of these terms relies on an alignment with the spatial position of the speaker and the listener (de Villiers, 2007). From our previous discussion, we can notice that a child's limited ability to use certain deictic terms reflect their burgeoning visual perception understanding, such that L1P is attained around when simple spatial deictic terms are preferred, and L2P is attained around when more complex terms become part of a child's mental lexicon.

Now that we've covered the less-than-simple relationship joint attention, perception, gesture, and language have to ToM, it's time to return to how all these relationships relate to video chat.

Video Chats and Joint Attention Challenges

Video chat is a unique social medium for children to navigate. It supports some important elements of social cueing, but it has limits, many of which result from lack of physical proximity between a child and his or her video-chat companion. Without physical proximity, problems arise from factors such as limitations to movement and field of vision, misalignment of eye gaze resulting from misaligned webcams and screens, and difficulty processing the actual vs. perceived size of items on screen (Parkinson & Lea, 2011). Because of the important role eye contact and gaze following play in communication for children, a misalignment of eye gaze resulting from poor

webcam placement could potentially lead to less social behavior from a young children inexperienced with video chat, thus decreasing the likelihood they will desire future interaction with that partner (McClure & Barr, 2017). Many of these physical challenges tie back to problems mentioned earlier from Barr (2013) regarding lack of perceptual stimulation of 2D images, and contextual or source mismatches present in a 2D context.

These physical challenges have a profound effect on children's ability to engage with joint attention on video chat. As mentioned earlier, children gain the ability to engage in joint attention bids for items outside their visual field during the age of 12- to 18-months (Butterworth & Jarrett, 1991). McClure and Barr (2017) argues that once a child's understanding of joint, visual attention develops to this point, being able to then transfer this understanding to video chat is an added complexity that is likely beyond their capabilities at that time. For instance, while on a video chat, only a very limited number of items in a chat participant's vicinity, dictated by the view of their webcam, are eligible for joint attention. As a result of this, if a child points to an object within their view, but outside the view of the webcam, their screen partner will not know at what the child is pointing. McClure et al. (2018) identified this difficulty, finding that most children between the ages of 6 and 15 months rarely initiated joint attention on video chat, and those that did were only able to initiate what they called *within-screen*, joint visual attention: directing a partner's attention to an object on their own side of the screen. However, some older children, between the ages of 16 and 24 months, were able to successfully initiate *across-screen,* joint visual attention: directing a partner's attention to an object on the partner's side of the screen. Additionally, all the studies mentioned earlier regarding social contingency supporting learning from video chat (Troseth, Saylor,

& Archer, 2006; Roseberry, Hirsh-Pasek, & Golinkoff, 2014; Myers et al., 2017) required that children jointly attend to the words being taught by the video-chat companion to learn said words. So while video chat adds an element of complexity to social interaction, children are capable of navigating video chat to engage in joint attention, presumably as their proficiency with joint attention improves. Unfortunately, there is very little research on children's video chat behavior outside studies relating to social contingency and word learning from video chat—which is where the current study comes in.

Purpose of Study

Based on the literature, we believe that a superior ability to engage in joint attention and successfully communicate on video chat is likely predominately dependent on children's ToM understanding, specifically having sufficient understanding of knowledge access. Though, to our knowledge, no study has explicitly examined this phenomenon, it seems evident that without proper understanding of both perceptual and conceptual perspective-taking, children will experience failed joint attention engagement while on video chat.

The purpose of this study is to examine how well young children (ages 2- to 4.5-years-old) are able to transfer their understanding of knowledge access to the context of a video chat, effectively establish joint attention, and communicate with their video-chat companion. Specifically, this study will answer the question, are children who show greater ToM ability, better able to determine what they and their video-chat companion each have visual access to, in order to successfully navigate IJA and RJA during various trials? Additionally, we will examine whether their ToM understanding reflects how well they are able to verbally communicate their understanding of the visual limitations of

video chatting. Given the selected age group for this study, a large majority of the participants should have attained at least a L1P understanding (Moll & Tomasello, 2006), and a smaller majority should have attained a L1C understanding (Pillow, 1989). This smaller majority should be comprised of children who have an understanding of knowledge access, and their ability should be evident in the language they use in the form of simple deictic terms. Children who have an understanding of false beliefs should have a L2P understanding (Perner et al., 2002) and at least a stronger grasp on conceptual understanding, though not quite L2C (Taylor, 1988), as well. And each child's understanding should be reflected in the language they use in the form of more complex deictic terms (de Villiers & de Villiers, 1974).

First and foremost, based on our interpretation of the literature, we believe that greater ToM understanding will correlate to superior joint attentional capability (IJA and RJA) on video chat in informative-declarative situations involving referents that only one chat participant has visual access to. As an extension of this, we also believe that children who show an understanding of false belief will display higher IJA and RJA skill in situations with visual limitations brought on by video chat, and they will better communicate about those limitations than children who show only an understanding of knowledge access, who in turn, will display higher skill than children with lesser understandings of ToM. This general hypothesis and sub-hypothesis will be tested by evaluating each participants' performance on the Scaling of Theory of Mind Tasks (STM) from Wellman and Liu (2004), as well as by using a ToM evaluation survey completed by participants' parents. Participants' STM scores will then be compared against their performance on six different joint attention trials, which will require them to

communicate through gesture and verbalization. Out of these six trials, there will be two main trials of focus that involve joint attentional bids to referents that only one chat participant has visual access to. By evaluating the participants' performance on these two trials, in particular, as well as the other four trials, it should be evident whether greater ToM capability leads to a superior ability to transfer understanding of knowledge access to video chat—therefore, showing understanding of the limitations inherent to the video chat context—and engage in joint attention with a video-chat companion.

Methods

Participants

The participants for this study were 31 English-speaking children (18 females) between the ages of 2 years, 1 month and 4 years, 7 months (M= 3 years, 4 months). Participants were recruited through posts in parenting Facebook groups, word of mouth, and through an advertisement posted on ChildrenHelpingScience.com. Demographic information about the studies participants can be found in Table 2. Data from an additional 6 participants were removed from the study due to regular exposure to a second language, which has been shown to influence perspective taking ability (Liberman, Woodward, Keysar, & Kinzler, 2017).

Procedure

The experimental procedure took place via Zoom Video Conferencing. Each child participated in a single 10- to 15-minute video chat that consist of three activities: an introduction, the STM tasks, and six Joint Attention Trials, outlined in Table 3. Prior to the start of the call, the CSB Survey and Media Exposure Survey were collected form the participants' parents. At the start of the call, there was be a brief introduction and then participants were show a clip from the children's television show, Daniel Tiger's Neighborhood. Following this clip, participants were asked questions about what happened in the clip in order to get them talking and build rapport with the experimenter.

Scaling of Theory of Mind Tasks

Following the introduction, the seven STM tasks were presented in order of difficulty, starting with the easiest and increasing in complexity with each task. Previous studies have shown consistent developmental progress along this scale, such that if a

Table 2

Demographic breakdown of participants

Variables	n	Percentage (%)
Gender		
Male	13	42.0
Female	18	58.0
Age		
2-years-old	10	32.3
3-years-old	14	45.2
4-years-old	7	22.6
Race		
White	28	90.3
Black	2	6.5
Asian	1	3.2
Parents' Education		
Some College	1	3.2
Bachelor's Degree	16	51.6
Graduate Degree	13	41.9
Household SES		
Low	1	3.2
Medium	4	12.9
High	26	83.9

child passes a more complex task, they should be able to pass a less complex task of the scale as well (Wellman & Lui, 2004), however, given the relevance of knowledge access to the focus of this study, all participants went through at least the first three task (diverse desire, diverse belief, and knowledge access) regardless of any incorrect answers given to the questions for the first two scenarios. After that, if a wrong answer was given the activity was ended and the experimenter moved on to the joint attention trials. During these tasks, the experimenter made sure participants stayed focused and repeated any parts of the story that the child missed, due to lack of attention.

