This is the preprint of an article submitted for publication. Please cite as:

Tsuji, S., Fiévét, Anne-Caroline, & Cristia, Alejandrina (preprint). Toddler word learning from contingent screens with and without human presence. Manuscript submitted for publication. Retrieved from: https://psyarxiv.com/7hxjk

Toddler word learning from contingent screens with and without human presence

Sho Tsuji[1,2], Anne-Caroline Fiévét[2], Alejandrina Cristia[2]

[1] The University of Tokyo

[2] Laboratoire de Sciences Cognitives et Psycholinguistique, Departement d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University

Author Note

Declarations of interest: none

The research was conducted in accordance with APA ethical standards in the treatment of the human study sample.

Correspondence concerning this article should be addressed to Sho Tsuji, International Research Center for Neurointelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 Japan, Contact: tsujish@gmail.com

Abstract

While previous studies have documented that toddlers learn less well from passive screens than from live interaction, the rise of interactive, digital screen media opens new perspectives, since some work has shown that toddlers can learn similarly well from a human present via video chat as from live exposure. The present study aimed to disentangle the role of human presence from other aspects of social interactions on learning advantages in contingent screen settings. We assessed 16-month-old toddlers' fast mapping of novel words from screen in three conditions: in-person , video chat, and virtual agent. All conditions built on the same controlled and scripted interaction. In the in-person condition, toddlers learned two novel word-object associations from an experimenter present in the same room and reacting contingently to infants' gaze direction. In the video chat condition,tthe toddler saw the experimenter in real time on screen, while the experimenter only had access to the toddler's real-time gaze position as captured by an eyetracker. This setup allowed contingent reactivity to the toddler's gaze while controlling for any cues beyond these instructions. The virtual agent condition was programmed to follow the infant's gaze, smile, and name the object with the same parameters as the experimenter in the other conditions. After the learning phase, all toddlers were tested on their word recognition in a looking-while-listening paradigm. Comparisons against chance revealed that toddlers showed above-chance word learning in the in-person group only. Toddlers in the virtual agent group showed significantly worse performance than those in the in-person group, while performance in the video chat group overlapped with the other two groups. These results confirm that in-person interaction leads to best learning outcomes even in the absence of rich social cues They also elucidate that contingency is not sufficient either, and that in order for toddlers to learn from interactive digital media, more cues to social agency are required.

Keywords: Interactive screen media, temporal contingency, video deficit effect, word learning, gaze-contingent eye-tracking

Toddler word learning from contingent screens with and without human presence

1. Introduction

The role of new forms of digital media play for toddler's language learning is a topic inciting heated discussions within and beyond the scientific community. Classical, passive screen media like TV or videos compete today with a multitude of novel digital formats such as tablets, smartphones, and ebooks, to which toddlers' are increasingly exposed. For instance, the percentage of US children below two years of age that have used interactive screen media like a smartphone increased from 10% to 38% between 2011 and 2014 alone (Rideout & Saphir, 2013). Apps promoting games for young children are top sellers in the App Store (Shuler, 2012), and more and more young children interact with family members and friends via video chat (McClure, Chentsova-Dutton, Barr, Holochwost, & Parrott, 2016), a trend that likely deepened with the recent sanitary crisis. These new formats introduce novel possibilities for toddlers to interact with digital media, notably devices and games that are responsive, as well as with a person on screen in real time. How these new digital formats impact early language learning is a continuous debate. The present study focuses on the interactivity inherent in many of these new formats and investigates to what extent this interactivity, in the presence or absence of a human interaction partner, can affect learning advantage from these digital media formats.

1.1 Learning from passive screen media

The conditions under which toddlers can or cannot learn from traditional, passive screen media like TV or video have been studied in depth. The *video deficit effect* describes the observation that toddlers systematically learn worse from screens than from live demonstrations in tasks ranging from imitation to object retrieval (for an overview, see Anderson & Pempek, 2005). In the domain of early language learning, it has been

demonstrated that repeated live exposure leads to better learning of speech sounds compared to closely matched video exposure in 9-month-old infants (Kuhl, Tsao, & Liu, 2003). Infants aged 5-6 and 9-10 months were also capable of word segmentation and learning a novel word-object association when taught by a live teacher, but not a televised one (Hakuno, Omori, Yamamoto, & Minagawa, 2017). Further, 1-to 2-year-old toddlers acquire vocabulary significantly better from a teacher interacting with them in real life than from a TV program targeted at teaching words (DeLoache et al., 2010; Krcmar, Grela, & Lin, 2007). In addition to these experimental studies, many observational studies have reported a negative relationship between young children's viewing of passive screens and their language skills (for an overview, e.g., Madigan, McArthur, Anhorn, Eirich, & Christakis, 2020).

