Published in final edited form as:
Child Dev. 2014 May ; 85(3): 956–970. doi:10.1111/cdev.12166.

Skype me! Socially Contingent Interactions Help Toddlers Learn Language

Sarah Roseberry[1], **Kathy Hirsh-Pasek**[2], and **Roberta Michnick Golinkoff**[3]

[1]University of Washington

[2]Temple University

[3]University of Delaware

Abstract

Language learning takes place in the context of social interactions, yet the mechanisms that render social interactions useful for learning language remain unclear. This paper focuses on whether social contingency might support word learning. Toddlers aged 24- to 30-months (N=36) were exposed to novel verbs in one of three conditions: live interaction training, socially contingent video training over video chat, and non-contingent video training (yoked video). Results suggest that children only learned novel verbs in socially contingent interactions (live interactions and video chat). The current study highlights the importance of social contingency in interactions for language learning and informs the literature on learning through screen media as the first study to examine word learning through video chat technology.

Young children's ability to learn language from video is a hotly debated topic. Some evidence suggests that toddlers do not acquire words from screen media before age 3 (Robb, Rickert & Wartella, 2009; Zimmerman, Christakis & Meltzoff, 2007), while others find limited learning or recognition in the first three years (Barr & Wyss, 2008; Krcmar, Grela & Lin, 2007; Scofield & Williams, 2009). Yet, a common finding in the literature is that children learn language better from a live person than from an equivalent video source (Krcmar, et al., 2007; Kuhl, Tsao & Liu, 2003; Reiser, Tessmer & Phelps, 1984; Roseberry, Hirsh-Pasek, Parish-Morris & Golinkoff, 2009). What makes social interactions superior to video presentations for children's language learning? We hypothesize that a key difference between the contexts of screen media and live interaction is social contingency between the speaker and the learner.

The “video deficit” (Anderson & Pempek, 2005), or the discrepancy between learning from a live person and learning from an equivalent media source, is a widely known phenomenon. Kuhl and colleagues (2003), for example, exposed 9-month-olds infants from English-speaking households to Mandarin Chinese through speakers on video or by live speakers. The researchers asked whether children would experience the same benefits in discriminating between foreign phonemes if their foreign language exposure came through

Correspondence concerning this article should be addressed to Sarah Roseberry, Institute for Learning and Brain Sciences, University of Washington, Mail Stop 357988, Seattle, WA 98195. Electronic mail may be sent to sarahr28@uw.edu.

the video or the live speakers. Results suggested that children who heard the speakers in a live demonstration learned to discriminate between the foreign language sounds whereas the video display failed to confer this advantage. Another example leads to the same conclusion, here with word learning. These researchers investigated children's ability to learn verbs, which some researchers have suggested are more difficult to master than nouns (Gentner, 1982; Gleitman, Cassidy, Nappa, Papafragou & Trueswell, 2005; but see Choi & Gopnik, 1995; Tardif, 1996). Could children learn these verbs from mere exposure to televised displays? In a controlled experiment, 30-month-olds learned better when an experimenter was live than when she appeared in the screen condition (Roseberry et al., 2009). Even though children older than three years gained some information from video alone, this learning was still not as robust as learning from live social interactions.

Given the overwhelming evidence that young children do not learn as much from video as they do from live interactions, what accounts for this discrepancy? One line of research, outside of the language literature, suggests that children do learn from video if the video format also allows them to engage in a contingent interaction (Lauricella, Pempek, Barr & Calvert, 2010; Troseth, Saylor & Archer, 2006). Troseth and colleagues (2006), for example, used an object retrieval task, in which an experimenter hid a toy, told the 24-month-olds where it was located, and then asked the toddlers to find the toy. Before revealing the location of the hidden toy, all children viewed a 5-minute warm-up of the experimenter on video. Half of the toddlers participated in a two-way interaction with the adult via closed-circuit video for the warm-up, whereas the other children viewed a pre-recorded video of the adult as they had interacted with another child. During the interaction via closed circuit video, the adult on video called children by name and engaged them in conversation about their pets and siblings. The pre-recorded, or *yoked*, video was not dependent on the child's responses and showed the experimenter asking about pets and siblings that were not relevant to the child for whom the video was played. When children searched for the hidden toy, only the children who experienced a social interaction with the adult via video found the toy at rates greater than chance. The researchers argue that socially contingent video training allowed toddlers to overcome the video deficit. These findings have recently been extended to show increased learning from interactive computer games relative to watching video (Lauricella et al., 2010).

Troseth and colleagues (2006) defined contingent interactions as a two-way exchange in which the adult on video established herself as relevant and interactive by referring to the child by name and by asking children specific questions about their siblings and pets. This view of social contingency posits that socially contingent interactions should be appropriate in content (Bornstein, Tamis-LeMonda, Hahn & Haynes, 2008) and intensity (Gergely & Watson, 1996). It is a departure from a narrower definition of contingency, which focuses solely on timing and reliability (Beebe et al., 2011; Catmur, 2011).

In the few studies that have investigated the role of contingency in language learning, timing and synchrony of interactions have been the focus. Bloom, Russell and Wassenberg (1987), for example, manipulated whether adults responded to 3-month-olds randomly or in a conversational, turn-taking manner. Here, the contingent interaction appeared as the parent listening while the infant vocalized and then immediately vocalizing in return. Results suggested that infants who experienced turn-taking interactions with an adult produced more syllabic, or speech-like, vocalizations. These findings have been extended with 5- and 8-month-olds who engaged in a contingent or non-contingent interaction with their mothers (Goldstein, King & West, 2003; Goldstein, Schwade & Bornstein, 2009). Infants learn quickly that their vocalizations affect their caregiver's response (Goldstein et al., 2009), and infants whose mothers were told to respond immediately to infant vocalizations, as opposed to responding randomly, produced more mature vocalizations (Goldstein et al., 2003).

Taken together, contingency has been implicated as an important catalyst for early language development, and its absence may be responsible for children's inability to use information presented on video. Yet, the role of social contingency in children's ability to learn words has not been explored. The current study examines social contingency as a cue for language learning. We define a socially contingent partner as one whose responses are not only immediate and reliable, but are also accurate in content (Csibra, 2010; Tamis-LeMonda et al., 2006; Troseth et al., 2006).

One method of investigating social contingency in children's language learning is through video chats. Video chatting is a new technology that provides a middle ground between live social interactions and screen media. This communication tool has some features of video and some features of live interactions. As a video, it provides a two-dimensional screen. As an interaction, it is a platform for socially contingent exchanges. To a slightly lesser degree, video chat offers the possibility of noting where the speaker is looking, although the speaker's eye gaze is somewhat distorted from the child's perspective.

Children use a speaker's eye gaze as an important communicative signal from early in life (Csibra, 2010). Infants prefer to look at eyes from birth (Batki, Baron-Cohen, Wheelwright, Connelan, & Ahluwalia, 2000), and even 3-month-olds prefer to look at photographs of faces with eyes that appear to make eye contact with them (Farroni, Csibra, Simion & Johnson, 2002). By 19- to 20-months, toddlers understand that eye gaze can be referential and can help them uncover the meanings of novel words (Baldwin, 1993). Novel labels typically refer to the referent in the speaker's purview (Baldwin, 1993; Bloom, 2002; Tomasello, 1995) and in fact, when the referent of a novel word is ambiguous, children are more likely to check speaker gaze to determine the correct referent (Baldwin, Bill & Ontai, 1996). One recent study suggests that older infants use eye gaze to learn labels for boring objects even when they would prefer to look at other interesting objects (Pruden, Hirsh-Pasek, Golinkoff & Hennon, 2006).

This study is the first to use video chats to test the role of social contingency in word learning, as well as additionally investigate whether children attend to the speaker's eyes, perhaps in an attempt to recruit information about the referent of the novel verb. Building on previous research that compares learning from video to learning from live interaction (Roseberry et al., 2009), we tested the efficacy of social contingency on language learning by asking whether language learning via video chats is similar to learning in live interactions or to learning from video. In this way, the current study seeks to inform both the literature on children's ability to learn from screen media as well as the literature on the social factors of children's language learning.