Table 3

Outline of procedure activities

Activity		Task
1. Introduction	i.	Clip of Daniel Tigers Neighborhood
2. STM Tasks	i.	Diverse Desires
	ii.	Diverse Beliefs
	iii.	Knowledge Access
	iv.	Contents False Belief
	v.	Explicit False Belief
	vi.	Belief Emotion
	vii.	Real-Apparent Emotion
3. Joint Attention Trials	i.	Through Screen Declarative Production (TSDP)
	ii.	Through Screen Declarative Comprehension (TSDC)
	iii.	Within Room Declarative Production (WRDP)
	iv.	Within Room Declarative Comprehension (WRDC)
	v.	Through Screen Imperative Production (TSIP)
	vi.	Through Screen Imperative Comprehension (TSIC)

Table 4

Breakdown of the different types of Joint Attention Trials

	Declarative	Imperative
Through Screen	TSDP[a] & TSDC[a]	TSIP & TSIC
Within-Room	WRDP[b] & WRDC[b]	N/a

[a] Trial involves an expressive declarative. [b] Trial involves an informative declarative

Joint Attention Trails

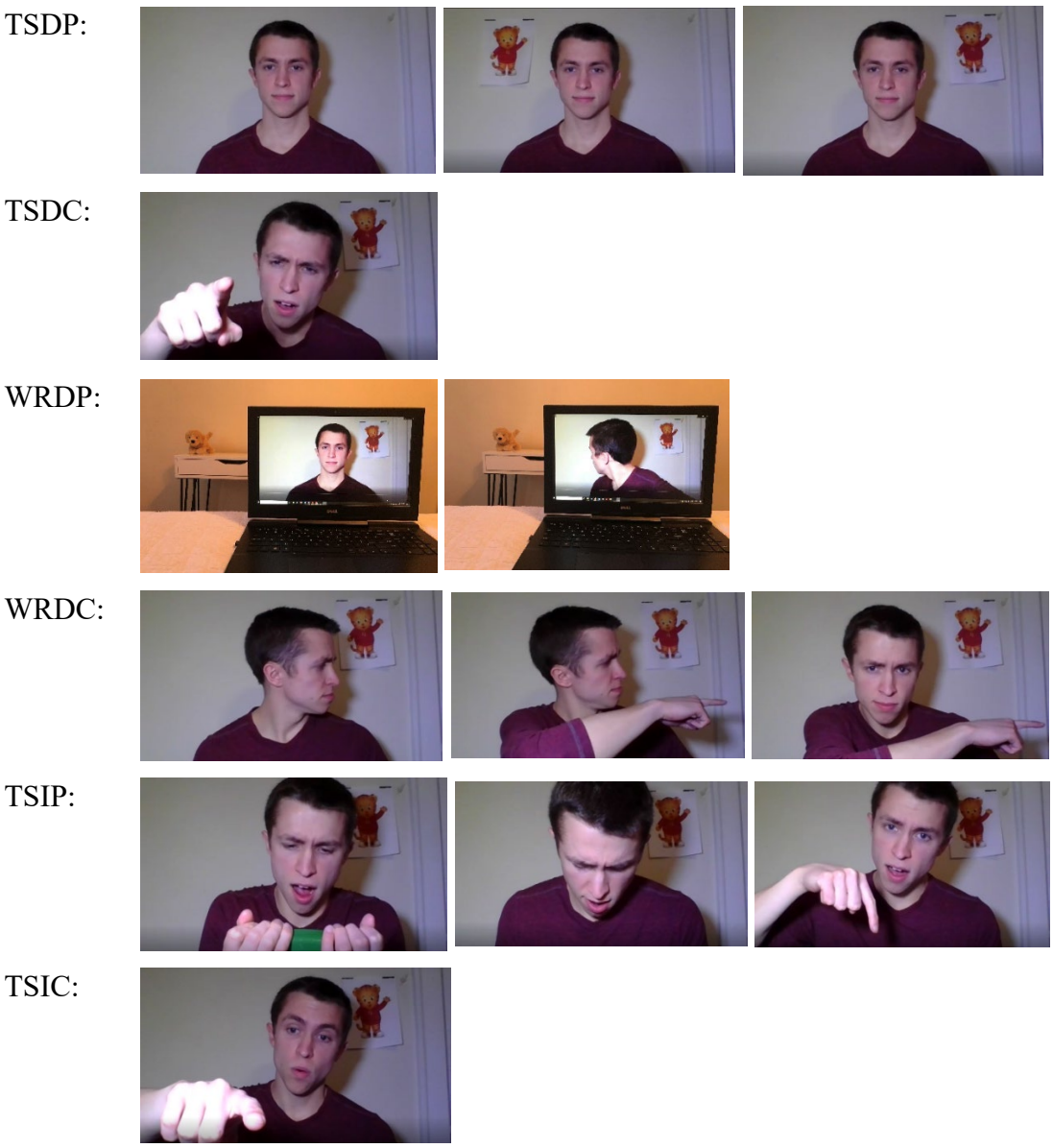
The participants were put through 6 different trial scenarios (see Table 3): the first four trials were of a declarative nature, such that two of the four trials required the child to produce a declarative: one expressive and one informative, in the form of a pointing gesture and/or verbalization, for IJA with the experimenter. The other two of the four

trials required the child to comprehend a declarative: one expressive and one informative, produced by the experimenter, for RJA with the experimenter. The last two trials were of an imperative nature such that one trial required the child to produce an imperative request, and the other trial required them to respond to an imperative request. Additionally, each trial varied the direction of the joint attentional bid being produced. Two of the declarative trials and the two imperative trials involved *through-screen* bids for joint attention, such that one imperative trial and one declarative trial required the child to produce a joint attentional bid through the computer screen, and one imperative trial and one declarative trial required them to comprehend the experimenters joint attentional bid through the screen. The other two declarative trials involved *within-room* bids for joint attention, such that one involved the child producing a bid for joint attention that referenced an item within their physical environment, and the other involved the child comprehending the experimenters bid for joint attention that referenced an item within the experimenter's physical environment. These two within-room trials were the main focus of this study given their complex visuospatial nature, with the simpler, through-screen, expressive-declarative trials acting as baselines to compare to for analysis. Note, there were no within-room, imperative trials because this could not be easily accomplished over video chat.

For the three trials that involved reference to items in the participants' physical environment, parental assistance was required. Prior to the video chat, parents were instructed via email to place a toy or object, such as a picture (referred to below as item 1), in plain view of the webcam, behind where the participant was sitting during the call. Next, they were asked to place a toy or object (referred to below as item 2) behind the

Figure 1

Images of each Joint Attention Trial from the participants' perspective



webcam, in plain view of the child, but out of reach from where they would be sitting. Then, they were asked to have a small toy handy during the call that they gave to their child when prompted by the experimenter. Lastly, they were asked to reply to the email to inform the experimenter of what objects they would be using for item 1 and item 2. To

follow, we outline each Joint Attention Trial. Images of each trail form the participants' perspective can be found in Figure 1.

Following conclusion of the STM tasks, the first joint attention trial was the Through-Screen, Declarative-Production (TSDP) trial. For this trial, the experimenter discretely used a pulley system, hidden off-screen, to move an image of Daniel Tiger along the wall, back and forth, behind the experimenter, in view of their webcam. As the image was moving the experimenter said, "Wait a second. I hear something, but I don't see anything. Is something happening?" The experimenter then continued to act confused and stare at the participant for approximately 10-15 seconds, giving them time to produce a pointing gesture and verbalization, before turning round to look at the picture.

Immediately following the TSDP, was the Through-Screen, Declarative-Comprehension (TSDC) trial. For this trial the experimenter stared at the screen, pointed through the screen, to the toy or object the parent had placed behind the child prior to the call and said, "What's that?" The experimenter continued to point for approximately 5 seconds before say, "What's that right there?" They then continued to point for approximately 10 more seconds giving the participant time to react by looking around the room and find the object.