Given the steep rise in young children's exposure to novel, interactive forms of screen media over the past decade, a key question for the field concerns to what extent we can generalize findings from traditional screen media to new digital formats. A particularly intriguing question concerns the role of the interactivity afforded by such new formats for early language learning, since, as discussed in the following section, infants are sensitive to environmental contingencies early on.

1.2 Contingent reactivity and early language learning

Infants are sensitive to environmental contingencies early on. Within their first few months of life, they develop sensitivity to the content and timing of contingent responsiveness, noticing, for instance, if an interaction partner suddenly stops responding or if responsiveness is irregular and delayed (for an overview, see Striano & Reid, 2006). The beneficial effects of caregivers' contingent responsiveness have been demonstrated for multiple aspects of language acquisition. For instance, infants in their first year of life increase the number and quality of vocalizations (Goldstein, King, & West, 2003; Goldstein

& Schwade, 2008; see Dunst, Gorman, & Hamby, 2010 for an overview) and gestures (Miller & Lossia, 2013) in response to contingently reacting caregivers. Contingent responsiveness has also been shown to have positive effects on perceptual narrowing in 6-month-olds (Elsabbagh et al., 2013) as well as other language milestones such as first words and combinatorial speech during the second year of life (Tamis-LeMonda, Bornstein, & Baumwell, 2001). Considering these results, it comes as no surprise that toddlers learn better from in-person interaction compared to passive video exposure. The following section discusses the effects of putting the contingent responsiveness found in natural interaction onto a screen via video chat.

1.3 Learning from video chat

Under certain conditions, video chat has been reported to lead to learning outcomes comparable to those obtained by in-person interaction, and superior to those in passive video viewing conditions (e.g., Roseberry, Hirsh-Pasek, & Golinkoff, 2014; Troseth, Saylor, & Archer, 2006). This is a fascinating finding in view of the reported video deficit effect, suggesting that a human interaction partner on screen can alleviate this deficit. The practical relevance of this finding is reflected in updated recommendations by the American Academy of Pediatrics, which recommend against any screen time for children under the age of 2, unless it is live video chat (Chassiakos, Radesky, Christakis, Moreno, & Cross, 2016). Similarly, parental reports suggest that caregivers who are cautious about exposing their young children to screens are still ready to let them participate in live video chat (McClure et al., 2016).

As mentioned above, young children learn better from a real person than from an asynchronous video of that same person (e.g., Kuhl et al., 2003; Hakuno et al., 2017; Roseberry, et al., 2014), despite the fact that these videos often contain a multitude of

engaging social cues such as smiling, direct gaze, or infant-directed speech, all of which are thought to boost learning in social situations. Two potentially critical remaining differences between in-person versus asynchronous video learning conditions would be (1) the fact that asynchronous video requires the child to process social cues via a screen, and (2) the lack of contingent responsiveness in the asynchronous video condition. Regarding the former, and as suggested in the context of the video deficit effect, it is possible that processing stimuli via a screen is harder (because cues of depth, for instance, would be missing). As to the latter, multiple studies now support the view that contingency (which may be a crucial cue to social agency) is critical for learning success. Comparing in-person live exposure, live exposure via closed circuit video, or pre-recorded video exposure, 2-year-old children were found to retrieve an object significantly better in both live conditions compared to the pre-recorded video condition (Troseth et al., 2006). However, the closed circuit video condition was only successful if preceded by several minutes of meaningful contingent interaction with the experimenter on screen. In a first version of this condition, in which toddlers learned significantly worse compared to live exposure, toddlers saw the experimenter talking to them and hiding the toys on screen, but had little occasion to experience the possibility to interact with the person they saw on screen. In contrast, once the hiding event was preceded by a phase consisting of interactive games and a conversation with questions, toddlers learned significantly better in the on-screen condition compared to asynchronous video presentation. Given that toddlers have little experience with synchronous video presentation, such a phase might be necessary to perceive a live agent on screen as a social agent. In the domain of language acquisition, two recent studies have reported similar effects. Three-year-old children have been found to learn novel verbs similarly well from a live exposure and a live video chat exposure condition, but not from yoked video (Roseberry et al., 2014). In a study assessing learning of novel nouns, toddlers 22 months and 25 months of age (but not

