Do you see what I see? Children's understanding of perception and physical interaction over video chat

Elizabeth Bennette (ehbennet@ucsd.edu),
Alison Metzinger (ametzing@ucsd.edu),
Michelle Lee (m.lee@ucsd.edu),
Adena Schachner (schachner@ucsd.edu)
University of California, San Diego, Department of Psychology
9500 Gilman Drive, San Diego, CA 92093-0109 USA

Abstract

How do children reason about people presented over video chat? Video chat is a representation, like a picture; but is also a real social interaction (the partner sees and hears you). Do children understand the nuanced affordances and limitations of video chat? We tested 4-year-old children's reasoning, asking if a person over video chat (vs. a live person; photograph) could see, hear, feel, and physically interact through the screen. Children judged that a person over video chat can see, but cannot feel nor receive an object, through the screen. The person over video chat was judged to hear more often than a photograph, but less often than a live person. Preschool children are not limited to considering a stimulus fully representational, or fully present; instead, they understand video chat as a medium that blurs the boundaries of representation and reality, allowing for a mixture of life-like affordances and picture-like limitations.

Keywords: technology; video chat; cognitive development; theory of mind; representation

Introduction

It has been more than 15 years since the introduction of video chat applications like Skype (in 2003) and FaceTime (in 2010). Over the past decade, video chat has revolutionized the way families integrate technology into their daily lives, becoming the primary means of relationship maintenance for geographically separated family members (Yarosh, et al., 2009, 2011; Madianou, 2013; Forghani & Neusteadter 2014). Video chat is treated as categorically different from other 'screen time' both by caretakers of young children (McClure, et al. 2014), and by the American Academy of Pediatrics (Chassiakos et al. 2016). Overall, data show that video chat is now a common part of children's experience from early in life (McClure, et al. 2014).

How do children understand interactions over video chat? Video chat presents a complex cognitive stimulus: It blurs the boundaries of representation and reality, allowing for a mixture of life-like affordances and picture-like limitations. In some senses, video chat is a representational stimulus – similar to a picture or television image, the person is not really there, and cannot feel touch or exchange objects through the screen. However, video chat is also a real social interaction, involving contingent interaction, shared attention, and communication; and during which the social partner can see and hear you, just as in an in-person interaction. Can young children understand the nuanced affordances and limitations of a person presented over video chat, switching between treating them as life-like or as representational as appropriate to each modality?

Children's Understanding of Photographs and Television as Representational

To shed light on children's understanding of videochat, we must first characterize children's understanding of representational stimuli, such as pictures and television. Previous work shows that children gradually come to understand the representational nature of pictures and television during infancy and the preschool years. By 9 months, infants transfer information from photographs to objects, thus showing some comprehension of pictures as related to their referents (Shinskey & Jachens, 2014). Other tasks provide evidence that 1-2 year-old children understand that pictorial representations are symbols of real-life objects (e.g. Preissler & Carey, 2004; Deloache et al., 1998; DeLoache & Burns, 1994; Suddendorf, 2003). However, there are limitations to this understanding: 9-month-old infants attempt to grasp objects presented as photographs as though they were real objects, suggesting that their comprehension is incomplete (Deloache et al., 1998; Pierroutsakos and Troseth, 2003). Explicit interview tasks suggest that children's explicit understanding of photographs may not be not robust until age 4 years (Flavell, Flavell, Green & Korfmacher, 1990).

Children's understanding of the representational nature of (non-live) television also develops during late infancy and the preschool years, despite television's dynamic imagery and realism, which is greater than photographs (Strause & Troseth, 2008). Nine-month-old infants interact with moving objects presented on a screen by attempting to grasp them, just as they do with photographs; by 15-19 months, this behavior declines and is replaced with pointing, for both photographs and television (Deloache et al., 1998; Pierroutsakos and Troseth, 2003). This change in behavior suggests that by the middle of the second year of life, children distinguish television and pictures from live stimuli.

To probe children's understanding of the nature of television in more detail, Flavell and colleagues (1990) conducted a foundational study asking whether children thought of television images as representations (similar to pictures), or real objects existing inside the television set. Flavell compared 3- and 4-year-old participants' reasoning about pre-recorded video to their reasoning about live, real stimuli, and stimuli in photographs (which were used as baselines for comparison). Children were then asked questions about the physical affordances of objects and people presented in each of these ways – that is, what possibilities for action it "afforded" (Gibson & Spelke, 1983). Some measures also pertained to the mental states and perceptual abilities of people seen over television: Could a person seen in a (pre-recorded) video see, hear, and know about the actions of the experimenter? This task showed an incomplete understanding of pre-recorded video at age three, and an adult-like understanding by four years of age. That is, by four, children judged that a person in a video was not truly present, and thus could not see, hear, or physically engage with them (Flavell et al., 1990).