We investigated one particular case of language acquisition – verb learning. Verbs are the building blocks of grammar and the fulcrum around which a sentence is constructed. Nearly thirty years of research demonstrates that verbs can be significantly more difficult to acquire than nouns for children learning English (Gentner, 1982; Gleitman et al., 2005; Golinkoff & Hirsh-Pasek, 2008; but see Choi & Gopnik, 1995; Tardif, 1996). Research is only beginning to uncover how children learn action words, so testing social cues with verb learning provides an especially strong test of the role of social contingency in language acquisition.

We hypothesize that if word learning relies on social contingency, children's learning from video chats will be more similar to learning from live interactions than to learning from video. In contrast, if the two-dimensional aspect of video chats prevents children from learning verbs, we suggest that no learning will take place from video chats, revealing once again, the “video deficit” (Anderson & Pempek, 2005). In this case, learning from video chats will resemble learning from video. Furthermore, the role of eye gaze in language learning is well established (Baldwin et al., 1996; Bloom, 2002; Tomasello, 1995), yet video chatting currently affords only contingent yet somewhat misaligned eye gaze. We hypothesize that if children attempt to recruit information from the speaker's eye gaze, they will look longer to the experimenter's eyes.

Method

Participants

Thirty-six children between the ages of 24- and 30-months (19 male, m = 26.52, SD = 1.74, range = 24.09 to 29.80) participated in the study. This age was chosen because 24-month-olds show robust verb learning from social interactions (Childers & Tomasello, 2002; Naigles, Bavin & Smith, 2005) but do not yet show evidence of verb learning from video displays (Krcmar et al., 2007; Roseberry et al., 2009). Children were randomly assigned to one of three training conditions: Twelve children participated in the video chat condition (m = 26.35, SD = 1.90, range = 24.09 to 29.80), 12 in the live interaction condition (m = 26.78, SD = 1.79, range = 24.09 to 28.90), and 12 in the yoked video condition (m = 26.42, SD = 1.64, range = 24.36 to 29.18). The yoked video condition showed participants pre-recorded video of the experimenter communicating in a video chat with another child (see Murray & Trevarthen, 1986; Troseth, et al., 2006). An additional 8 participants were excluded from the current data set for fussiness (2), bilingualism (1), experimenter error (2), prematurity (2) and technical difficulties (1). Of the excluded participants, 3 were from the video chat condition (1 for fussiness, 1 for technical difficulties, 1 for prematurity), 3 were from the live condition (1 for fussiness, 1 for bilingualism, and 1 for experimenter error) and 2 were from the yoked video condition (1 for experimenter error and 1 for prematurity). All children were full-term and were from monolingual English-speaking households.

Design and Variables

To determine whether language learning in video chats is similar to learning from live interactions or from yoked video, we used a modified version of the Intermodal Preferential Looking Paradigm (IPLP; Hirsh-Pasek & Golinkoff, 1996). The IPLP is a dynamic, visual multiple-choice task for children. Here, the dependent variable is comprehension, as measured by the percentage of gaze duration to the action that matches the novel verb during the test trials.

Additionally, we collected eye-tracking data to determine whether children looked at the experimenter's eyes during screen-based training (i.e., video chat and yoked video training). The dependent variable here is percentage of looking time towards the experimenter's eyes.

Apparatus

Video-based portions of the current study (i.e., Introduction Phases, Salience Phases, Video Chat Training Phases, Yoked Video Training Phases, Test Phases; everything except the Live Interaction *Training Phases*) used a Tobii X60 eye tracker to collect eye gaze data during video exposure. Children's eye gaze was recorded through a sensor box positioned in front of a 32.5-inch computer monitor. This captured eye gaze within a virtual box of space (20 cm × 15 cm × 40 cm) as determined by the machine. Children sat on their parent's lap in a chair approximately 80 cm from the edge of the computer table. The height of the chair was adjusted for each participant dyad so children's eyes were located 90 cm to 115 cm above the ground on the vertical dimension and in the middle of the sensor bar's 40 cm horizontal dimension. Before the study began, a gauge appeared on the screen to confirm that the child's eyes were properly centered in the virtual space for detection. Each child's fixations were calibrated through a short child-friendly video of an animated cat accompanied by a ringing noise in each of five standardized locations on the screen.

The video chat conditions of the study used the same apparatus, but the computer monitor was also equipped with a high-quality web cam (Logitech Quickcam Vision Pro) and an external microphone (Logitech USB Desktop Microphone). The web cam was attached to the computer monitor at the top center of screen and was angled slightly downward to capture video of the child's head and torso. The stand-alone microphone was placed on the desk to the right of the computer monitor and was calibrated to capture the child's speech. During the video chat interactions, an experimenter appeared on video chat from another room, in which they were seated in front of a 21.5-inch computer with a built-in camera and microphone. Both computers were equipped with Skype, a video chat software that connects users via the Internet. Using Skype software, the experimenter and participant communicated with each other in real time as the computers transmitted audio and video back and forth. This provided the basis for a contingent interaction.

All conditions used overhead fluorescent lighting to facilitate gaze capture and parents wore opaque glasses to ensure that they did not see the video and could not influence their child's looking patterns.

Stimuli

Each child was trained and tested on four novel verbs (Table 1). All four referent actions were transitive, meaning the actions required an object or character to be acted upon. Each action was labeled with a nonsense word.

To train children on each of the novel verbs, an experimenter performed the referent action with the designated prop while labeling the action with the novel verb. For example, *meeping* refers to turning a dial on an object. When the experimenter performed *meeping* for the child, she said,

> "Look at what I can do with this toy! I'm *meeping* it! Do you see me *meeping* the toy? Wow, I'm *meeping* it! Watch me *meeping* the toy! I'm *meeping* it. I am *meeping* the toy! Would you like to see that again? Let me show you one more time. Cool! I'm *meeping* it! Do you see me *meeping* the toy? Wow, I'm *meeping* it! Watch me *meeping* the toy! I'm *meeping* it. I am *meeping* the toy!"

Each novel verb was uttered 12 times in full sentences. The script was identical for each action, except for the particular novel verb (i.e., *blicking*, *twilling*, *frepping*, *meeping*) and the specific object used to demonstrate the action.

All verbs were tested using video clips edited from *Sesame Beginnings* (Hassenfeld, Nosek & Clash, 2006a; 2006b). Each test clip required children to extend their knowledge of the novel verb to a new actor and to a new object performing the action. For example, an experimenter demonstrated *meeping* during training by turning the rotors on a toy helicopter, but the test clip of *meeping* showed Elmo's dad turning the dial on an old-fashioned radio. Although the rotors and radio dial were perceptually similar, and the experimenter and Elmo's dad made similar turning motions, children nevertheless had to extend their knowledge of *meeping* to a new actor performing the action (i.e., from a human actor to a puppet actor) and to a new object (i.e., from a helicopter to a radio).

Procedure

The experiment was divided into five phases that always occurred in the same order and consisted of the Introduction/Salience Phase, Training Phase 1, Testing Phase 1, Training Phase 2, Testing Phase 2, with training condition as the only between-subjects variable (Figure 1). Training and testing of novel verbs was divided into two segments based on pilot data suggesting that children lost focus on the task when four verbs were presented serially. Importantly, each child only experienced one mode of training during the experiment. That is, a child in the video chat condition participated in a video chat interaction during both Training Phase 1 and Training Phase 2. Each child was exposed to four verbs throughout the study.

Introduction phase—A character (e.g., Cookie Monster) appeared first on one side of the screen and then on the other side for 6 seconds each. The left/right presentation order was counterbalanced. This introduction ensured that toddlers expected to find information on both sides of the screen.

Salience phase—Children saw previews of the exact test clips to be shown in the *Test Phase*. Measuring looking time to this split-screen presentation before training allowed detection of *a priori* preferences for either member of a pair of clips. Lack of an *a priori* preference in the *Salience Phase* indicates that children had no natural preference to look at one video clip or the other. This assures that differences in looking time during the *Test Phase* are due to the effects of the *Training Phase*. Because each child was trained on four novel verbs, the *Salience Phase* presented four 6-second test clips; one for each of the novel verbs in which the novel action for the novel verb was paired with a second novel action. Each of the four trials in the *Salience Phase* was separated by a 3-second “centering trial” that showed a video of a laughing baby accompanied by child-friendly music. This was designed to recall the child's attention to the center of the screen. The order of presentation of the salience videos for each verb was counterbalanced.