Following the TSDC was the Within-Room, Declarative-Production (WRDP) trial. For this trial, the experimenter said, "I wonder what else is in the room with you. Do you see [name of item 2]? Where is it?" If the participant didn't look or find the item right way, the experimenter said, "Do you see it? Where is it?" If the participant then pointed, the experimenter turned around in the direction of the point to look behind themselves and say, "Where? I don't see it," while shifting his gaze back and forth

between the participant and the direction of their point. The experimenter did this for approximately 10 seconds, giving the child time to react and provide a verbal explanation that the experimenter could not see the item, then the experimenter said, “Oh, okay”. If the child did not point after finding the item, the experimenter tried prompt them further then waited 10 seconds before saying, “Oh, okay.”

Next was the Within-Room, Declarative-Comprehension (WRDC) trial. For this trial, the experimenter casually looked off to the side and then suddenly pointed to the side at an item off-screen (i.e. a stuffed giraffe), and said, “Wait, what’s that?” The experimenter then continued to point and say, “Do you know what that is?” while shifting his gaze back and forth between the item and the child. After the child gave a verbal explanation that they couldn’t see the item or after about 10 seconds, the experimenter reached off screen to grab the item and said “Oh, it’s my friend George the giraffe!” while holding it up to the screen for the child to see.

After that was the Through-Screen, Imperative-Production (TSIP) trial. For this trial, the experimenter looked at a new off-screen item (i.e. a little, green canoe), excitedly grabbed, bring it on-screen momentarily, and then held it in front of them below the view of the webcam, so the participant could not see the item, and said, “Whoa, look at this! This is so cool!” Then after 5 seconds of looking at the item the experimenter said, “Do you know what this is?” If the child verbally requests to see the item, the experimenter acted confused and said, “What? See what?” waited approximately 10 seconds before holding up the item to the camera and saying, “Oh, it’s a little green canoe.”

Last of all was the Through-Screen, Imperative-Comprehension (TSIC) trial. For this trial, the experimenter said, “I really wish you had a toy with you right now. Maybe [mom/dad] has one for you.” at which point the parent was directed to hand the participant a small toy. Once the toy was produced the experimenter acted very excited, and interested in the item. If the child did not hold up the toy, the experimenter then asked “Can I see it?” to see how the child would respond. After this, the participant was thanked, and the call was concluded.

Material and Measures

Participants’ ToM understandings were assessed using a modified version of the STM tasks, found in Wellman and Liu (2004). This assessment consisted of seven different scenarios involving various characters. Each scenario was designed to evaluate a particular element of a child’s ToM comprehension. The elements in order of increasing complexity were diverse desires: understanding that two people might have different desires about the same object, diverse beliefs: understanding that two people might have different beliefs about the same object, knowledge access: understanding that two people might not have access to the same information, contents false belief: understanding that someone might have a false belief about the contents of a container when the child knows the actual contents, explicit false belief: understanding that someone might act a certain way based off a mistaken belief, belief emotion: accurately judging how someone might feel based off a mistaken belief, and real-apparent emotion: understanding that someone can feel one way but look like they feel another way on the outside. Printed pictures, as well as props such as small boxes, toys, and stuffed animals were used to tell the seven different scenarios. For each scenario, the child was asked a comprehension or memory

questions about the scenario, along with the target question, to make sure they understood the story (Wellman & Lui, 2004). Participants were required to correctly answer both the target and memory/comprehension question in order to pass each level of the scale. They were then scored from 0-5 based off the highest level they passed, such that a score of 0 was given to individuals who did not pass any scenarios, and a score of 5 was given to individuals who passed the real-apparent emotion scenario. Each score corresponded to passing a single scenario with the exception of scoring a 4. Individuals were given a four for passing at least one of the contents false belief, explicit false belief, or belief emotion scenarios. This is because these three tasks have been shown to be comparably difficult for children to complete and therefore signify the approximate same level of ToM comprehension (Wellman & Lui, 2004).

A secondary assessment of ToM was obtained using a translated and abbreviated version of the Children's Social Adjustment Scale (EASE) parental questionnaire used by Comte-Gervais, Giron, Soares-Boucaud, & Poussin (2008). The original questionnaire consists of 50 questions: 25 questions seeking to understand the child's ability to attribute mental states to others (ToM related), for example, 'Are they able to lie to avoid being reprimanded, or to get something?' or 'Are they able to realize on their own that they have hurt someone?', and 25 questions focusing on gauging the child's adoption of standard social behavior, such as 'Do they know they shouldn't say bad words?' or 'Do they spontaneously use conventional courtesies (e.g. "please" and "thank you")?'. Given that we are only interested in ToM related behavior, we used only the 25 ToM-related question, which will be referred to as the Children's Social Behavior Survey (CSB) from here on. For each question, parents were instructed to indicate the frequency with which

their child displays the behavior: often (2), occasionally (1), or never (0). Each question was scored and totaled together to produce an overall CSB score, which ranges from 0-50.

In addition to the CSB survey, parents were given a Media Exposure Survey in order to assess each participants' experience with media devices, in general, and video chatting, in particular. Three variables, were drawn from responses to this survey: media exposure (ME) level, video chat exposure (VCE) level, and video chat handling proficiency (VCH). ME level was determined base off the number of hours a participant would spend streaming videos or playing on apps/ video games in a day. Less than 1 hour was consider low exposure, 1 to 2 hours was considered moderate exposure, and 2 or more hours was considered high exposure. VCE level was more complicated in that it was determined based off the age children first started using video chat, the frequency with which they use it, the duration they use it for, and the general involvement they show on the calls. Each participant was categorized as either no exposure, low exposure, moderate exposure, or high exposure. Frequency was weighted highest when scores, such that a child who was somewhat involved in short, daily calls was marked as high exposure, whereas a child who was somewhat involved in longer calls only once a weak or less was marked as moderate exposure. VCH was based off how often children would handle the phone while on a phone-based, video chat, and of those occasions in which they're handling the phone, how often they would point the camera at objects to show the other call participant. They were ranked between 0 and 6, with 0 being no display of proficiency and 6 being a high display.

For the joint attention trials, different objects and toys were used by the experimenter, such as a picture of Daniel Tiger, hanging on the wall, a small green canoe, and a stuffed giraffe. Other toys and object used in some of the trials were provided by participants' parents (see the following section for more information). An important component of some of these trials involved the child's ability to produce a pointing gesture, which was defined as an extension of the arm and hand that could involve either pointing of the index finger or whole-hand reaches that satisfy a relevant communicative intention for each trial.

Behavioral Coding

Following the completion of the video chat the recording was reviewed and coded. For each joint attention trail, the participant was evaluated on either their IJA or RJA skill while also effectively communicating. The two within-room trials were of particular interest because they also required participants to take into account the visual access and knowledge access of the experimenter. Because of the varying elements of focus in the trials, each trial had a unique coding scale which outlined for each trial in Table 4. The main focus of each trial was the following:

TSDP: The main focus of this trial was to evaluate the participants' IJA skill on video chat in a straight forward scenario. In evaluating performance on this trial we asked a two-part question, 1) does the participant attempt to produce a joint attentional bid, referencing the moving Daniel Tiger picture via a gesture and/or verbalization, in a social contingent and communicative manner (i.e. gaze shifting between the image and the experimenter, and/or providing synchronous verbal response), and, if so, 2) what deictic

Table 5

Behaviors and corresponding scores for each Joint Attention Trial

Trail	Score	Behavior
TSDP	0	Participant failed to produce a joint attentional bid
	1	Participant referenced Daniel Tiger through gesture or verbalization
	2	Participant used a simple spatial deictic term: *there*
	3	Participant used a more advanced spatial deictic term: *behind you*
	4	Participant said "behind you" (or some variation) within a complex construction
TSDC	0	Participant turned to search to either side of their self
	1	Participant searched on the screen for the referent
	2	Participant searched for the referent in front of or around their self, within relative view of their webcam
	3	Participant turned around to search
WRDP	0	Participant failed to communicate that joint visual attention was not possible
	1	Participant expressed a negative verbal reaction to seeing the experimenter turn
	2	Participant provided a simple descriptive location of the object
	3	Participant explained that the experimenter couldn't see the object without explicitly saying so
	4	Participant explicitly said "you can't see it".
WRDC	0	Participant turned to the side to search in their physical environment
	1	Participant remained watching the screen but didn't say anything
	2	Participant watched the screen and said "I can't see"
	3	Participant said "it's off screen", "out of view", or some alternative equivalent
TSIP	0	Participant didn't say anything
	1	Participant asked what the item was
	2	Participant said they can't see the item
	3	Participant explained to the experimenter that they were holding the object too low
TSIC	0	Participant failed to show the object
	1	Participant attempted to hold up the object for the experimenter

communicative skills do they utilize? In response to these questions they received a score between 0 and 4.