The Current Study

Here we aim to characterize 4-year-old children's understanding of video chat, by adapting established methods previously used to explore children's understanding of television (Flavell et al., 1990). In particular, we explore children's reasoning about the other person's ability to see what the child sees, hear what the child hears, feel a touch through the screen, and engage in a physical interaction through the screen (four modalities). For each of these modalities, we compare children's reasoning about video chat to their reasoning about a person present live, and about a person in a photograph. As in previous work (Flavell et al 1990), these conditions served as baselines for comparison, to examine whether children's judgments about video chat resemble their responses regarding reality (live interaction) or representation (photographs).

It may be that young children are limited to considering a stimulus either representational, or fully present and real; in this case, children may overestimate the capacities of people they talk to over video chat, believing them to share all affordances of live social interactions including tactile sensation and physical interaction. Children may also underestimate the affordances of video chat, construing all screen-based images as having the affordances of representations like photographs or non-live television. Alternatively, it is possible that even at preschool age, children understand the mixture of life-like affordances and picture-like limitations that video chat involves. If so, this would suggest that children are able to understand video chat as a unique medium that blurs the boundaries of representation and reality.

We thus aimed to characterize the accuracy of children's understanding of video chat, asking if children understood the particular ways video chat affords, and does not afford, perceptual and physical access. In doing so we aimed to inform the broad cognitive question of how children comprehend the semi-representational nature of video chat.

Method

Participants

N=44 4-year-old children participated (21 male, 23 female; Mean age=4 years; 5.5 months, SD=3.76 months; Range 4;0 to 5;0). Children were from the San Diego metro area, and were recruited via email and phone call invitations from a database of local interested families. An additional 4 children were tested but excluded for reasons determined a-priori: Extreme distraction and inattention (1); and technical difficulties (3).

Design

Each participant completed 12 trials for a within-subject design, with one trial for each of four modalities (whether another person could see, hear, feel or physically interact), across three presentation types (when the person was presented live, in a photo, or over video chat). Questions were blocked by presentation type, with the order of blocks counterbalanced across participants. Question orders within each block were counterbalanced across participants using a Latin square design.

Stimuli and Procedure

All testing took place in a child-friendly testing room in the lab, with the participant seated opposite the primary experimenter across a child-size table.

Warm-up Trials Two warm-up questions familiarized the participant with the situation and format of test questions (as in Flavell et al., 1990). Children were first shown a clear glass partially filled with water, and asked: "When I turn this glass upside down, what will happen: Do you think the water will spill out? Or will the water stay in the glass?". After the child's answer, the experimenter turned the glass upside down over a bowl (providing feedback). Second, children were shown a laminated photo of a glass of milk, and asked the same question regarding "turning the picture of a glass of milk upside down". After the child's answer, the experimenter again provided feedback by turning the picture upside down.

Test Trials Children then completed 3 blocks of 4 test trials, one block for each presentation type (live, photo, video chat).

At the start of each block, a new person was introduced. In the live block, a person entered the room and sat at the table. In the photo block, a laminated photo of a person was placed on a bookstand at the same location at the table. In the video chat block, a person was called by the experimenter using FaceTime on an iPad, and the iPad was placed on the same bookstand at the same location at the table. Each person was a different female individual, none of whom had interacted with the participant prior to their role in the procedure.

The introduction of people over live, photo, and video chat was closely matched, with the introduction of "my friend [Name]" (Live, Video chat), or "a picture of my friend

[Name]" (Photo). To prevent the live or video chat researchers from providing direct evidence that they could see, hear, etc. through their behavior, we implemented two procedures: (1) All researchers maintained a neutral-positive facial expression with forward gaze direction, and produced minimal body movement (with some small movements to prevent appearing unnaturally still). (2) To prevent children from over-interpreting unresponsive behavior, interactions were prefaced with the explanation that there was a rule in the game that the other person "has to sit there quietly, no matter what happens, so that she can't give away the answers."