Training phases—Children participated in a series of two *Training* and *Testing Phases*. To most effectively train and test four novel verbs, each *Training Phase* presented two novel verbs and the child was immediately tested on those verbs in the subsequent *Testing Phase*. The remaining two verbs were paired and presented in the second *Training* and *Test Phases*. Presentation and testing order were counterbalanced across participants such that each verb appeared equally often in both the first and second training and test positions.

Video chat training—Children assigned to the video chat *Training Phases* sat on their parent's lap in front of the computer monitor, webcam, and microphone. An experimenter (Experimenter 1) was seated in front of a computer in another room ready to interact with the child via Skype. Since the video chat format necessitated that Experimenter 1 and the participant be in different rooms, the video chat training sessions required the assistance of a second experimenter (Experimenter 2) to operate the participant's computer. Immediately following the *Introduction* and *Salience Phases*, Experimenter 2 initiated a Skype call with Experimenter 1. Once the call was answered, Experimenter 2 started recording the events with the Tobii eye tracker, maximized the Skype screen and hid all of the Skype toolbars. When this process was complete, the child's computer monitor contained a full-screen video of Experimenter 1 and nothing else. Experimenter 2 signaled to Experimenter 1 that the *Training Phase* could begin and moved behind a partition for the duration of the Training Phase.

The Video Chat *Training Phases* used a fixed script to ensure that all participants received the same verbal information. Experimenter 1 began the video chat training with a warm-up period; she greeted the child by name, asked a question related to the child's initial time in the laboratory playroom (e.g., “Did you like playing with the blocks?”), and invited the child to play a short game. The game consisted of the experimenter posing questions to the child, such as, “Can you point to your eyes? Where are your eyes?” Experimenter 1 responded contingently to the child after asking the questions, such that if the child successfully pointed to his or her eyes, Experimenter 1 would clap her hands and say “Great job! You pointed to your eyes!” If the child did not immediately respond, Experimenter 1 would prompt the child, saying, “I can see your eyes, can you point to them for me?” If the child remained unresponsive after 3 prompts, Experimenter 1 acknowledged the lack of response moved to another question (e.g., “That was a tricky question. Let's try another one.”). This warm-up procedure was meant to establish the experimenter as a socially contingent partner and to demonstrate the interactive nature of video chat (Troseth et al., 2006). The experimenter continued to engage the child with questions for 60 seconds, as measured by a timer on the experimenter's computer.

To begin training the novel verbs, the experimenter said, “I have some fun toys to show you. Let me show you some cool toys!” For each of the novel verbs presented during a training session, the experimenter produced the prescribed toy or doll and performed the action for 60 sec. while labeling it 12 times in full sentences. Although the experimenter focused her attention on the child on the screen, she occasionally looked toward the action she was producing. At the conclusion of the demonstration, the experimenter repeated the process for the second novel verb in the *Training Phase*. The order of verb presentation was counterbalanced.

The *Training Phase* lasted approximately 3 minutes: One minute for the initial warm-up period and one minute to present each of two novel verbs. In the Video Chat training phase, as in all training phases, parents wore opaque glasses and were asked to refrain from interacting with their child; all parents complied.

Live interaction training—Children assigned to live interaction training moved from their parent's lap to a chair opposite Experimenter 1 at a child-sized table in the same room to begin the *Training Phase*. The parent remained in the chair in front of the computer during the live interaction, still wearing opaque glasses and oriented away from the experimenter and child.

Live interaction *Training Phases* used the common script. As in the other conditions, the training phase began with a warm-up period. The experimenter began by greeting the child by name, asking questions relevant to the child's playroom experience, and inviting the child to play a short game demonstrating the contingent nature of the live interaction. After this initial sequence, the experimenter presented the two novel verbs. The props needed to perform the actions were hidden in an opaque container next to the table. Experimenter 1 retrieved the toys as necessary and then placed them back into the container. Thus, children's exposure to the toys was limited to the one-minute demonstration of the novel verb, as in the other training conditions. Again, each *Training Phase* was approximately three minutes in length: one minute for the warm-up game and two minutes total for the presentation of both novel verbs.

Yoked video training—Children assigned to yoked video training were seated on their parent's lap in front of the computer monitor during the training session, just as in the video chat condition. These children viewed a video extracted from eye tracker recordings of previous video chat trainings. Yoked videos were a subset of all video chat trainings selected to be representative of counterbalanced verb presentation order as well as total length of training. Thus, four sets of videos (each set included both Training Phase 1 and Training Phase 2) were taken from the video chat condition to be used as training in the yoked video condition. Yoked videos were used to give children the exact experience – minus social contingency – of video chats. The critical aspect of yoked video training is that children did not actually experience a socially contingent interaction with the experimenter in the video. Rather, the experimenter's responses were recorded and did not change, regardless of how the child tried to interact with the experimenter.

Within the video, the training session was identical to the other conditions: a one-minute warm-up period followed by two minutes of novel verb training that contained 12 presentations of each novel verb.

Test Phases—Following each *Training Phase*, all children participated in the same *Test Phase* that examined their knowledge of the verbs. The video-based *Test Phases* consisted of four trials per verb and were identical for all children, regardless of the type of training they received. The four test trials for a particular verb were presented sequentially, followed by four test trials for the second trained novel verb. Thus, each *Test Phase* contained eight total test trials.

In each of the trials, a split-screen simultaneously presented two novel clips. One of the clips displayed the same action that children saw during the *Training Phase* (e.g., *blicking*). Importantly, *all* conditions required children to extend their knowledge of the novel verb to a new actor and a new instantiation of the event. As extension is more demanding than fast mapping (e.g., Seston, Golinkoff, Ma & Hirsh-Pasek, 2009), the inclusion of extension trials provides a strong test of verb learning.

Children viewed the same video clips for all test trials for each verb; the trials differed in the audio that children heard, which asked children to look at different events. Of the four *Test Trials* for a given verb, trials 1 and 2 were designed to examine children's ability to generalize the trained verb to an action performed by a novel actor (e.g., from experimenter to puppet). In this *Extension Test*, pre-recorded audio asked children to find the action labeled by the novel verb. For example, if children were trained that the novel verb *blicking* referred to bouncing up and down on the knee, one test clip showed Elmo's dad bouncing Elmo up and down on his knee (the matching action) and the other clip showed Cookie Monster's mom rocking Cookie Monster (the non-matching action). The audio asked children, “Where is *blicking*? Can you find *blicking*? Look at *blicking*!” If children learned the target verb, they should look at the matching action, or Elmo's dad bouncing Elmo, during each of the first two test trials.

Test Trials 3 and 4 together constituted a *Stringent Test of verb learning* by asking whether children had truly mapped the novel verb to the particular novel action (Hollich, Hirsh-Pasek & Golinkoff, 2000). Here, we examine whether children will accept any new verb for the action presented during training or whether children know that only the trained verb should label the trained action. Based on the principle of mutual exclusivity (Markman, 1989), children should prefer attaching only *one* verb to any given action. Thus, Test Trial 3, the *New Verb trial*, asked children to find a novel action that was not labeled during training (“Where is *glorping*? Can you find *glorping*? Look at *glorping*!”; *glorping*, *spulking*, *hirshing*, and *wezzling* were terms used for non-trained verbs). If children learned the target verb (e.g., *blicking*), they should not look toward the action previously labeled *blicking* during the *new verb trial*. They may look instead toward the non-matching action (e.g., Cookie Monster's mom rocking Cookie Monster), inferring that if the *blicking* action had already been named, then *glorping* must refer to the other, non-matching action. Or children may show no preference to either action, indirectly indicating their unwillingness to attach a new label to the previously labeled action. Test Trial 4, the *Recovery trial*, asked children to renew their attention to the trained action (e.g., Elmo's dad bouncing Elmo) by asking for it again by name (“Where is *blicking*? Can you find *blicking*? Look at *blicking*!”).

In sum, children who learned a novel verb should look more toward the matching action during the *Extension Test* (trials 1 and 2), look away from the matching action in Test Trial 3 (the *new verb trial*), and resume looking to the matching screen during Test Trial 4 (the *recovery trial*). This v-shape in visual fixation time during the *Extension Test*, Test Trial 3 and Test Trial 4 forms a quadratic pattern of looking, in which we can examine the relative differences between these three data points to determine whether looking to the trained action in Test Trial 3 is relatively less than in the *Extension Test*; and whether looking to the trained action is relatively more in Test Trial 4 than in Test Trial 3.