TSDC: The main focus of this trial was to evaluate the participants' RJA skill on video chat. The question this trial sought to answer was how does the participant perceive a point directed at the screen, with the intended reference of something behind them? In response to this question they received a score between 0 and 3.

WRDP: The intention of this trial was to be a more complex version of the TSDP, by evaluating the participants' ability to understand what the experimenter has visual access to when attempting IJA. Specifically, does the participant understand that the experimenter cannot see an off-screen object the participant tries to reference? In evaluating this trail, we again ask a two-part question, 1) does the participant successfully communicate in some way that joint *visual* attention cannot be attained, and, if so, 2) to what degree does the participant manage to communicate their understanding of visual access? In response to these questions they again received a score between 0 and 4.

WRDC: The intention of this trial was to be a more complex version of the TSDC, by evaluating the participants' ability to understand they don't have visual access to what the experimenter is referencing when attempting RJA. In order to evaluate this trial, we asked how does the participant react to the experimenter's gesture? In response to this question they again received a score between 0 and 3.

TSIP: The goal of this trial was to again see how well the participant could communicate that they don't have visual access to a clearly referenced object. For this, participants were given a score between 0 and 3. Additionally, though this is a production trial, it does involve RJA.

TSIC: The goal of this trial was to see if the participant would try to accommodate the experimenters visual access to an object. For this, participants were simply given either a 1 or 0, indicating success or failure. Though this is a comprehension trial, it does involve IJA.

Analysis

To test our hypothesis that greater ToM comprehension supports better transfer of knowledge access to video chat, Pearson's correlation coefficient was calculated between STM score and each joint attention trial score, as well as between trails, and between trials and variables from the Media Exposure Survey. We also grouped STM scores into broader levels of ToM: *low*, *moderate*, and *high*, in order to conduct ANOVA tests comparing differences across ToM groups. The low category consisted of children who did not pass any of the STM tasks higher than the diverse belief task (score of 2 or below). The moderate category consisted of children whose highest level pass was the knowledge access task (score of 3). And the high category consisted of children who pass the content false belief task or higher (score of 4 or 5). This was done in order to more specifically test our hypothesis that understanding of knowledge access and understanding of false belief are two important milestones that impact video chat competency. Additionally, for 7 out of the 31 participants the WRDP trial was not successfully conducted, largely due to accidental interference from parents or children immediately reaching to grab the referenced item to bring it on screen. For 2 other participants, the TSDC trial was not successfully conducted. The 7 participants were excluded from any analysis related to the WRDP trial and the 2 participants were

excluded from any analysis related to the TSDC trial, but all participants were included in all other analysis.

Results

General analysis

We first evaluated the strength of the relationship between the ToM measures— the STM and CSB scores— and the participants' ages. There were significant, moderate correlations between age and both STM and CSB scores ($r = 0.541$, $p < 0.005$ & $r = 0.538$, $p < 0.005$, respectively). We then tested the relationship between STM score and CSB, and, interestingly, there was no significant relationship.

Relationship of ToM and Joint Attention Trial Performance

Looking at how ToM related to performance on each of the 6 joint attention trials, all 3 of the production trials, the TSDP, WRDP, and TSIP, had moderate correlations ($r = 0.547$, $r = 0.683$, and $r = 0.500$, respectively) and significant relationships ($p < 0.005$, $p < 0.01$, and $p < 0.01$, respectively) to STM scores (see Table 5). Additionally, on the TSDP, all but 3 participants had successfully IJA (92%), on the WRDP, only 7 participants successfully communicated about joint attention (28%), and, on the TSIP, 20 participants successfully communicated to the experimenter (65%). On the other hand, out of the other 3 comprehension trials, TSDC and WRDC had no significant relationship to STM score. Neither did performance on the TSIC, however, all but 1 participant successfully completed this trial. On the WRDC, 17 participants successfully maintained their gaze on the computer screen (55%). Looking then at CBS scores, the TSDC trial was the only trial to show a significant, weak correlations ($r = 0.494$, $p < 0.01$). The rest of the trials showed no significant relationship.

Next, we conducted one-factor ANOVA tests, comparing ToM level: *low*, *moderate*, or *high*, as the factor variable to see if there were significant differences in

Table 6

Correlation of performance on joint attention trials to STM and CSB scores

	TDSP	TSDC	WRDP	WRDC	TSIP	TSIC
STM	0.55***	0.21	0.68***	0.11	0.49**	0.01
CSB	0.34	0.49**	0.07	0.09	0.31	0.04

$^{**}p < 0.01$, $^{***}p < 0.005$

group performances for each of the trials (see Table 6 & Table 7). On any trials for which the ANOVA produced significant findings, we additionally conducted Levene's test to check for homogeneity of variance in order to assess the validity of the ANOVA test, before then conducting a Tukey multiple pairwise-comparisons of means between each level. For the TSDP, the ANOVA showed a significant difference between groups ($p < 0.01$), which was validated by Levene's test showing no significant difference in variance across groups. The Tukey comparison then showed a significant difference between only the low (M= 1.78, SD = 1.31) and high (M= 3.71, SD = 0.488) groups ($p <0.005$). For the TSDC, no significant difference between ToM levels was found. For the WRDP, the ANOVA showed a significant difference between groups ($p < 0.001$), which was validated by Levene's test showing no significant difference in variance across groups. The Tukey comparison then showed a significant difference between only the low (M = 0.25, SD = 0.683) and high (M= 2.6, SD = 1.52) groups ($p < 0.01$). For the WRDC, the ANOVA showed a significant difference between groups ($p < 0.05$), which was validated by Levene's test showing no significant difference in variance across groups. The Tukey comparison then showed a significant difference between only the moderate (M = 1.33, SD = 0.816) and high (M= 0.286, SD = 0.488) groups ($p < 0.05$), however, it's important to call attention to the fact that the moderate group's score was high than the high group,

Table 7

Average score for each production joint attention trial grouped by level of ToM

Group	TSDP			WRDP			TSIP		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Low	18	1.79	1.31	16	0.25	0.683	18	0.556	0.616
Moderate	6	2.17	1.47	3	1	1.73	6	1.17	0.753
High	7	3.71	0.488	5	2.6	1.52	7	1.57	0.787

Table 8

Average score for each comprehension joint attention trial grouped by level of ToM

Group	TSDC			WRDC			TSIC		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Low	16	1.44	1.26	18	0.611	0.608	18	1.78	0.548
Moderate	6	2.33	0.816	6	1.33	0.816	6	1.67	0.516
High	7	1.71	0.951	7	0.286	0.488	7	1.86	0.378

as well as the low group (though not significant). For the TSIP, the ANOVA showed a significant difference between groups ($p < 0.01$), which was validated by Levene's test showing no significant difference in variance across groups. The Tukey comparison then showed a significant difference between only the low ($M = 0.556$, $SD = 0.616$) and high ($M = 1.57$, $SD = 0.787$) groups ($p < 0.01$). Lastly, for the TSIC, no significant difference between ToM levels was found.