younger) exposed to video chat for a week showed better word learning than a group exposed to yoked video over the same period of time (Myers, LeWitt, Gallo, & Maselli, 2017). Both of these studies implemented contingent responsiveness whereby experimenters involved children in an interaction that was meaningful, relevant, and appropriate in content, and thus rich in social cues. In the study by Roseberry and colleagues (2014), for instance, children in the video chat condition engaged in a pre-learning interaction phase, in which they were called by their name, asked questions about the toys they played with, and given affectionate feedback on their actions by the on-screen human. Children in the yoked condition watched a video extracted from the live interaction videos of the video chat condition. Together, these results suggest that such *social* contingency, thus a type of contingency that identifies the interaction partner as a social agent, is necessary to support learning. Converging evidence comes from another recent study where a positive learning effect from video chat was not obtained: 24- and 30-month old toddlers succeeded in learning a novel word-object association from live, but not live video chat interaction (Troseth, Strouse, Verdine, & Saylor, 2018). The authors indeed suggest divergent results are possibly due to a lack of information identifying the on-screen interaction partner as a potential social agent, which was established via a pre-learning interaction in previous studies (such as Troseth's object retrieval study, or even the week-long exposure to video in Myers' work). In addition, the task might have been more difficult for toddlers in the video chat group. The test phase was always administered with real objects in the 3-D world. Thus, only toddlers in the video groups had to transfer the 2-D content learned on a screen to recognizing word-object associations in the real 3-D world, whereas toddlers in the in-person condition did not have to make such a transfer.

Overall, these studies show that young children can learn better from interactive than from passive screens, at least if the interaction partner is a real person and the toddler is able

to identify her as a potential social agent. May interactive digital screen media with these features allow learning? The following section summarizes research in this domain.

1.4 Learning from interactive media in the absence of humans

Only a few studies have assessed to what extent temporal contingency on screen in the absence of a human interaction partner would enhance word learning in toddlers. In a recent touch-screen study, the pre-recorded video of a woman labelling novel objects hidden in various boxes was presented (Kirkorian, Choi, & Pempek, 2016). Twenty-four-month-old children in a specific contingent condition with instructions to touch a box on screen in order to see the object showed a word-learning advantage, but not children in a general contingent condition ("touch the screen") or a non-contingent condition ("watch the screen", where the video was advancing regardless of touching). This finding indicates that the mere addition of a temporally contingent element - the display proceeding to the next screen upon touch instead of automatically - can support word learning. The video did, however, contain a human teacher on screen, and an experimenter interacted with participants at various points during the experiment. Thus, it is still an open question whether the temporal contingency manipulation would be successful in the absence of a broader social context containing human agents. A more recent study controlled for these factors by displaying a virtual agent that was contingently reacting to 12-month-old infants' gaze via gaze-contingent eye-tracking, and teaching them novel word-object associations (Tsuji, Jincho, Mazuka, & Cristia, 2020). The contingent reactions displayed by the on-screen avatar included mutual gaze and gaze following, but no broader social context such as a prolonged preceding interaction phase. Note that, in comparison to previous studies, this avatar, while having human-like features like a face and extremities, did not resemble a human being. This study showed a learning advantage for infants learning from this avatar compared to an avatar that

did not react contingently, thus suggesting that contingent responsiveness can support learning even in the absence of a human interaction partner and a rich set of social cues.

Indeed, such a result is consistent with reports of infants' early sensitivity to on-screen temporal contingencies. For instance, 6- to 8-month infants learned to trigger the appearance of a novel stimulus on screen with their eye movements in a few trials' time, illustrating their sensitivities to contingency encountered on-screen (Wang et al., 2012). Contingency on screen also triggers social-like reactions, as illustrated in a study where 8-month-old infants were exposed to an amorphous object on screen, which either did or did not perform movements contingent on infants' gaze. Only if the object had been contingent, infants would later gaze-follow its turning direction (Deligianni, Senju, Gergely, & Csibra, 2011). The authors suggested that such relatively abstract temporal contingencies serve as an amodal cue to communication that might generate referential expectations by themselves. If so, the question as to why those video chat studies that were temporally contingent, but did not contain broader social contexts failed to lead to learning needs to be revisited, as perhaps the on-screen presentation did not contain sufficient cues to social agency.