Counterbalancing determined which verb was tested first and second in each *Test Phase*. After both novel verbs were tested, children participated in the second series of *Training* and *Test Phases* (Figure 1). In sum, the entire protocol lasted 10 minutes (54 s for the Introduction and Salience Phases, 3 min for each of two Training Phases, and 93 s for each of two Test Phases), plus minimal time for transitions between phases.

Coding

We first coded children's overall attention to the *Training Phases*. For children in the video chat and yoked video conditions, data was recorded by the Tobii eye tracker. For children in the live interaction training condition, we used video recordings of these interactions. Video data was coded off-line by a trained coder. A second coder re-coded 25% of the attention measures for the live interaction training to establish reliability. Attention to the *Training Phase* was simply defined by the amount of time the child looked toward the experimenter or the action (as opposed to the wall or door of the room, for example).

Next, children's gaze direction and duration to the left and right sides of the split screen was coded for the *Salience* and *Test Phases*. Gaze direction and duration during the *Salience* and *Test Phases* was used to calculate the percentage of time children looked to either side of the screen.

Finally, we explored children's eye gaze during the *Training Phases*. Although video chats approximate live social interactions in many ways, we were interested in children's patterns of attention to the experimenter's eyes in the video chat and yoked video conditions. To examine patterns of looking, we defined the experimenter's eyes as an Area of Interest (AOI). Because the video feed for each child's training session was unique (with the exception of the yoked videos, which included some repetitions), this AOI was individually defined for each verb presentation to each child. To capture movement of the experimenter's head over time with a static AOI, a trained coder watched the videos, marked the extremities of eye movement, and then used these measurements to define the AOI. The specified eye gaze AOI was used to determine total fixation to the experimenter's eyes during training. This was converted to a percentage of total looking time.

Results

Preliminary analyses were conducted to establish null effects of several variables for which we hypothesized no difference due to training condition: salience preferences, gender, total time of training phases and child attention during training. To determine potential effects of gender and whether children had an *a priori* preference for either test clip during the *Salience Phase*, a preliminary 2 (gender) × 4 (percentage looking to the matching action during salience for each novel verb: *blicking*, *frepping*, *twilling*, *meeping*) multivariate analysis of variance (MANOVA) was conducted. No gender effects emerged, F(1, 36) = .77, p = .55, η_p^2 = .09 and results indicated no bias toward either test clip for any of the novel verbs (*blicking*, F(1, 36) = .05, p = .82, $\eta_p^2 < .01$; *frepping*, F(1, 36) = .07, p = .80, $\eta_p^2 <.01$; *twilling*, F(1, 36) = .37, p = .55, η_p^2 = .01; and, *meeping*, F(1, 36) = 1.82, p = .19, η_p^2 = .05).

A final preliminary analysis asked whether attention to training and the duration of training phases differed across conditions. A one-way ANOVA compared children's attention to training and length of training by condition (i.e., video chat, live interaction, yoked video). No group differences were found for either children's attention to training, F(2, 35) = .02, p = .99, $\eta_p^2 < .01$, or in length of training F(2, 35) = .25, p = .78, η_p^2 = .02. Together, these analyses indicate that group-level differences at test cannot be attributed to time spent in training.

To determine whether learning from video chat more closely resembled learning from live interactions or learning from videos, we examined novel word comprehension during the Test Phases and children's references to the experimenter's eyes during the Training Phases.

Did toddlers show evidence of novel verb comprehension?

Extension Test of Verb Learning—In Test Trials 1 and 2, children were asked to find the matching action when they heard the novel verb. Since these trials were identical, data from Test Trials 1 and 2 were averaged for each of the verbs. Using the mean of two test trials increases the reliability of children's responses. Each child was taught four novel verbs and because previous studies have found an order effect such that children learn verbs better later in the experiment (Roseberry et al., 2009), a 3 (training type: video chat, live interaction, yoked video) × 4 (verb order) analysis of variance (ANOVA) was used to determine the effect of training style and verb order on learning. Results indicate a main effect of training condition, F(2, 36) = 12.04, $p < .001$, η_p^2 = .50, but no main effect of verb order, F(3, 36) = .45, p = .72, η_p^2 = .05, and no interaction effect, F(6, 36) = .91, p = .50, η_p^2 = .19. This suggests that training condition impacted children's ability to learn novel verbs and that learning was not affected by verb order. Further analyses considered all verbs regardless of order.

To decipher the effect of different training styles, we conducted planned paired-sample *t*-tests comparing children's looking times to the matching action versus the non-matching action for each type of training. After Bonferroni corrections for multiple comparisons, results indicated that children who were trained through video chats and live interactions looked significantly longer toward the matching action, t(11) = 7.06, $p < .001$ (m = 66.90, SD = 8.29 to the matching action; m = 33.1, SD = 8.29 to the non-matching action), and t(11) = 5.87, $p < .001$ (m = 64.19, SD = 8.37 to the matching action; m = 35.81, SD = 8.37 to the non-matching action), respectively. These mean looking times did not differ from each other, t(22) = .80, p = .43 (m = 66.90, SD = 8.29 to the matching action in the video chat condition; m = 64.19, SD = 8.37 to the matching action in the live interaction condition). In contrast, children trained through yoked video did not look longer toward either side of the screen, $t(11) = .05$, $p = .96$ (m = 50.12, SD = 8.57 to the matching action, m = 49.88, SD = 8.57 to the non-matching action; Figure 2a). Thus, children only learned the novel verbs after video chat or live interaction training, but not when they were trained with yoked video.

Stringent Test of Verb Learning—The *Stringent Test* of verb learning is composed of three data points: the average score of Test Trials 1 and 2 (the *Extension Test*), Test Trial 3, the *New Verb trial*, and Test Trial 4, the *Recovery trial*. Notably, this test relies on the pattern of results between the three data points, as opposed to the specific percentages of looking times. Because we hypothesized that the *Stringent Test* would detect robust learning when children succeeded in the *Extension Test*, we further analyzed looking patterns only for conditions in which children successfully extended the novel verb (Roseberry et al., 2009); video chat and live interaction training conditions.

Data were analyzed using a repeated measures ANOVA with training type (video chat or live interaction) as the between-subjects factor and a series of three data points (extension test, new verb trial, recovery trial) as the within-subjects factor. Results suggest no main effect of training type, $F(1,22) = 1.86$, $p = .19$, $\eta_p^2 = .08$, but a significant within-subjects quadratic contrast emerged, $F(2,22) = 30.08$, $p < .001$, $\eta_p^2 = .58$, indicating that children with contingent training (either via video chat or live interaction) learned the novel verbs, even by the standards of the strong test of verb learning (Figure 2b). Furthermore, paired samples *t*-tests with Bonferroni corrections revealed that children in the video chat condition had no preference for either side of the screen in Test Trial 3, $t(11) = 1.79$, $p = .10$ (m = 42.12, SD = 15.26 to the matching action; m = 57.88, SD = 15.26 to the non-matching action), but looked significantly longer toward the matching screen in Test Trial 4, $t(11) = 6.57$, $p < .001$ (m = 74.60, SD = 13.00 to the matching action; m = 25.40, SD = 13.00 to the non-matching action). The same pattern emerged for children in the live interaction condition, $t(11) = 1.76$, $p = .11$ (m = 41.45, SD = 16.83 to the matching action; m = 58.55, SD = 16.83 to the non-matching action), and $t(11) = 3.50$, $p = .005$ (m = 65.68, SD = 15.51 to the matching action; m = 34.32, SD = 15.51 to the non-matching action), respectively. Finally, a comparison of percentage looking time between children in the video chat and live interaction conditions during the *Stringent Test* revealed no differences in Test Trial 3, $t(22) = .10$, $p = .92$ (m = 42.12, SD = 15.26 to the matching action in the video chat condition; m = 41.45, SD = 16.83 to the matching action in the live interaction condition), or in Test Trial 4, t(22) = 1.53, $p = .14$ (m = 74.60, SD = 13.00 to the matching action in the video chat condition; m = 65.68, SD = 15.51 to the matching action in the live interaction condition). Thus, after contingent training (either live or via video chat), children succeeded in the *Extension Test*, showed no preference in the *new verb trial*, and again preferred the matching screen in the *recovery trial*.