Intra-trial Relationships

We then did basic comparisons of performances between trials to see if there were any significant correlations. Again, only significant relationships were seen for the 3 production trials: we found a moderate correlation between the WRDP and TSIP ($r =$

0.522, $p < 0.01$), a moderate correlation between TSDP and WRDP ($r = 0.598$, $p < 0.01$), and a moderate correlation between TSDP and TSIP ($r = 0.532$, $p < 0.005$).

Relationship of other variables

We next wanted to look for relationships between the variables collected from the Media Exposure Survey—VCE, ME, and VCH—and joint attention trial performance. First looking at VCE, surprisingly, we found no significant relationship to performance on any of the trials. Next looking at ME, we found no significant relationship to performance on the TSDP, TSDC, WRDC, TSIP, nor TSIC. There was a significant, weak negative correlation to WRDP performance ($r = 0.4119$, $p < 0.05$). With this finding, we then modeled STM and ME together in a linear model, assuming interaction between the variables, and found a significant, moderately-strong correlation to WRDP performance (adjusted $r = 0.761$, $p < 0.0005$). Lastly, looking at handling proficiency, we found no significant relationship to performance on the TSDP, WRDP, WRDC, nor TSIC. There was a significant, weak correlation to TSDC performance ($r = 0.458$, $p < 0.05$), and a significant, moderate correlation to TSIP performance ($r = 0.537$, $p < 0.005$). With this second finding, we then modeled STM and VCH proficiency together in a linear model, assuming interaction between the variables, and found a significant, moderately-strong correlation to TSIP performance (adjusted $r = 0.701$, $p < 0.0001$).

Discussion

In this study, we sought to evaluate young children's ability to transfer understanding of knowledge access to a video-chat context in order to participate in joint attention, based on their level of ToM capability. For our general hypothesis, we proposed that ToM would be positively correlated to informative-declarative episodes of both IJA and RJA on video chat. We found this to be true only for IJA in the WRDP trial, partially confirming our hypothesis. For the WRDP trial, STM score significantly correlated with performance. However, we found no significant relationship between STM score and RJA in the WRDC trial.

For our sub-hypothesis, we proposed there would be group-performance differences between children with low ToM (STM scores of 2 and below), children with moderate ToM (those who had attained an understanding of knowledge access; STM scores of 3), and children with high ToM (those who had attained an understanding of false belief or greater; STM scores of 4 and above), with each successive group performing better than the last. We found the high-ToM group did perform significantly better than the low-ToM group in the WRDP trail. The mean score for the moderate group fell in between the low and high groups, but was not found to be significantly different from either group. This lack of significance may be the result of our small sample size of individuals in the moderate (3 participants) and high groups (5 participants)—a majority of participants were in the low group (16 participants)—as well as overall sample size. We believe a larger more equally distributed sample of participants will provide even more promising results. However, while these current findings do not entirely affirm our belief that trial performance follows a clear ToM-

progression, it does suggest that ToM is an important factor that correlates to children's proficiency in IJA skill for visuospatially complex, video-chat scenarios, such as in the WRDP trail.

Tying in perceptual and conceptual perspective-taking, our results suggest that attaining L1C understanding, bolstered by L1P understanding, as well as false belief understanding, is what creates distinction in IJA performance on video chat. Children in the low-ToM group should have attained, or be in the process of attaining, L1P understanding (Moll & Tomasello, 2006), but have yet to attain L1C or false belief understanding. Children in the moderate-ToM group are a step above in that they should have attained a L1C understanding, but not yet a false belief understanding (Pillow, 1989). And children in the high-ToM group should have attained both L1C and false belief understanding. Breaking down the WRDP trial, what's required of the participant, first, is that they determine that the experimenter cannot see the item of reference (use of L1P), and therefore, the experimenter does not have access to knowledge of the items location (use of L1C). Then the child must, at a minimum, identify that the experimenter misinterpreted their gesture, or, more productively, identify that despite not having access to the item (reality), the experimenter has a false understanding that he could turn to look in the direction of the child's gesture (a false belief). After determining all of this, the child then must communicate information about the item to the experimenter, hopefully correcting the experimenter's false understanding in the process. It's not necessary for children to identify the experimenter's false belief to succeed on this trial—evident from the handful of lower-ToM participants who were successful—but we argue it is a more constructive route to success that higher-ToM participants likely use, based on their more

detailed explanations. Thus, through this interpretation of our result, we see this progression of skills at work for video-chat IJA, but so far that's only half the picture.

Looking at group-performance differences for the WRDC trial, we found that the moderate-ToM group performed significantly better than the high-ToM group. This was a very interesting and unexpected result, and while it appears inexplicable at first glance, it outlines a difference between IJA and RJA for children, that we'll dive in to now.

IJA v. RJA

Expanding our view to first look at all six joint attention trials, only the 3 production trials: the TSDP, WRDP, and TSIP, showed significant relationships to ToM (as measured by the STM tasks), which were all of moderate correlation. Additionally, in comparing performances between trials, again, only performance in the production trials showed significant correlations to one another, whereas, the comprehension trails showed no significant relationships to any other trials. We did find it interesting that the TSIP and WRDP performances had a significant relationship, yet the TSIP and WRDC performances did not. We say this because although the TSIP trial involves provoking the child to produce an imperative, the experimenter does so by producing a expressive-declarative in reference to an item off screen, therefore, making it an RJA situation that is conceptually similar to the WRDC trial. However, one major difference is that in the TSIP trial, the object of reference is intentionally shown very briefly on screen, so it is clear to the child that the object being referenced is in the room with the experimenter, which is not the case in the WRDC. This difference is important because it gives the lack of significant relationship between the TSIP and WRDC, significance in and of itself because it suggests that poor performance on the WRDC, and potentially the other

comprehension tasks, was likely not the result of an inability to conceptualize the experimenter's unique, visual access, but perhaps instead a failure to take it in to account. It does, however, still remain curious that children managed to avoid this mistake in production tasks.

The curiosity only grows when we more closely compared the WRDP and WRDC trials. As mentioned, there was no significant relationship between these trials, nor were we able to find a significant relationship performing a two-sample t-test comparing WRDC scores against children who were successful in the WRDP vs. those who failed. One possible source of uncertainty in this study might have been due to the juxtaposition of these two trials, or the lack of trial-order randomization overall. This may have more of an impact on the STM evaluation, which we'll talk more about later, but it could have influenced performance on the WRDC, which took place immediately following the WRDP. In the WRDP, the participant's focus is drawn away from the computer screen, and into the surrounding room. Then, a moment later, in the WRDC, focus is shifted through the screen, into the experimenter's room. This may have been a drastic shift for some children to make in that moment, thus making the WRDC more challenging; we believe, however, that this is not a confounding element of this trial, but instead a central element. The ordering of these six joint attention trials we used was intended to mimic a naturally flowing conversation as opposed to individual, structured tests. We concluded, that to best evaluate the participants, this emulation of natural discourse was more important than randomizing trial order. The attention shift between the WRDP and the WRDC is a very plausible situation to occur on video chat, making this an even more applicable test of the child's video chat, communication proficiency. Additionally, in the

WRDC, the experimenter very intentionally turned his head 90 degrees to his left before pointing, and then looking back to the child. He then continued to shift his gaze between the referent to his left and the participant. Thus, for the child to successfully engage in joint attention with the experimenter for this trial, he or she must understand that the experimenter is clearly looking to the side; however, in doing so, the experimenter is not capable of looking in the participant's room outside the view of the webcam, and therefore, he must be looking at something in his own room.