1.5 The present study

Previous studies have established that toddlers can learn from digital media that allow temporally contingent responsiveness, but typically they also include social contingency involving humans, which may indicate that a rich and meaningful social context is crucial for learning success from such media.

In the present study, we assess to what extent toddlers can learn from situations with off- or on-screen contingent responsiveness in the absence of a broader set of cues to social agency. In a baseline group, we assess toddlers' learning of novel word-object associations from a live human interaction partner. However, the in-person interaction is completely

scripted and allows a rather minimal amount of social contingency. Concretely, the experimenter followed a script to name the displayed objects, and was instructed to say nothing else. We also instructed the experimenter to show contingent reactions like smiling and gaze following only during specific periods during a trial. Despite the reduction of social contingency, the in-person interaction cannot exclude the possibility that the experimenter would show spontaneous reactions while facing the toddler. We avoided this in a second condition, called the video chat group, by providing the experimenter only the information that the eye-tracking machine had. Thus, instead of seeing the toddler displayed on screen in real time, the experimenter saw the toddler's gaze position in real time, and was instructed to react accordingly. The third group of toddlers saw a virtual agent identical to the one used in Tsuji et al. (2020). Script and reactions of the virtual agent were matched to those of the experimenter in the video chat group. Test trials were administered on-screen for all three groups, thus, in our case the in-person group would be the only one facing a potential disadvantage due to having to transfer the learned content to a 2-D screen.

Overall, the comparisons of these three groups allowed us to assess whether toddlers would learn the novel word-object associations under these three conditions, and to what extent learning would differ between groups. Investigating these questions would allow us to assess the role of in-person presence and human presence in an interactive digital media context in the absence of a broader set of cues to social agency.

2. Methods

Data and analysis scripts as well as full analysis outputs can be found in our repository on the Open Science Framework (https://osf.io/e3ksb/wiki/home/?view_only=c6a6065a02f64bbb87984b55f70aaf19).

2.1 Participants

Ninety-six normally developing French-learning toddlers from the Paris region were included in the analysis, 32 in each of the three experimental groups, the in-person group (13 female, mean age = 505 days, range 492-518 days), the video chat interaction group (18 female, mean age = 501 days, range 490-518 days), and the virtual agent interaction group (15 female, mean age = 504 days, range 489-517 days). Fifteen additional toddlers (in-person: 5, video chat: 4, virtual agent: 6) were tested but excluded from analysis due to contributing too little data after our data cleaning criteria (described further below) were applied. The study was approved by the Ethics Committee of (name masked for anonymous review). Infants were recruited from the laboratory participant pool. Caregivers signed an informed consent prior to their inclusion in the study. Data were collected in 2017-2018.

2.2 Stimuli

Three groups of toddlers were taught novel word-object associations under different stimulation conditions. The in-person group was taught by a real person sitting in front of them in the experimental room; the video chat interaction group was taught by a real person interacting with them in real time via video chat on the eyetracker screen, and the virtual agent interaction group was taught by a cartoon-like virtual agent contingently reaction to them on the eyetracker screen. The virtual teacher was designed to have human-like facial and body features including eyes, a mouth, a torso, and extremities (see Figure 1). It was identical to the virtual agent used in Tsuji and colleagues (2020). In all three groups, the teacher first named two familiar objects, a baby bottle “biberon” and a dog “chien”. We used a fluffy toy dog and an actual baby bottle (or photos thereof, depending on condition) as stimuli. Toddlers were subsequently exposed to two pairs of novel word-object associations.

Which word was paired with which object was counterbalanced across infants. The novel objects were soft, smooth objects of similar size (see Figure 1). The names for these two objects [lagi, toʁba] were bisyllabic phonotactically legal French non-words. These non-words were matched on frequency of CV1 and CV2. All target novel and familiar words were embedded in carrier sentences that were either produced in real time (for the teaching phases of the live and video chat groups) or prerecorded by a female native speaker of French in infant-directed register. The full phrases are documented in our OSF repository (Supplementary Information A).