Do children look at the experimenter's eyes during training?

To investigate whether children referenced the experimenter's eyes during training, we conducted an Independent Samples *t*-test to detect differences in percentage of looking time between children in the video chat and yoked video training conditions, or the training groups for which we had eye gaze data. Results did not reveal a significant effect of training group, t(22) = 1.52, *p* = .14 (m = 22.42, SD = 12.86 in the video chat condition; m = 14.08, SD = 14.00 in the yoked video condition), suggesting that children in video chat training and yoked video training did not differ in the amount of time they looked at the experimenter's eyes. Even though allocation of looking did not differ on the group level, individual differences in children's attention to eyes during training may be related to their performance at test. Pearson's correlations revealed that the percentage of children's looking time directed toward the experimenter's eyes was significantly correlated with looking time during the *Extension Test*, *r*(24) = .66, *p* < .001. This suggests that the more a child focused on the experimenter's eyes during the *Training Phase*, the longer they looked toward the matching action during the *Extension Test* of verb learning.

Discussion

Discussions of young children's ability to learn language from live interactions but not from screen media appear repeatedly in the literature (e.g., Childers & Tomasello, 2002; Krcmar et al., 2007; Kuhl et al., 2003; Naigles et al., 2005; Roseberry et al., 2009; Scofield & Williams, 2009; Zimmerman et al., 2007). The current study used a new technology, video chats, to begin to understand the mechanism behind this dichotomy. We asked whether language learning via video chats is similar to learning in live interactions or to learning from video.

We found that toddlers learned novel words both from video chats and from live interactions, suggesting that socially contingent interactions are a powerful catalyst for word learning. Impressively, children who learned in these contingent environments extended the novel verbs to new instances of the action, a task that is more demanding than simply fast mapping verb meanings to actions, and children resisted applying a second novel label to the same action. Additionally, we found some evidence that children who attended to the experimenter's eyes during training learned the novel words better. This research has important implications for the role of social cues in language learning as well as for how children process screen media.

Social Mechanisms of Language Learning

Of the possible mechanisms that facilitate young children's language acquisition, the current study highlights the role of social contingency. Video chat technology allowed us to compare learning from socially contingent screen media to learning from socially contingent live interactions and non-contingent video. The results unequivocally suggest that language learning is improved by social contingency.

Based on work by Troseth and colleagues (2006), we defined socially contingent adults as social partners whose responses were immediate, reliable, and accurate in content. Although this differs from the traditional definition of contingency, in which synchrony of timing is the only requirement, the characteristics of social interactions -- such as turn-taking -- support a broader definition of social contingency. According to Csibra (2010), “turn-taking is a kind of interactive contingency that is qualitatively different from temporal synchrony or simultaneous mirroring reactions” (p. 150). This type of interactive, or social, contingency requires an iterative pattern of back-and-forth responses that contain complementary, not identical, content (Csibra, 2010). Critically, socially contingent interactions must maintain these elements over time, as research demonstrates that children are sensitive to temporal delays (Henning & Striano, 2011), alterations in the reliability of response (Goldstein et al., 2009), and changes in the accuracy of content (Scofield & Behrend, 2008).

Why was social contingency so useful for young children? One benefit of social contingency may be the trust it establishes between the speaker and the child (Koenig & Harris, 2005; Sabbagh & Baldwin, 2001; Scofield & Behrend, 2008). A warm-up period began all training phases and allowed children to determine whether the speaker was reliable or unreliable. The experimenter asked the child questions (e.g., “Can you point to your shirt?”), praised the child for a correct answer (e.g., “That's right!”), corrected the child for incorrect answers (e.g., “Your shirt is not red. It's yellow!”), and offered child-specific hints as prompts (i.e., “I can see the animal on your shirt. What kind of animal is it?”). Although this interaction was likely beneficial for children in contingent training conditions, children in non-contingent yoked video training may have experienced the opposite effect; these children experienced another child's warm-up interaction. When children understood that the experimenter's questions and responses did not depend on the addressee, the experimenter proved herself unreliable. This may have been particularly important once the experimenter began the novel verb training, which used a fixed script. When the experimenter produced the novel verbs while performing the actions, the novel verb label always occurred in conjunction with the matching action, which is a type of temporal synchrony (Brand & Tapscott, 2007; Tomasello & Barton, 1994). Termed “acoustic packaging” (Hirsh-Pasek & Golinkoff, 1996), this technique has been shown to facilitate children's language learning (Tomasello & Barton, 1994). Although temporal synchrony between the label and the action existed in all three training conditions, only social contingency seemed to facilitate word learning in the present study.

A second critical question that arises from the current findings is how we know that the results can be attributed to contingency and not another social cue, like eye gaze or joint attention, particularly in light of the fact that our results do indicate that children who looked longer at the experimenter's eyes also learned the novel words better.

Previous research suggests that attention to eye gaze is beneficial for language learning (e.g., Baldwin et al., 1996; Brooks & Meltzoff, 2005), and that eye gaze is a critical component of joint attention, or a social interaction in which both partners focus their attention on a common object or event (Adamson et al., 2004; Baldwin, 1991; Bruner, 1981; Moll & Tomasello, 2007). Although the current study measured children's attention to the experimenter's eyes and not eye gaze following, children are known to establish eye contact immediately before gaze following. In fact, children may not follow a speaker's eye gaze unless they have experienced mutual eye gaze first (Senju & Csibra, 2008; Senju, Csibra & Johnson, 2008). It may be that our measurement of attention to the speaker's eyes captured the first part of this process. The correlation between attention to eyes and language learning is consistent with prior research linking eye gaze following and learning language (Baldwin, 1991, 1993; Baldwin et al., 1996; Brooks & Meltzoff, 2005; Dunham, Dunham & Curwin, 1993).

Although attending to the speaker's eyes was useful for some toddlers and even though some children may have also engaged in joint attention with the speaker, it is unlikely that children in the current study relied solely on eye gaze or joint attention to learn novel words. If these factors drove language learning, we would have expected children in both the video chat and yoked video training conditions to learn the novel words since there was no difference in looking to the experimenter's eyes between conditions. However, this finding did not emerge; instead, the real predictor of success in the current study was whether or not children personally experienced a socially contingent interaction.

In a sense, it is surprising that attention to eyes during video chat training was helpful to children at all, since the experimenter's eye gaze on video appeared to be cast downward due to the relative placement of the camera and the computer screen. Thus, even though the experimenter's eyes in video chats moved appropriately, and were properly aligned for objects and actions within the experimenter's environment, eye gaze was nevertheless distorted from the child's perspective. That is, when the experimenter looked at the child on her screen, the experimenter appeared to be looking toward the bottom of the screen from the child's point of view. Given that children are adept at noting eye gaze and not simply body posture or head orientation (Brooks & Meltzoff, 2005), children in the current study attended to misaligned eye gaze. Although infants show a preference to direct, as opposed to averted eye gaze (Blass & Camp, 2001; Hood et al., 2003; Farroni et al., 2007), little is known about how toddlers might correct for mismatches in gaze alignment.

Screen Media for Children

In addition to uncovering social contingency as a possible tool for children's learning, we also replicated the finding that children younger than three years do not learn language from video alone (Krcmar et al., 2007; Kuhl et al., 2003; Robb et al., 2009; Roseberry et al., 2009) in the yoked control. Decades of research demonstrate that children older than three years learn robustly from video (Reiser et al., 1984; Rice & Woodsmall, 1988; Singer & Singer, 1998). The dual representation hypothesis may account for this developmental shift in children's ability to learn from video, suggesting that toddlers have trouble using television as a source of information (DeLoache, 1987; Troseth & DeLoache, 1998). According to the dual representation hypothesis, young children are so attracted to the salient, concrete aspects of the television that they cannot also understand its abstract, symbolic nature (DeLoache, 1987; Troseth & DeLoache, 1998). Further support for the dual representation hypothesis comes from experiments in which children are convinced that the events on video are realistic and occurring live. Under these circumstances, children under three years can learn from screen media (Roseberry, Hirsh-Pasek & Golinkoff, 2010; Troseth & DeLoache, 1998). Video chats might help young children circumvent dual representation through a platform for socially contingent interactions. Although older children may be able to compensate for the lack of contingency in traditional videos, this format may be particularly useful for the youngest children.