On each test trial, the experimenter showed a stimulus, and asked a question of the same format: "Does [Name] [see, hear, etc.] [the stimulus], or does she *not* [see, hear, etc.] it?" Question wording for each modality was identical across all presentation types (live, photo, video chat).

Sight modality Sight stimuli were three different stuffed animals of the same size (monkey, alligator, rhino), placed on the center of the table, within direct line of sight of the child and the other person/photo. Children were asked: Does [Name] see the stuffed animal? Or does she *not* see it?

Hearing modality Sound stimuli were hand bells, rung briefly by the experimenter (~3 seconds); across the three sound trials the handbells varied in color and acoustic pitch, and were otherwise identical. Children were asked: Does [Name] hear the sound of the bell? Or does she *not* hear it?

Tactile modality Tactile stimuli involved the experimenter poking the other person's arm with an index finger; poking the same body location in the photo; or poking the same body location on the iPad screen. Children were asked: Does [Name] feel me poking her? Or does she *not* feel it?

Physical interaction modality Children were shown a hair tie (grey, blue, pink across trials), and asked: Could I use this hair tie to put [Name's] hair in a ponytail? Or could I *not* use this hair tie to put her hair in a ponytail? All experimenters had long hair, worn down, to make this possible for all individuals (if they had been present live).

Additional Exploratory Questions At the end of the session, several exploratory questions were asked, with the aim of further characterizing children's understanding of video chat. These questions were: "If we look behind the iPad will [Name] be there? Or will she not be there?" and "If we open up the iPad what would we find inside? Would we find [Name], wires, or something else?" If the child responded with "something else", they were asked to specify their answer, which was recorded.

Results

To ask whether participants' responses changed based on presentation type and modality, we used logistic regression to predict participants' responses (yes/no) with the predictors of presentation type (live/photo/video chat), modality (sight, hearing, touch, physical exchange), the interaction of these two factors, and subject (as a random factor). There was a significant interaction between presentation type and modality (nested model comparison, $\chi^2(6)=19.07$, $p=0.004$). There were also effects of presentation type and modality (contrasting with video chat as the reference level, live differed $p<0.0001$; with hearing as the reference level, sight differed $p<0.0001$). We explore these patterns further below.

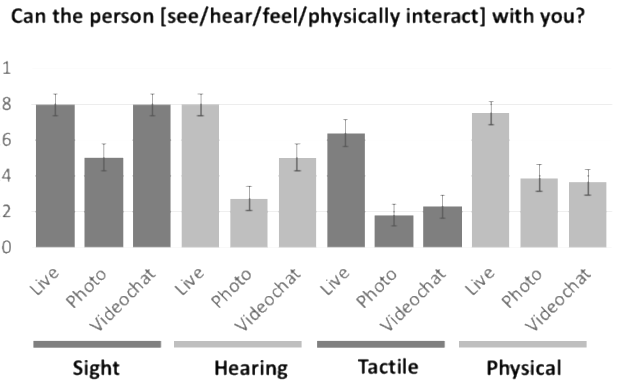

Figure 1: Results. Four-year-old children's judgements of the affordances and limitations of a person over video chat, as compared to the same judgments of a person present live, and a person in a photograph. Y-axis shows proportion of children answering "yes". X-axis shows the type of question asked, regarding both modality (e.g. sight) and presentation type (e.g. live). Error bars are standard error. Children accurately understood most affordances of video chat, in spite of the resemblance of video chat to reality in some modalities, and to a representation in others.

Do children's judgments of the abilities and limitations of a partner over video chat resemble their judgments regarding live interaction, photographs, or neither? To ask how children's judgments about video chat differed from the two baseline conditions (live, photo), we performed a logistic regression for each modality separately, and asked how responses differed by presentation type within each modality.

Sight by the Video Chat Partner

Most children judged a person capable of seeing an object over video chat (79.5% of participants), and there was no difference between children's judgments about capacity to see over video chat vs. live interaction (both were 79.5%; $p=1.0$; logistic regression, comparing live to video chat as the reference level). In contrast, children accurately judged a photo capable of sight less often (50.0% vs. 79.5%; $\beta=-2.42$, $p=0.0018$).

Hearing by the Video Chat Partner

Only 50.0% of participants judged the video chat partner capable of hearing. However, they accurately judged the photo capable of hearing less often (27.3% vs. 50.0%, $\beta=-1.53$, $p=0.013$), and the live partner capable of interaction more often (79.5%% vs. 50.0%, $\beta=2.08$, $p=0.002$), suggesting that they distinguish hearing in video chat from hearing in live interaction and photographs.