Based on the results for these two trials, it appears that the WRDC was an easier trial than the WRDP because out of the 24 participants who completed the WRDP trial, only 7 were successful, whereas 13 out of the 24 participants were successful in the WRDC. Only 3 of the children who passed the WRDP trail, also passed the WRDC trial, which is not unexpected given that there was no significant relationship found between performance on these trials. However, what is interesting is that 3 out of the 4 children who passed the WRDP but not the WRDC, were part of the high-ToM group. It's surprising that these 3 children, who not only passed, but performed well on the WRDP (scoring 3 or higher)—thus showing a strong understanding of the visual limitations of video chat, accurately accounting for what the experimenter cannot see—then immediately made the exact opposite mistake by not accounting for the experimenter's visual access, and assuming the experimenter could see and reference a different object in the room with the participant. It is largely because of these 3 children's performances that the high-ToM group performed worse as a group on the WRDC than the moderate-ToM group: only 2 out of the 7 high-ToM group children passed the WRDC, whereas 5 out of the 6 moderate-ToM group children and 10 out of 18 low-ToM group children passed as

well. From comparing these two trials, it's clear that, despite both trials relying on the participant to account for the experimenter's visual access, there is some sort of disconnect between the variables that influence performance on each of these trials. A potential culprit for the missing variable is executive function.

Executive Function and Joint Attention

Executive function is one's own ability to control their thoughts and behaviors through inhibition of reflexive response in order to plan actions, execute those actions as planned, and detect errors or self-correct as needed (Zelazo, Carter, Reznick, & Frye, 1997). Moses (2001) argues that executive function likely affects both the emergence and expression of ToM understanding. To attain belief understanding, children must have at least a minimal ability to distance themselves from immediate stimuli in order to reflect on their thoughts and actions, while inhibiting either misleading information or reality (in the case of false belief). Expression of ToM, on the other hand, involves one's ability to override reflexive reference to reality and one's own knowledge. Thus, failure on ToM tasks, such as false belief, may not reflect a child's actual understanding, but may reflect a failure to properly translate that understanding into successful action (Carlson, Moses, & Hix, 1998). A central component of executive function is inhibitory control (IC): the ability to inhibit reflexive response to extraneous stimuli in order to perform an intended action (Rothbart & Posner, 1985). Pointing to the true location of an object is a highly practiced and reinforced response for young children from activities such as picture book reading. Therefore, overcoming this habitualized response in false belief tasks requires sufficient IC (Carlson et al., 1998). We believe it's not too much of a stretch to then argue that perhaps gaze following, and RJA becomes habitualized in the same way such that, in

the case of the WRDC trail, children need to inhibit the desire to respond to the joint attention bid by turning to look in their own room, instead maintaining their focus on the computer screen.

IC develops over the first six years of life, undergoing large improvements from the age of 3- to 6-years-old (Diamond & Taylor, 1996). This timeline heavily overlaps with children's development of more advanced ToM understanding, suggesting that the two might be related, which appears to be the case. Carlson and Moses (2001) found that IC and ToM share a significant correlation, suggesting that development of IC plays a crucial role in facilitating the emergence and expression of ToM understanding in children. Based on this, children with high ToM should likely also have high IC—which is an assumption we were operating under for this study—though it's not a necessary condition. Thus, in comparing WRDP performance to the WRDC performance, there seems to be a clear lapse in IC taking place for the 3 high-ToM individuals who passed the WRDP, yet failed the WRDC, and potentially others. Why is that? It likely isn't because of the attention shift taking place between the two trials, mentioned earlier, since younger, lower-ToM proficient children, who presumable also have low IC, managed to succeed on the WRDC trial—this was the case for 10 out of 18 participants, only 1 of which succeeded on the WRDP. But this raises another question, why did so many low-ToM participants pass the WRDC, but fail the WRDP? We wonder if perhaps there were imperfections in the evaluation criteria. One possible issue is that evaluation of the WRDP was much stricter than the WRDC, in that it requires a clear communication of understanding, whereas evaluation of the WRDC was a little more subjective and less reliant on children's vocalizations. Thus, a confounding element we did not foresee is

that there likely were some false positives in the WRDC results. We coded all participants who maintained focus on the computer screen as successes; however, it's possible that some children interpreted the experimenters gesture as referencing some aspect of the computer screen (i.e. the Zoom window, or the edge of the screen itself), which is inappropriate in this situation.

For this reason, we decided to do a post-hoc review and recode for the WRDC trial, this time using TSDC performance as reference, for instances when the child maintained focus on the screen but did not say anything. If they received a 2 or 3 on the TSDC, signifying that they had interpreted the experimenter's gesture as coming out of the screen, rather than being directed to the screen, they were marked as a success. We assume that if participants in this category did not understand the limitation of video chat, they would have turned to search within their environment rather than watch the screen. In addition to these, we then also marked participants who expressly communicated their inability to see the experimenters reference as successes. Doing this cut the number of successes down from 17 participants to 12. After reanalyzing the data, there was still a significant difference between moderate- and high-ToM groups for the WRDC ($p < 0.01$) and the only change was a significant, moderate correlation between TSDC and WRDC ($r = 0.547$, $p < 0.01$); however, this is likely given that we used TSDC performance to code the WRDC. Importantly though, the 5 out of the 17 participants who were no longer marked as successes, were all low-ToM individuals. So we believe there were likely some false positives. Additional, this recoding did take the number of low-ToM participants who passed the WRDC down from 10 to 6, 1 of which succeeded on the WRDP. Yet, despite this change, 6 is still a third of all the low-ToM participants, whose

successful WRDC performance is unexpected based on their assumed IC. This is particularly true in light of the 3 high-ToM individuals we've been discussing. Therefore, we suspect it is possible these 6 low-ToM participants had surprisingly high IC, and the 3 high-ToM participants had surprisingly low IC. For these reasons, future studies regarding children's joint attention on video likely would benefit from evaluating or controlling IC through the use of a battery such as the one used in Carlson and Moses (2001), possibly along with evaluation of executive control in general. Doing this might help explain variation in results such as those we see with the 6 low-ToM and 3 high-ToM participants.

On top of concerns regarding IC, it's important to point out that IJA (used in the WRDP) and RJA (used in the WRDC) do vary in the internal regulation system that control them as a result of their varied relationship with executive control (Mundy & Van Hecke, 2008). Mundy, Card, and Fox (2000) suggests that RJA and IJA have many common, but also some distinct neurophysiological correlation, with RJA largely being regulated by temporal and parietal systems involved in attention disengagement and orientation, whereas, IJA is regulated by frontal, temporal, and parietal systems involved with goal-directed social behavior, working memory, and dual-task processing. So while success on both of the WRDP and WRDC rely on a similar understanding of the visual limitations inherent to video chat, different neurological systems regulate expression of this understanding, thus potentially adding to variety in performance results. On top of evaluating IC, incorporating the use of EEG could be incredibly beneficial in deciphering how children interpret IJA and RJA situations on video chat by comparing brain activation data to non-video chat IJA and RJA episodes.

Media Related Variables

Beyond ToM, we foresaw that a possible confounding variable affecting participants' performances might be prior experience with video chat and other forms of media. Children who have a greater exposure to video chatting, thus receive more practice communicating on video chat, which we assumed should lead those children to be more adept at video chat communication in general, and perhaps also be more skilled at navigating complex joint attention scenarios. This was an idea shared by McClure et al. (2018), who suggested that video chat is a unique joint attentional experience that offers complex perspective-taking scenarios that may alter children's developmental trajectories. However, to our surprise, this did not seem to be the case based on our data. The only significant relationship we found regarding video chat exposure was to TSIP performance, which had only a weak correlation.