The success of video chatting is important for young children given the increasing popularity of this medium. By one estimate, video chatting has increased by 900 percent since 2007, with more than 300 million minutes of video chats taking place on Skype daily

(Scelfo, 2011). Regardless of its popularity, however, video chatting technology affords researchers the opportunity to disentangle contingency from other social cues. Our findings suggest that language learning occurs in the socially contingent interactions made possible via video chat. Researchers in other domains have also capitalized on the popularity of video chatting, using the medium with children of various ages to investigate the possibility of virtual play dates among school-aged children (Yarosh, Inkpen & Brush, 2010), and to test children's ability to maintain attachment relationships to caregivers across virtual space (Tarasuik, Galligan & Kaufman, 2010).

Although our findings support the utility of video chats, they also hint at how adept children are at distinguishing real contingency from other types of interactions. Our results from the yoked video condition indicate that simply posing questions to children and pausing for the answer did not result in language learning if the children were not able to interact contingently with the person on video. In fact, it may be more beneficial for young children to witness two characters interacting with each other on screen than for the characters to attempt to talk with children directly. Toddlers seem to learn better from watching a social interaction on video than from being directly addressed through video (O'Doherty, Troseth, Shimpi, Goldenberg, Akhtar & Saylor, 2011). This has important implications for children's media. Many children's television shows have attempted to incorporate interactions into their shows by posing questions to the unseen audience, waiting for an answer, and then responding (Anderson, Bryant, Wilder, Sontomero, Williams & Crawley, 2000; Fisch & McCann, 1993). For example, the host of *Blue's Clues* will often address the camera (i.e., “Do you see a clue?”), pause for a few seconds, and then respond (i.e., “There it is! You're right!”; Troseth et al., 2006). Our results suggest that children as young as 24 months can distinguish this one-sided “ask and wait” model from actual contingent interactions.

One caveat, however, is that the current study used children's *names* once at the beginning of the warm-up period that began the training phase (Troseth et al., 2006). Consequently, children in the yoked video condition heard the wrong name during the prerecorded warm-up. Research suggests that children recognize their name by 4.5 months and presumably expect to hear it from adults (Mandel, Jusczyk, & Pisoni, 1995). As children may have been confused when the experimenter addressed them by the incorrect name, it is possible that an “ask and wait” model that does not use names would produce more learning than the yoked video condition of the current study. Yet, because social contingency is likely established by reinforcement over time (Csibra, 2010) and because children are attuned to changes in the reliability of the speaker (Scofield & Behrend, 2008), there may be a second interpretation. It is possible that children could have compensated for the one inaccurate piece of information (i.e., the name used at the beginning of the warm-up) if subsequent content had been timely, reliable, and accurate. That is, additional socially contingent interaction might have provided children with enough data to determine that the speaker was accurate in all but one instance and therefore sufficient as a social partner. Future research should examine this possibility.

As the entertainment industry becomes more technologically advanced, the ability to incorporate live interactions into media would transform passive viewing experiences into socially contingent learning situations. Thus, children's learning from media may not be a product of the medium per se (i.e., video chat, video or live interaction), but rather the type of interaction children experience with the media.

Conclusions

This study takes the first step toward uncovering the mechanisms responsible for why children can learn from social interactions, but not from video. Socially contingent interactions, like those in video chats and live interactions, provided toddlers with sufficient social information to learn language. The results of this study not only addresses contingency as a critical social cue, but also highlights the capability of screen media to capitalize on the power of social contingency.

Acknowledgments

This research was supported by NICHD grant 5R01HD050199 and NSF BCS-0642529 to the second and third authors.

We thank Russell Ritchie for his assistance in data collection and Tilbe Göksun for valuable discussions on study design.

References

Adamson LB, Bakeman R, Deckner DF. The development of symbol-infused joint engagement. Child Development. 2004; 75:1171–1187.10.1111/j.1467-8624.2004.00732.x [PubMed: 15260871]

Anderson DR, Pempek TA. Television and very young children. American Behavioral Scientist. 2005; 48:505–522.10.1177/0002764204271506

Anderson DR, Bryant J, Wilder A, Santomero A, Williams M, Crawley AM. Researching Blue's Clues: Viewing behavior and impact. Media Psychology. 2000; 2:179–194.10.1207/S1532785XMEP0202_4

Baldwin D. Infants' contribution to the achievement of joint reference. Child Development. 1991; 62:875–890.10.2307/1131140 [PubMed: 1756664]

Baldwin DA. Infants' ability to consult the speaker for clues to word reference. Journal of Child Language. 1993; 20:395–418.10.1017/S0305000900008345 [PubMed: 8376476]

Baldwin, DA.; Bill, B.; Ontai, LL. Infants' tendency to monitor others' gaze: Is it rooted in intentional understanding or a result of simple orienting?; Paper presented at the International Conference on Infant Studies; Providence, RI. 1996 May.

Barr R, Wyss N. Reenactment of televised content by 2-year olds: Toddlers use language learned from television to solve a difficult imitation problem. Infant Behavior & Development. 2008; 31:696–703.10.1016/j.infbeh.2008.04.006 [PubMed: 18514319]

Batki A, Baron-Cohen S, Wheelwright S, Connelan J, Ahluwalia J. Is there an innate gaze module? Evidence from human neonates. Infant Behavior and Development. 2000; 23:223–29.10.1016/S0163-6383(01)00037-6

Beebe B, Steele M, Jaffe J, Buck KA, Chen H, Cohen P, et al. Maternal anxiety symptoms and mother-infant self- and interactive contingency. Infant Mental Health Journal. 2011; 32:174–206.10.1002/imhj.20274

Blass EM, Camp CA. The ontogeny of face recognition: Eye contact and sweet taste induce face preference in 9- and 12-week-old human infants. Developmental Psychology. 2001; 37:762–74.10.1037/0012-1649.37.6.762 [PubMed: 11699751]

Bloom K, Russell A, Wassenberg K. Turn taking affects the quality of infant vocalizations. Journal of Child Language. 1987; 14:211–227.10.1017/S0305000900012897 [PubMed: 3611239]

Bloom P. Mindreading, communication and the learning of names for things. Mind & Language. 2002; 17:37–54.10.1111/1468-0017.00188

Bornstein MH, Tamis-LeMonda CS, Hahn C, Haynes OM. Maternal responsiveness to young children at three ages: Longitudinal analysis of a multidimensional, modular, and specific parenting construct. Developmental Psychology. 2008; 44:867–874.10.1037/0012-1649.44.3.867 [PubMed: 18473650]

Brand RJ, Tapscott S. Acoustic packaging of action sequences by infants. Infancy. 2007; 11:321–332.10.1111/j.1532-7078.2007.tb00230.x

Brooks R, Meltzoff AN. The development of gaze following and its relation to language. Developmental Science. 2005; 8:535–543.10.1111/j.1467-7687.2005.00445.x [PubMed: 16246245]

Bruner JS. The social context of language acquisition. Language and Communication. 1981; 1:155–178.10.1016/0271-5309(81)90010-0

Catmur C. Contingency is crucial for creating imitative responses. Frontiers in Human Neuroscience. 2011; 5:1–2.10.3389/fnhum.2011.00015 [PubMed: 21283556]

Childers JB, Tomasello M. Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. Developmental Psychology. 2002; 38:967–978.10.1037/0012-1649.38.6.967 [PubMed: 12428708]

Choi S, Gopnik A. Early acquisition of verbs in Korean: A cross-linguistic study. Journal of Child Language. 1995; 22:497–529.10.1017/S0305000900009934 [PubMed: 8789512]

Csibra G. Recognizing communicative intentions in infancy. Mind & Language. 2010; 25:141–168.10.1111/j.1468-0017.2009.01384.x

DeLoache JS. Rapid change in the symbolic functioning of very young children. Science. 1987; 238:1556–1557.10.1126/science.2446392 [PubMed: 2446392]

Dunham PJ, Dunham F, Curwin A. Joint-attentional states and lexical acquisition at 18 months. Developmental Psychology. 1993; 29:827–831.10.1037/0012-1649.29.5.827

Farroni T, Massaccesi S, Menon E, Johnson MH. Direct gaze modulates face recognition in young infants. Cognition. 2007; 102:396–404.10.1016/j.cognition.2006.01.007 [PubMed: 16540101]

Fisch SM, McCann SK. Making broadcast television participate: Eliciting mathematical behavior through Square One TV. Educational Technology Research and Development. 1993; 41:103–109.10.1007/BF02297360

Gentner, D. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In: Kuczaj, S., editor. Language development: Language, thought, and culture. Vol. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1982. p. 301-334.