2.3 Procedure

For both the teaching and test phases in all three conditions, toddlers were seated on a caregiver's lap in a sound-attenuated room. Except for the teaching phase of the in-person group, where they faced a live experimenter, toddlers faced a screen with an EyeLink 1000 eye-tracker (SR Research Ltd., 2005) throughout the experiment. One experimenter was monitoring the study from outside. The caregiver wore headphones with masking music. In all groups, toddlers were exposed to eight teaching trials (two with familiar objects, and three each for each novel object), and six test trials (two with familiar objects, and two each for each novel object).

In the *in-person group*, the teaching trials were administered by one of two trained experimenters who were female native speakers of French, and who were present in the same room as the toddler. During these live trials, the caregiver's chair was positioned in a 90 degree angle away from the eyetracker screen (that was switched off). The experimenter was seated in front of the caregiver and toddler with a small table in front of her, and a box with the objects hidden behind a curtain (Figure 1). She started the experiment by waving at the

toddler and greeting them with a phrase translatable to “Hello! Would you like to play with me?” Subsequently, she initiated each teaching trial by pulling out one of the objects and placing it on the table in front of the toddler. The teaching trials were subdivided into a pre-naming phase, in which the infants had time to experience the interactivity of the situation, and a naming phase, in which the teacher named familiar or novel objects. In the pre-naming phase of each teaching trial, the experimenter was instructed to visually interact with the infant in a semi-naturalistic fashion for a few seconds. If the toddler looked at her face, she was instructed to make eye contact and smileback, and if the toddler looked elsewhere, she followed her gaze. When the toddler focused attention on the object, she was instructed to take it into her hand and wiggle it slightly. She then would turn to the object and name it three times, pointing at, looking at, and turning towards the object during each naming instance, and turning back her gaze to the infant between naming instances. After the teaching phase, the caregivers’ headphones were removed temporarily and they were asked to stand up in order to turn the chair to face the eyetracker screen.

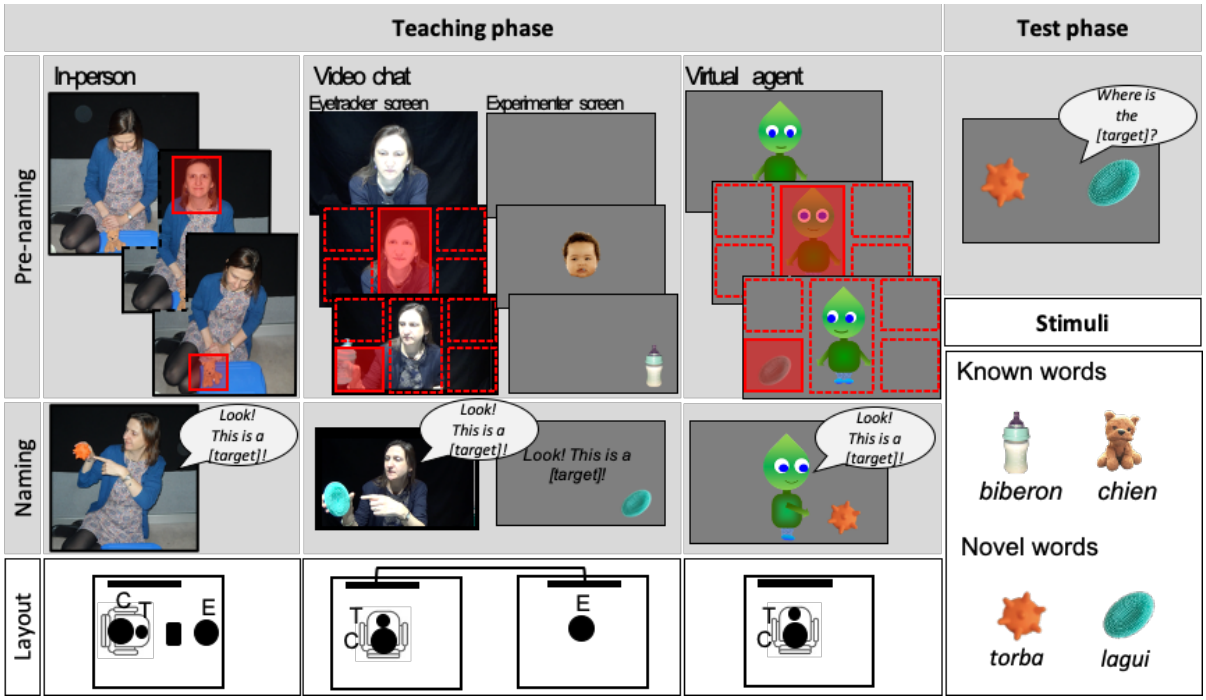

Figure 1. Experimental design. The teaching phase in each group was divided in a prenaming