Alternatively, we also looked at media exposure and found a significant negative correlation between media exposure and WRDP performance. It was only a weak correlation, but after combining STM score and media exposure into a linear model, we found a stronger correlation to WRDP score. Exposure to television has been shown to be negatively correlated to ToM understanding in preschool aged children (Nathanson, Sharp, Aladé, Rasmussen, & Christy, 2013). Looking at our data, we did see a negative correlation between STM and media exposure, but it was not significant. Regardless, it was interesting to find this correlation in WRDP performance. We wonder if, as an extension of Barr's (2013) third factor, due to the symbolic nature of screens, perhaps children with greater media exposure have greater difficult dissociating the experimenter, as a 3D being who is represented in a 2D display, from the 2D display itself. Therefore,

when the experimenter turns, this action is inaccurately interpreted as the experimenter looking behind the participant's computer screen. Or does greater exposure to non-contingent video interfere with a child's ability to assign mental states to the experimenter on the call. As mentioned, because of the lack of social contingency, pre-recorded video is interpreted by children as a less reliable and inaccurate source of information (Roseberry et al., 2014). Perhaps, this distrust in non-contingent media can bleed over to video chat for children with high media exposure.

A third variable we analyzed was children video chat handling proficiency, for which we did find a significant relationship to both TSDC and TSIP performance. In handling the phone while on a video chat, we would assume children likely gain a better understanding for what is in view of their webcam. This might account for the correlation to TSDC, in which performance would benefit from the child having a greater awareness of the visual access to the area around behind them that the experimenter has. However, following this logic, we would expect there to be correlation to other trials such as the WRDP and WRDC. We found only a weak correlation, so we are uncertain of the effect the relationship has. For TSIP, we assume that greater handling proficiency gives children a better understanding of how the view of objects can be manipulated either by being brought on or taken off screen or by moving the camera. Therefore, perhaps they are better able to communicate to the experimenter a need for the object to be brought in to view.

Based on these findings relating to media based variables, it would be worthwhile for future video-chat related studies to take a more comprehensive dive into analyzing the effect of different types of media exposure on children's behavior. The depth of the

Media Exposure Survey we use in this current study was not extensive. In the future, use of a long, more detailed survey or having parent log media behavior for a handful of weeks, would provide superior data. A future study focusing on media base variables could done in the form of a longitudinal study observing how children's exposure to media and video chat affect development of future video chat joint attentional ability (such as IJA and RJA), alongside ToM.

Inconsistency between ToM Evaluations

As mentioned, we did not randomize trail-order across participants, which may have impact participants' performance on the STM tasks. For instance, it was very common for children to be shy and quiet at the start of the call, but gradually open up and get more engaged. Because of this shyness, it's possible that children performed worse on the easier beginning STM tasks when they were capable of answering them correctly, thus misrepresenting their actual ToM understanding. A cartoon of Daniel Tiger was shown at the start of the video and then discussed to help alleviate this shyness, but it likely would have been helpful to include more warmup activities. Fortunately, some of this uncertainty is alleviated in analyzing performance based on the low, moderate, and high grouping, since every child was put through at least the first three STM tasks. Because of this, the first two tasks effectively acted as further warm up, in preparation for the knowledge access task, such that a child who understood knowledge access still had the opportunity to then to show their understanding, despite potentially having gotten the first two trials wrong. Any child who did not yet have understanding of knowledge access, would be bucketed in the Low-ToM group regardless of whether or not they needed the first two trials to warm up more.

Aside from task ordering, another possible element of uncertainty may have to do with the reliability of our ToM evaluation. Comparing STM score to CSB, we surprisingly did not find a significant correlation, suggesting that one or both measures were inaccurate. Looking at STM scores, it's possible video chat added additional mental load to participants, thus affecting their scores. Though video chat is resistant to the effects of the 'video deficit', it less unyielding to the effects of the 'transfer deficit' that Barr (2013) spoke of. As outlined in two of the factors Barr mentions, because of its lack of perceptual stimulation and the symbolic nature of screens, video chat interpretation places an additional mental load on children participating in a social interaction via video chat. This additional mental load could have compromised STM performance, and led participants to produce scores indicating a lower ToM then they actually had. However, the CSB score itself is not necessarily a perfect measure either, as it relies on parental-reporting, which is subject to bias. Additionally, both measures show significant correlation to age, as is expected of ToM understanding (Tomasello et al., 1993), so it's unclear which score is the more accurate indicator of actual ToM understanding. Therefore, we chose to trust in the STM score as it was evaluated under our supervision.

Technical Limitations

There were some technical issues encountered while conducting the study. Frequently while review the recordings, the researcher's audio can be heard cutting in and out, such that words are sometimes truncated or cut completely. Additionally, this challenge was present on the participants' end, requiring the researcher to ask children to repeat themselves or ask the parent for confirmation on their response. In addition,

children sometimes spoke too quietly for their mics to pick up the speech, which again led the research to seek confirmation on responses.

Aside from audio issues, another issue with conducting this study virtually was that we had less control of some elements of the study and had to rely on parents to help prepare for the study through setting up items. In several case, parents had to be reminded to have the objects in place. And once or twice, an item was not where the parent said it would be, such as the item intended to be behind the child (item 1), forcing the experimenter to instead point to something else behind the child or nothing in particular. Additionally, there were a few occurrences where item 2 was placed too close to the child, allowing them to grab it, interfering with the WRDP trial. As well as another occurrence or two where the parent actually grabbed the toy for the child and handed it to them. Conducting this study in a controlled laboratory environment would have removed many of these challenges.

Implications for Future Study

Another important recommendation for future studies, beyond those already suggested, is using hybrid testing, that is, using both in-person and video chat testing, in order to compare children's normal behavior to their behavior on video chat. For instance, aside from alleviating challenges with reliance on parent, in-person testing would allow for a more comprehensive evaluation of ToM while also removing any of the uncertainties present in this study regarding ToM evaluation. It would also more easily allow for testing children's verbal ability by using an assessment, such as the Peabody Picture Vocabulary Test (PPVT-R: Dunn & Dunn, 1981). Additionally, the Early Social Communication Scales (ESCS) from Mundy et al. (2003) could be

conducted in-person to get baselines of joint attention behavior and then used to compare against behavior displayed on video chat.

Additionally, something we did not talk about during our earlier discussion on IC, bilingual children have been shown to have superior IC to monolingual children of the same age (Kapa & Colombo, 2013) and infants who received multilingual exposure showed better perspective taking than monolingual children (Liberman et al., 2017). In the current study, language exposure was controlled; however, there is a plethora of options for future studies on children and video chat involving bilingualism. For instance, performance on tasks in this study could be compared between bilinguals and monolingual or bilinguals of varying proficiency. Or alternatively, it would be interesting to compare performance on the joint attention trials across languages. Liu, Wellman, Tardif, and Sabbagh (2008) found that for children from different language and cultural backgrounds, ToM follows the same developmental progression, but on different timetables. It would be interesting to compare the performance of individuals who have the same level of ToM understand but have different language and cultural backgrounds.

Conclusion

In conclusion, this study evaluated children's IJA and RJA ability on video chat based on their ToM. Our results indicate that ToM supports IJA, but not RJA, when on video chat, and that specifically false belief understanding and L1P are likely important factors that influence performance. These findings suggest that other variables, potentially related to IC and executive control overall, more strongly impact RJA capability on video chat, whereas variables, such as media exposure, have influence of video-chat related IJA. These difference between IJA and RJA potentially stem from

variation in neurological regulation. This possibility, as well as other mention here, should be examined further in future studies.

References

American Academy of Pediatrics, Media use by children younger than 2 years, Pediatrics 128 (2011) 1040–1045.

American Academy of Pediatrics: Council on Communications and Media. (2016). Media and young minds. Pediatrics, 138, e20162591. doi: 10.1542/peds.2016-2591

Anderson, D. R., Huston, A. C., Schmitt, K. L., Linebarger, D. L., Wright, J. C., & Larson, R. (2001). Early childhood television viewing and adolescent behavior: The recontact study. Monographs of the society for Research in Child Development, i-154.