Gergely G, Watson JS. The social biofeedback theory of parental affect-mirroring: The development of emotional self-awareness and self-control in infancy. The International Journal of Psychoanalysis. 1996; 77:1181–1212.

Gleitman LR, Cassidy K, Nappa R, Papafragou A, Trueswell JC. Hard words. Language Learning and Development. 2005; 1:23–64.10.1207/s15473341lld0101_4

Goldstein MH, King AP, West MJ. Social interaction shapes babbling: Testing parallels between birdsong and speech. Proceedings of the National Academy of Sciences. 2003; 100:8030–8035.10.1073/pnas.1332441100

Goldstein MH, Schwade JA, Bornstein MH. The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. Child Development. 2009; 80:636–644.10.1111/j.1467-8624.2009.01287.x [PubMed: 19489893]

Golinkoff RM, Hirsh-Pasek K. How toddlers begin to learn verbs. Trends in Cognitive Science. 2008; 12:397–403.10.1016/j.tics.2008.07.003

Hassenfeld, J.; Nosek, D.; Clash, K. Beginning together [Motion picture]; Available from Sesame Workshop; One Lincoln Plaza, New York, New York, 10023. 2006a.

Hassenfeld, J.; Nosek, D.; Clash, K. Make music together [Motion picture]; Available from Sesame Workshop; One Lincoln Plaza, New York, New York, 10023. 2006b.

Henning A, Striano T. Infant and maternal sensitivity to interpersonal timing. Child Development. 2011; 82:916–931.10.1111/j.1467-8624.2010.01574.x [PubMed: 21410930]

Hirsh-Pasek, K.; Golinkoff, RM. The origins of grammar: Evidence from early language comprehension. Cambridge, MA: MIT Press; 1996.

Hollich GJ, Hirsh-Pasek K, Golinkoff RM, Hennon E, Chung HL, Rocroi C, Brand RJ, Brown E. Breaking the language barrier: An emergentist coalition model for the origins of word learning. Monographs of the Society for Research in Child Development. 2000; 65(3, Serial No. 262)10.1111/1540-5834.00090

Hood BM, Macrae CN, Cole-Davies V, Dias M. Eye remember you! The effects of gaze direction on face recognition in children and adults. Developmental Science. 2003; 6:69–73.10.1111/1467-7687.00256

Koenig MA, Harris PL. Preschoolers mistrust ignorant and inaccurate speakers. Child Development. 2005; 76:1261–1277.10.1111/j.1467-8624.2005.00849.x [PubMed: 16274439]

Krcmar M, Grela B, Lin K. Can toddlers learn vocabulary from television? An experimental approach. Media Psychology. 2007; 10:41–63.

Kuhl PK, Tsao F, Liu H. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. PNAS. 2003; 100:9096–9101.10.1073/pnas.1532872100 [PubMed: 12861072]

Lauricella AR, Pempek TA, Barr R, Calvert SL. Contingent computer interactions for young children's object retrieval success. Journal of Applied Developmental Psychology. 2010; 31:362–369.10.1016/j.appdev.2010.06.002

Mandel DR, Jusczyk PW, Pisoni DB. Infants' recognition of the sound patterns of their own names. Psychological Science. 1995; 6:314–317.10.1111/j.1467-9280.1995.tb00517.x

Markman, EM. Categorization and naming in children: Problems of induction. Cambridge, MA: MIT Press; 1989.

Moll H, Tomasello M. How 14- and 18-month-olds know what others have experienced. Developmental Psychology. 2007; 43:309–317.10.1037/0012-1649.43.2.309 [PubMed: 17352541]

Murray L, Trevarthen C. The infant's role in mother-infant communication. Journal of Child Language. 1986; 13:15–29.10.1017/S0305000900000271 [PubMed: 3949895]

Naigles LR, Bavin EL, Smith MA. Toddlers recognize verbs in novel situations and sentences. Developmental Science. 2005; 8:424–431.10.1111/j.1467-7687.2005.00431.x [PubMed: 16048515]

O'Doherty K, Troseth GL, Shimpi PM, Goldenberg E, Akhtar N, Saylor MM. Third-party social interaction and word learning from video. Child Development. 2011; 82:902–915.10.1111/j.1467-8624.2011.01579.x [PubMed: 21418054]

Pruden SM, Hirsh-Pasek K, Golinkoff R, Hennon E. The Birth of Words: Ten-Month-Olds Learn Words Though Perceptual Salience. Child Development. 2006; 77:266–280.10.1111/j.1467-8624.2006.00869.x [PubMed: 16611171]

Reiser RA, Tessmer MA, Phelps PC. Adult-child interaction in children's learning from “Sesame Street”. Educational Communication & Technology Journal. 1984; 32:217–223.

Rice M, Woodsmall L. Lessons from television: Children's word-learning while viewing. Child Development. 1988; 59:420–429.10.2307/1130321 [PubMed: 3359862]

Robb MB, Richert RA, Wartella E. Just a talking book? Word learning from watching baby videos. British Journal of Developmental Psychology. 2009; 27:27–45.10.1348/026151008X320156 [PubMed: 19972661]

Roseberry, S.; Hirsh-Pasek, K.; Golinkoff, RM. Honey, we shrunk the Sesame characters! Going beyond symbols to increase language learning Paper presented in S Roseberry and K Hirsh-Pasek (chairs). Why can't young children learn from television? Two potential explanations; The XVIIth International Conference on Infant Studies; Baltimore, MD. 2010 Mar.

Roseberry S, Hirsh-Pasek K, Parish-Morris J, Golinkoff RM. Live action: Can young children learn verbs from video? Child Development. 2009; 80:1360–1375.10.1111/j.1467-8624.2009.01338.x [PubMed: 19765005]

Sabbagh MA, Baldwin D. Learning words from knowledgeable versus ignorant speakers: Link between preschoolers' theory of mind and semantic development. Child Development. 2001; 72:1054–1070.10.1111/1467-8624.00334 [PubMed: 11480934]

Scelfo, J. Video chat reshapes domestic rituals [Electronic version]. The New York Times. 2011 Dec 21. Retrieved December 23, 2011 from www.nytimes.com

Scofield J, Behrend DA. Learning words from reliable and unreliable speakers. Cognitive Development. 2008; 23:278–290.10.1016/j.cogdev.2008.01.003

Scofield J, Williams A. Do 2-year-olds disambiguate and extend words learned from video? First Language. 2009; 29:228–240.10.1177/0142723708101681

Senju A, Csibra G. Gaze following in human infants depends on communicative signals. Current Biology. 2008; 18:668–671.10.1016/j.cub.2008.03.059 [PubMed: 18439827]

Senju A, Csibra G, Johnson MH. Understanding the referential nature of looking: Infants' preference for object-directed preference. Cognition. 2008; 108:303–319.10.1016/j.cognition.2008.02.009 [PubMed: 18371943]

Seston R, Golinkoff RM, Ma W, Hirsh-Pasek K. Vacuuming with my *mouth*? Children's ability to comprehend novel extensions of familiar verbs. Cognitive Development. 2009; 4:113–124.10.1016/j.cogdev.2008.12.001 [PubMed: 20161104]

Singer, JL.; Singer, DG. Barney & Friends as entertainment and education: Evaluating the quality and effectiveness of a television series for preschool children. In: Asamen, JK.; Berry, GL., editors. Research paradigms, television, and social behavior. Thousand Oaks: Sage Publications; 1998. p. 305-367.

Tarasuik, J.; Galligan, R.; Kaufman, J. Maintaining familial relationships via video communication; Poster presented at the The XVIIth International Conference on Infant Studies; Baltimore, MD. 2010 Mar.

Tardif T. Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabulary. Developmental Psychology. 1996; 32:492–504.10.1037/0012-1649.32.3.492

Tomasello, M. Pragmatic contexts for early verb learning. In: Tomasello, M.; Merriman, WE., editors. Beyond the names for things: Young children's acquisition of verbs. Hillsdale, NJ: Lawrence Erlbaum Associates; 1995. p. 115-146.