and a naming phase. In the prenaming phase, the toddler experienced the teacher's contingent responsiveness. In the in-person condition, toddler (T) and caregiver (C) faced the experimenter (E). The teacher was instructed to look up and smile when the toddler looked at her face, and to follow their gaze otherwise. Red shaded rectangles indicate the toddler's present gaze focus. In the video chat condition, T and C faced the eyetracker screen displaying E in one experimental room, while E faced a screen displaying screen prompts in another room. In the virtual agent group, T and C faced the eyetracker screen displaying the virtual agent For both video chat and virtual agent groups, red dotted rectangles schematically indicate the regions of interest on screen that determined the teacher's reaction. The teacher looked up and smiled when infants looked at an area of interest around her face, and followed gaze to areas of interest in the four corners of the screen. During the naming phase, the teacher repeatedly turned towards, looked at, and pointed at the named object. In test trials, two objects were presented side by side on screen while one of them was named. Both teaching and test phases started with the presentation of known word-object associations before moving to novel ones.

In the *video chat group*, one of the same two female experimenters was seated in front of a screen in an adjacent sound-attenuated cabin, while the toddler and caregiver were facing the eyetracker screen in the experimental room. On this screen, the toddler saw the experimenter in real time. In contrast, the experimenter's screen did not display the toddler herself, but provided instructions based on the toddler's gaze direction (as detected by the eyetracker, see Figure 1). Before each experiment, the experimenter carefully placed herself so that her face as well as her hands when she was holding an object were displayed on the same position on the eyetracker screen each time. The objects were otherwise hidden out of sight of the toddler. The experimenter started the experiment by waving at the toddler and greeting them. Subsequently, she initiated each trial by pulling out one of the objects and

holding it in her hand in view of the toddler. In the prenaming portion of the teaching phase,if the toddler looked at the center of the screen (where she saw the experimenter's face), the experimenter saw the face of a smiling infant in the center of the screen, and was instructed to look up and smile. If the toddler looked to one of the corners of the screen, the experimenter saw a square on the corresponding side of the screen, and was instructed to shift their gaze to that location. If that location coincided with the object location, the experimenter saw an object instead, which was her promit to wiggle the object slightly. This manipulation thus ensured the experimenter's contingent responsiveness while stripping her response off any spontaneous reactions that could have been evoked by actually seeing the toddler's face. After a few seconds, the screen displayed a prompt with a carrier sentence, as well as the picture of the object to be named on one side of the screen. The experimenter was instructed to read out the carrier sentence while pointing at and naming the object in the same way as in the in-person group. After the naming phase, the toddler's screen display switched to the computerized test phase.

In the *virtual agent group*, toddlers saw a virtual teacher on screen (Figure 1). This virtual teacher waved and greeted toddlers in the same way as the real experimenter in the live video chat group. Specifically, during the pre-naming phase of each trial, the teacher reacted as follows. When the toddler looked at the teacher's face she would look up and smile, and when the toddler looked to one of the corners, the teacher would follow toddlers' gaze. If the toddler looked to the object, the object would gently pulsate. The virtual teacher thus showed the same reactions based on the same cues as the real teacher in the video chat interaction group. In the naming phase, the teacher would turn towards and name the objects in the same way as the teacher in the other groups. All carrier sentences were matched across the three groups.

After the teaching phase, all toddlers received an identical test phase in all three groups. The test phase started with the two familiar objects being displayed side by side on the screen, and being named each once in two subsequent trials. After that, each of four test trials displayed the two previously learned objects side by side on the screen (side counterbalanced). After two seconds, infants heard a sentence in which one of the objects was named (looking-while-listening procedure, Fernald, Zangl, Portillo, & Marchman, 2008).