Tactile Sensation by the Video Chat Partner

Most children judged that a person over video chat could not feel touch through the screen (77.3% no; 22.7% yes), and there was no difference between children's judgments about video chat vs. a photograph (22.7% vs. 18.2%, β=-1.17, p=0.33). In contrast, children accurately judged a person capable of feeling touch more often when they were present live (63.6% vs 22.7%, β=10.27, p=0.0002).

Physical Interaction with the Video Chat Partner

Most children judged that they could not physically interact with a person over video chat through the screen (63.6% no; 36.4% yes); there was no difference between children's judgments about physical interaction capability for photo vs. video chat (38.6% vs 36.4%, β=0.2, p=0.75). Participants accurately judged physically interaction more possible for live interaction than over video chat (75.0% vs 36.4%, β=3.57, p=0.0004).

Discussion

Overall, these data show that preschool children are not limited to considering a stimulus either fully representational, or fully present. Instead, they understand video chat as a medium that blurs the boundaries of representation and reality, allowing for a mixture of life-like affordances and picture-like limitations. Thus, 4-year-old children judge that a person over video chat can see objects through the screen (as if they were present live), but cannot feel another person's touch nor physically interact through the screen (just as a representation, like a photograph, also could not). Children also more frequently judged that person over video chat was able to hear, versus a photograph (although they judged a live person able to hear more consistently).

In previous work on children's understanding of television, Flavell and colleagues posed the question of whether young children think of television images as more like pictures, or real objects (Flavell et al., 1990). Implicit in this binary framing is the idea that people categorize any given stimulus as either representational, or present and real. Technological change (and the advent of video chat) provided a natural experiment to test this framework: The current work shows that this binary categorization is not a necessary component of children's thinking.

The current dataset raises novel questions regarding unique ways that video chat technology may give children insight into others' mental states. For example, when reasoning about what a person can see over video chat, children have access to a uniquely informative source of information: The viewfinder, a small box present in the corner of the screen which directly displays the visual scene – as the other person sees it on their screen. Thus, in a real sense, the video chat interface simplifies the problem of perspective taking (e.g. Birch et al., 2017), reducing the need for mental state reasoning in the visual modality by providing a direct representation of the other person's visual access. In our dataset, we noted that children often made use of this feature: Just after the sight question was asked, many children leaned over to peer closely at the video chat screen, seemingly checking the viewfinder to see if the object in question was within view. This observation suggests another way that children have a nuanced understanding of video chat: They appear able to appropriately make use of its unique features as novel sources of information about the mental states of another person.

In addition, these data raise questions about how children reason about hearing over video chat. Children's answers for this modality were divided: Although children more frequently judged that a person over video chat could hear than a person in a photo, only approximately half of children judged that the person over video chat could hear through the screen.

We see two possible interpretations of this finding. It may be that it is more difficult for children to reason about what another person can hear, versus what another person can see or feel. Several studies have found that children have a more difficult time understanding how hearing leads to knowing, vs. how seeing leads to knowing (O'Neill & Gopnik, 1992; Pillow, 1993, Weinberger & Bushnell, 1994), although findings are mixed (e.g. Schmidt & Pyers, 2014). A general difficulty reasoning about auditory access may explain children's greater uncertainty about the auditory modality in video chat.

We find another interpretation more plausible: Children's divided answers for hearing may be rational, and reflect an accurate understanding of the technological features and challenges of video chat. There are several common reasons why a person over video chat may be unable to hear: The mute button may be activated; the volume on the other device may be turned down; or there may be a problem with the call's connection. In the visual modality, if the call's connection is severed or the video signal muted, this is immediately apparent on the screen. However, for hearing, problems with the connection are not immediately obvious, and must instead be inferred based on whether the other person responds to sounds – evidence which was intentionally not made available in our study. Thus, it is possible that children's uncertainty about hearing in video chat reflects a rational, adult-like inference – because there is, in fact, less certainty about whether the person over video chat can hear.

Overall, these data demonstrate for the first time that by preschool age, children grasp the semi-symbolic nature of video chat: Children reason about the perceptual and physical capacities of video chat partners, and recognize their unique mixture of life-like affordances and photo-like limitations. This study provides a foundation for future work exploring how understanding of video chat develops earlier in life. Anecdotal reports suggest that unlike 4-year-olds, toddlers may dramatically misunderstand the nature of video chat, overestimating the extent to which the person is really present (e.g. putting raisins behind an iPad for grandpa to eat later; LaFrance, 2015). In future work, we plan to compare and contrast younger toddlers' understanding of video chat, to characterize the origins of children's nuanced understanding of video chat, and other semi-representational technology.