Tomasello M, Barton M. Learning words in nonostensive contexts. Developmental Psychology. 1994; 30:639–650.10.1037/0012-1649.30.5.639

Troseth GL, DeLoache JS. The medium can obscure the message: Young children's understanding of video. Child Development. 1998; 69:950–965.10.2307/1132355 [PubMed: 9768480]

Troseth GL, Saylor MM, Archer AH. Young children's use of video as socially relevant information. Child Development. 2006; 77:786–799.10.1111/j.1467-8624.2006.00903.x [PubMed: 16686801]

Yarosh, S.; Inkpen, KI.; Brush, AB. Proc Of CHI. ACM; 2010. Video playdate: Toward free play across distance; p. 1251-1260.

Zimmerman FJ, Christakis DA, Meltzoff AN. Associations between media viewing and language development in children under age 2 years. Journal of Pediatrics. 2007; 151:364–368.10.1016/j.jpeds.2007.04.071 [PubMed: 17889070]

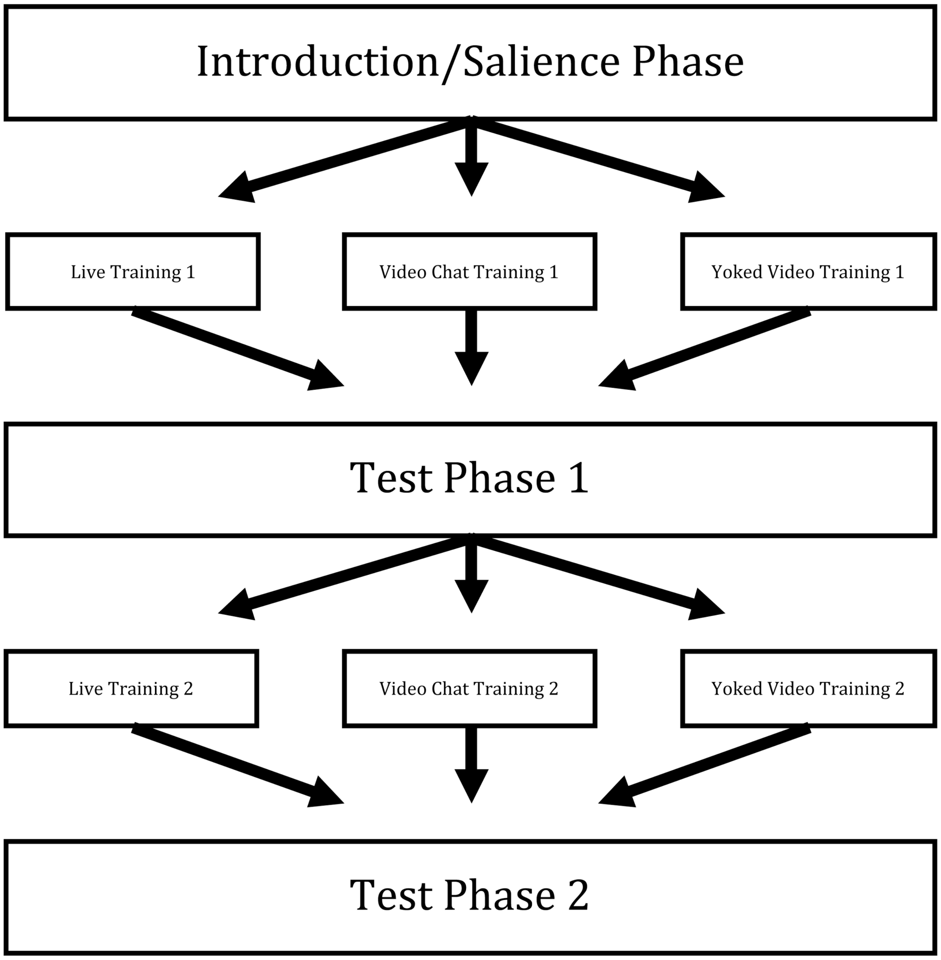
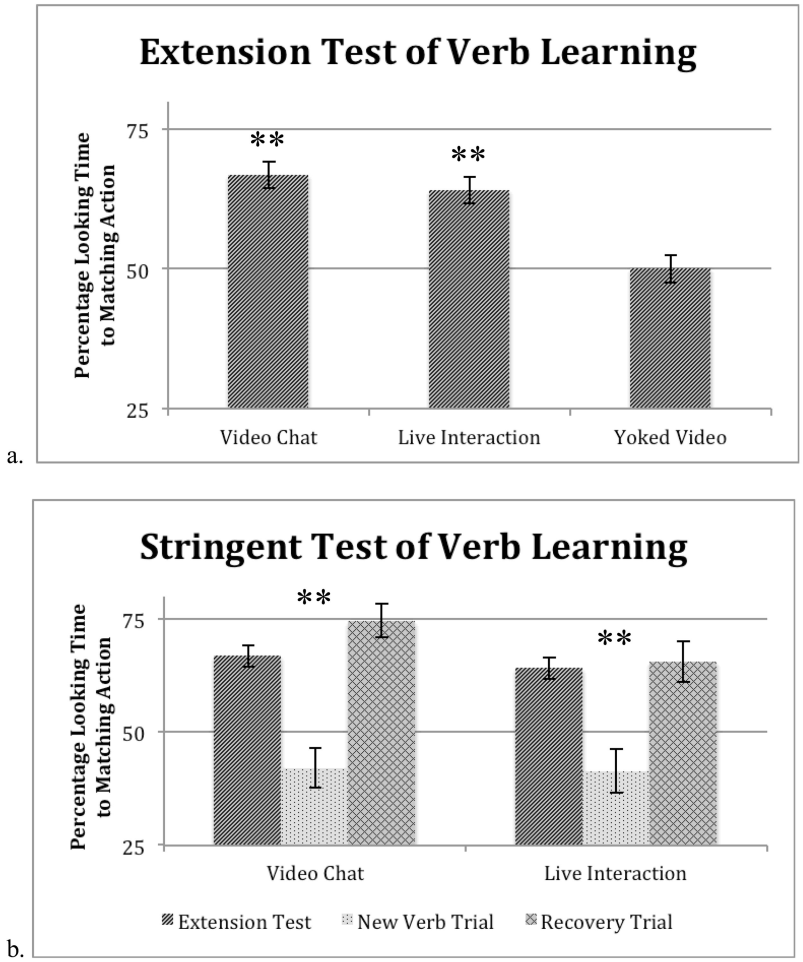


Figure 1.

Flow chart depicting the study design. The only difference between conditions was the mode of training (i.e., Live Training, Video Chat Training, or Yoked Video Training).



Extension Test of Verb Learning

Stringent Test of Verb Learning

Figure 2.

In the Extension Test of Verb Learning (a), children trained via video chat and live interaction looked significantly longer toward the matching action at test, $p < .01$, whereas children trained through yoked video performed at chance; their looking times to either side of the screen were not different, $p > .05$. In the Stringent Test of Verb Learning (b), conducted only for the Video Chat and Live Interaction conditions, children looked longer toward the matching action during the Extension Test, looked away during the new verb trial and then resumed looking to the matching action during the recovery trial $p < .01$. Error bars represent the standard error of the mean. ** $p < .01$.

Table 1

The Novel Verbs, Their English Approximations, and Descriptions of the Familiarization and Test Actions.

Novel Verb	English Equivalent	Description of Familiarization	Description of Test	
			Matching Action	Non-matching Action
Blicking	Bouncing	Adult moves a doll up and down on knee	Elmo's dad moves baby Elmo up and down on knee	Cookie Monster's grandma holds baby Cookie Monster in her arms and twists side to side
Twilling	Swinging	Adult holds a doll in arms and rotates from side to side	Elmo's dad holds baby Elmo and rotates from side to side	Baby Big Bird holds a teddy bear and talks while wiggling
Frepping	Shaking	Adult moves a rattle in hand from side to side rapidly	Prairie Dawn moves a box in hand from side to side rapidly	Elmo places his hands on a block as he stands up
Meeping	Turning	Adult turns the rotor blades on a toy helicopter	Elmo's dad turns the dial on a radio	Baby Cookie Monster runs with a toy airplane in hand