Toddlers' gaze was calibrated with an infant-friendly 5-point calibration. Their gaze was recorded with a sampling rate of 500 Hz. Calibration was performed at the beginning of the experiment for the video chat and virtual agent groups, and before the test phase for the in-person group. The screen-based parts of the experiment were administered using E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Preceding each screen-based trial, an attention getter (the picture of a flower) appeared centrally on screen, and the trial was initiated by the experimenter once the toddler's gaze was fixated on it.

2.4 Data Cleaning

Exclusion criteria for the eye-tracking data in test trials are identical to those in Tsuji et al. (2020). We focused on the time window between 400-2400 ms after target word onset for analysis of word recognition trials. This time-window was chosen to be close to previous studies using comparable designs (e.g., Mani & Plunkett, 2007), and accounts for the fact that toddlers need several hundred milliseconds to initiate a gaze shift (Fernald et al., 2008). Toddlers who did not complete the task were immediately excluded from analysis. Subsequently, data marked as saccades were excluded as recommended by the manufacturer (SR Research Ltd., 2005)[1]. Next, based on the remaining data points, we excluded trials in which infants were looking away from the screen for more than 75% of the time window of

[1] Since Tsuji et al. (2020) used another eyetracker, this step slightly differs.

analysis. This criterion excluded 14.9% of trials (10.6% for the in-person condition, 11.3% for the video chat condition, and 21.7% for the virtual agent condition). Based on the remaining trials, we further excluded four infants that had less than two trials left in the test phase (cf. Experiment 1). The remaining 96 infants had an average of 3.6 trials (SD = 0.62) left per group (3.7 in-person , 3.6 video chat, 3.7 virtual agent).

2.5 Data Analysis

All analyses were conducted in R version 3.6.1 (R Core Team, 2019) with the packages eyetrackingR version 0.1.8 (Dink & Ferguson, 2018) and lme4 version 1.1-21 (Bates, Maechler, Bolker, & Walker, 2015). Figures were made with ggplot2 version 3.2.1 (Wickham, 2016).

We performed two types of data analysis. First, we performed a classical time window analysis, in which we aggregated the mean proportion of looks per group in the time window of interest and ran a regression model on the difference between groups. Empirical logit transformation was performed on the proportions to accommodate the categorical nature of the data (fixating the target picture or not) in a way that is robust to values at or near the boundaries (0 and 1) (Barr, 2008). The in-person group served as the baseline against which the other two groups were compared, and the model took the form lm(Elog ~ Group). In order to also compare the video chat and virtual agent group against each other, we subsequently releveled the data. Second, as in Tsuji et al. (2020), we fitted a growth curve analysis (GCA) modeled after Mirman (2014). GCA accounts for the dynamic nature of gaze data by not only assessing overall differences in looking times but additionally differences in the shape and latency of the gaze curve. The time course of the word recognition effect was captured with third-order orthogonal polynomials and with fixed effects of condition on all time terms, as well as random effects of participant and trial on all time terms. Data were

grouped into 100ms bins. The original model took the form lmer(Elog ~ Group * (ot1+ot2+ot3) + (ot1+ot2+ot3|Subject) + (ot1+ot2+ot3|Trial), where ot1, ot2, and ot3 refer to the linear, quadratic, and cubic orthogonal polynomials. Due to convergence issues, we successively removed the random effects on me terms such that the final model took the form lmer(Elog ~ Group * (ot1+ot2+ot3) + (1|Subject) + (1|Trial)). Statistical significance (p-values) was assessed using normal approximation (i.e., treating the t-value as a z -value). For both window analysis and GCA, it is not only of interest whether groups differ, but also whether each individual group leads to above-chance word recognition. We therefore also inspected the model intercepts of two additional models in which the video chat or the virtual agent condition served as the comparison level.

3. Results

3.1 Known words

As a validity check, we first analyzed word recognition of the two known word-object associations, using our two pre-established analysis approaches. In the *window analysis*, the intercept term for the in-person group was significant, indicating above-chance word recognition (b = 0.981, SE = 0.471, t(94) = 2.08, p = .040, d = 0.286). The differences between the in-person condition and the video chat (b = -0.364, SE = 0.683, t(94) = -0.53, p = .595, d = 0.12) or virtual agent (b = -0.275, SE = 0.662, t(94) = -0.42, p = .678, d = 0.10) groups were non-significant. The *GCA analysis* likewise showed a significant intercept term for the in-person group (b = 1.007, SE = 0.383, t = 2.63, p = .009), and no differences between the in-person condition and the video chat (b = -0.353, SE = 0.538, t = -0.656, p = .512) or virtual agent (b = -0.080, SE = 0.554, t = -0.14, p = .886) groups. Thus, toddler's

recognition of known words did not differ between groups, demonstrating that toddlers across groups showed comparable understanding of the experimental task.