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

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. Psychological Science, 15, 415–419.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. Natural theories of mind: Evolution, development and simulation of everyday mindreading, 1, 233-251.

Bates, E. (1976). Language and context: The acquisition of pragmatics. Academic Press.

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. (2013). Memory constraints on infant learning from picture books, television, and touchscreens. Child Development Perspectives, 7(4), 205-210

Brito, N., Barr, R., McIntyre, P., & Simcock, G. (2012). Long-term transfer of learning from books and video during toddlerhood. Journal of experimental child psychology, 111(1), 108-119.

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. British journal of developmental psychology, 9(1), 55-72.

Butterworth, G. (2001). Chapter Eight Joint Visual Attention in Infancy. Blackwell handbook of infant development, 213.

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. Child development, 69(3), 672-691.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. Child development, 72(4), 1032-1053.

Colonnesi, C., Rieffe, C., Koops, W., & Perucchini, P. (2008). Precursors of a theory of mind: A longitudinal study. British Journal of Developmental Psychology, 26(4), 561-577.

Colonnesi, C., Stams, G. J. J., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. Developmental Review, 30(4), 352-366.

Cochet, H., Jover, M., Rizzo, C., & Vauclair, J. (2017). Relationships between declarative pointing and theory of mind abilities in 3-to 4-year-olds. European Journal of Developmental Psychology, 14(3), 324-336.

Comte-Gervais, I., Giron, A., Soares-Boucaud, I., & Poussin, G. (2008). Evaluation de l'intelligence sociale chez l'enfant. L'information psychiatrique, 84(7), 667-673.

Delgado, C. E., Mundy, P., Crowson, M., Markus, J., Yale, M., & Schwartz, H. (2002). Responding to joint attention and language development.

de Villiers, P. A., & de Villiers, J. G. (1974). On this, that, and the other: Nonegocentrism in very young children. Journal of Experimental Child Psychology, 18(3), 438-447.

de Villiers, J. G., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. Children's reasoning and the mind, 189, 226.

de Villiers, J. (2007). The interface of language and theory of mind. Lingua, 117(11), 1858-1878.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". Developmental psychobiology, 29(4), 315-334.

Dunn, L. M., & Dunn, L. M. (1981). Peabody picture vocabulary test-revised. American guidance service, Incorporated.

Dunn, J., & Cutting, A. L. (1999). Understanding others, and individual differences in friendship interactions in young children. Social development, 8(2), 201-219.

Hains, S. M., & Muir, D. W. (1996). Effects of stimulus contingency in infant-adult interactions. Infant Behavior and Development, 19(1), 49-61.

Flavell, J. H. (1978). The development of knowledge about visual perception. In Nebraska symposium on motivation. University of Nebraska Press.

Flavell, J. H., Shipstead, S. G., & Croft, K. (1978). Young children's knowledge about visual perception: Hiding objects from others. Child Development, 1208-1211.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. Developmental psychology, 17(1), 99.

Flavell, J. H., Beilin, H., & Pufall, P. (1992). Perspectives on perspective taking (pp. 107-139). Hillsdale, NJ: Erlbaum.

Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. Child development, 567-582.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. Psychological science, 16(5), 367-371.

Kapa, L. L., & Colombo, J. (2013). Attentional control in early and later bilingual children. Cognitive development, 28(3), 233-246.

Kristen, S., Sodian, B., Thoermer, C., & Perst, H. (2011). Infants' joint attention skills predict toddlers' emerging mental state language. Developmental psychology, 47(5), 1207.

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, 100(15), 9096-9101.

Liberman, Z., Woodward, A. L., Keysar, B., & Kinzler, K. D. (2017). Exposure to multiple languages enhances communication skills in infancy. Developmental science, 20(1), e12420.

Linebarger, D. L., & Walker, D. (2005). Infants' and toddlers' television viewing and language outcomes. American behavioral scientist, 48(5), 624-645.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. Developmental psychology, 44(2), 523.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. Child development, 74(4), 1130-1144.

McClure, E. R., Chentsova-Dutton, Y. E., Barr, R. F., Holochwost, S. J., & Parrott, W. G. (2015). "Facetime doesn't count": video chat as an exception to media restrictions for infants and toddlers. International Journal of Child-Computer Interaction, 6, 1-6.

McClure, E., & Barr, R. (2017). Building family relationships from a distance: Supporting connections with babies and toddlers using video and video chat. In Media Exposure During Infancy and Early Childhood (pp. 227-248). Springer, Cham.

McClure, E. R., Chentsova-Dutton, Y. E., Holochwost, S. J., Parrott, W. G., & Barr, R. (2018). Look at that! Video chat and joint visual attention development among babies and toddlers. Child Development, 89(1), 27-36.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. Child development, 78(2), 622-646.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. British Journal of Developmental Psychology, 24(3), 603-613.

Moses, L. J. (2001). Executive accounts of theory-of-mind development. Child development, 72(3), 688-690.

Mundy, P., Card, J., & Fox, N. (2000). EEG correlates of the development of infant joint attention skills. Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 36(4), 325-338.

Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A., & Seibert, J. (2003). Early social communication scales (ESCS). Coral Gables, FL: University of Miami.

Mundy, P., & Van Hecke, A. (2008). Neural systems, gaze following, and the development of joint attention.

Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2017). Baby FaceTime: Can toddlers learn from online video chat? Developmental Science, 20(4), e12430.

Nathanson, A. I., Sharp, M. L., Aladé, F., Rasmussen, E. E., & Christy, K. (2013). The relation between television exposure and theory of mind among preschoolers. Journal of Communication, 63(6), 1088-1108.

Parkinson, B., & Lea, M. (2011). Video-linking emotions. In A. Kappas & N. C. Kraemer (Eds.), Face-to-face communication over the Internet: Emotions in a web of culture, language and technology (pp. 100–126). Cambridge: Cambridge University Press.

Perner, J., Stummer, S., Sprung, M., & Doherty, M. (2002). Theory of mind finds its Piagetian perspective: Why alternative naming comes with understanding belief. Cognitive development, 17(3-4), 1451-1472.

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. Journal of experimental child psychology, 47(1), 116-129.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14- and 18-month-olds. Developmental psychology, 33(1), 12.

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. Child development, 85(3), 956-970.

Rothbart, M. K., & Posner, M. I. (1985). Temperament and the development of self-regulation. In The neuropsychology of individual differences (pp. 93-123). Springer, Boston, MA.

Salo, V. C., Rowe, M. L., & Reeb-Sutherland, B. C. (2018). Exploring infant gesture and joint attention as related constructs and as predictors of later language. Infancy, 23(3), 432-452.

Scelfo, J. (2011). Video chat reshapes domestic rituals. New York Times.

Stern, D. (1985). The interpersonal world of the infant, New York: Basic Books

Tarasuik, J. C., Galligan, R., & Kaufman, J. (2011). Almost being there: video communication with young children. PloS one, 6(2).

Taylor, M. (1988). Conceptual perspective taking: Children's ability to distinguish what they know from what they see. Child Development, 703-718.

Tomasello, M. (1988). The role of joint attentional processes in early language development. Language sciences, 10(1), 69-88.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. Behavioral and brain sciences, 16(3), 495-511.

Tomasello, M. (1995). Joint attention as social cognition. Joint attention: Its origins and role in development, 103130, 103-130.

Tomasello, M. (2003). Constructing a Language. Harvard University Press.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. Child development, 78(3), 705-722.

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(3), 786-799.

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.

Tulving, E. (1983). Elements of episodic memory.

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes. Harvard university press.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. Cognition, 30(3), 239-277.

Wellman, H. M., & Bartsch, K. (1994). Before belief: Children's early psychological theory. Children's early understanding of mind: Origins and development, 1994, 331-354.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. Child development, 72(3), 655-684.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. Child development, 75(2), 523-541.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103-128.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. Review of general psychology, 1(2), 198-226.