Acknowledgments

We thank Jessica Ni, Shruti Nishith, Alyssa Quy, and Amanda Calcetas for their assistance with stimulus design and data collection.

References

Ames, M. G., Go, J., Kaye, J. J., & Spasojevic, M. (2010, February). Making love in the network closet: the benefits and work of family videochat. In *Proceedings of the 2010 Association for Computing Machinery Conference on Computer-supported Cooperative Work*, 145-154.

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

Birch, S. A. J., Li, V., Haddock, T., Ghrear, S. E., Brosseau-Liard, P., Baimel, A., & Whyte, M. (2017). Perspectives on perspective taking: how children think about the minds of others. In *Advances in child development and behavior* 52, 185-226.

Chassiakos, Y. L. R., Radesky, J., Christakis, D., Moreno, M. A., & Cross, C. (2016). Children and adolescents and digital media. *Pediatrics*, 138(5), e20162593.

DeLoache, J. S., & Burns, N. M. (1994). Early understanding of the representational function of pictures. *Cognition*, 52(2), 83-110.

DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengren, K. S., & Gottlieb, A. (1998). Grasping the nature of pictures. *Psychological Science*, *9*(3), 205-210.

Flavell, J. H., Flavell, E. R., Green, F. L., & Korfmacher, J. E. (1990). Do young children think of television images as pictures or real objects? *Journal of Broadcasting & Electronic Media*, 34(4), 399-419.

Forghani, A., & Neustaedter, C. (2014). The routines and needs of grandparents and parents for grandparent-grandchild conversations over distance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 4177-4186.

Gibson, E. J., & Spelke, E. S. (1983). The development of perception. In J. H. Flavell and E. M. Markman (Eds.), *Handbook of child psychology, Vol. 3: Cognitive development* (1-76). New York: Wiley.

Hill, J.A.C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, 6, 287-317.

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology*, 10(1), 41-63.

LaFrance, A. (2015, September 10). Do Babies Know the Difference Between FaceTime and TV? *The Atlantic.* Retrieved from https://www.theatlantic.com/technology/archive/2015/09/do-babies-know-when-theyre-skyping/404650/

Madianou, M., & Miller, D. (2013). *Migration and new media: Transnational families and polymedia.* Routledge.

McClure, E., Chentsova-Dutton, Y, Barr, R, Holochwost, S & Parrott, W. (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.

O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology*, 27, 390– 397.

Pierroutsakos, S. L., & Troseth, G. L. (2003). Video verite: Infants' manual investigation of objects on video. *Infant Behavior and Development*, 26(2), 183-199.

Pillow, B. H. (1993). Preschool children's understanding of the relationship between modality of perceptual access and knowledge of perceptual properties. *British Journal of Developmental Psychology*, 11, 371– 389.

Preissler, A. M., & Carey, S. (2004). Do both pictures and words function as symbols for 18-and 24-month-old children?. *Journal of Cognition and Development*, 5(2), 185-212.

Schmidt, E., & Pyers, J. (2014). First-hand sensory experience plays a limited role in children's early understanding of seeing and hearing as sources of knowledge: Evidence from typically hearing and deaf children. *British Journal of Developmental Psychology*, *3*2(4), 454-467.

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

Shinskey, J. L., & Jachens, L. J. (2014). Picturing objects in infancy. *Child Development*, 85(5), 1813-1820.

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

Suddendorf, T. (2003). Early representational insight: Twenty-four-month-olds can use a photo to find an object in the world. *Child Development*, 74(3), 896-904.

Weinberger, N., & Bushnell, E. W. (1994). Young children's knowledge about their senses: Perceptions and misconceptions. *Child Study Journal*, 24, 209– 235.

Yarosh, S., & Abowd, G. D. (2011). Mediated parent-child contact in work-separated families. *In Proceedings of the SIGCHI Conference on Human Factors in Computing System*s, 1185-1194.

Yarosh, S., Chieh, Y., & Abowd, G. D. (2009). Supporting parent–child communication in divorced families. *International Journal of Human-Computer Studies*, 67(2), 192-203.