Unexpectedly, visual inspection of gaze trajectories showed a below-chance proportion of looks to target before naming in the video chat condition (Fig. 2). At this point, we do not have an explanation for this pattern; however, since it did not affect toddler behavior during the time-window of analysis, we will not discuss it further.

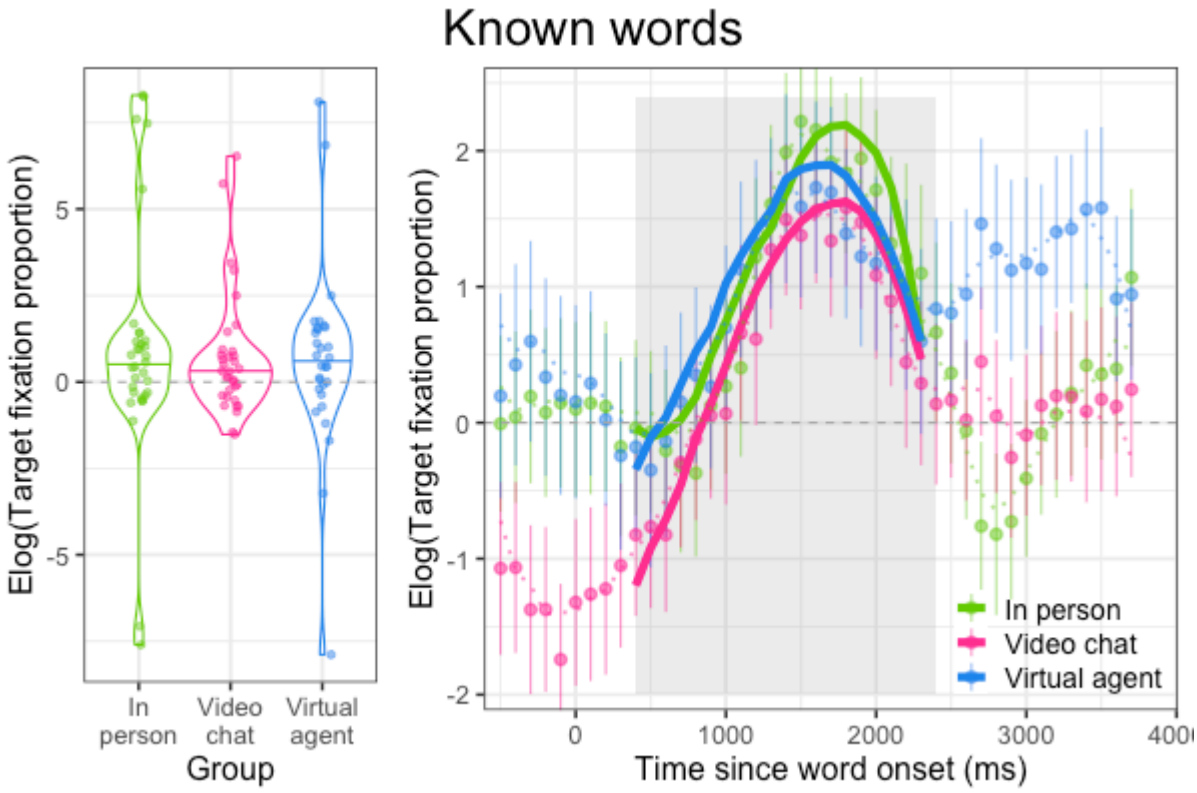

Figure 2. Empirical logit of mean target fixation proportion to correct word-object association in test phase for known words. Density plots on the left side indicate density estimates over mean target fixation proportion in 400-2400 ms time-window of analysis. Points represent means by participant, solid horizontal line represents median. Dashed line indicates chance level. Plot on the right represents time-course of looks to target around the time-window of analysis (shaded in grey). Gaze proportions were binned into 100 ms units. Dashed line indicates chance level. Circles and error bars represent the observed mean and

±1SE of the mean over each time bin. Solid lines represent model fits derived from GCA model reported in the main text.

3.2 Novel words

As with the known words, we investigated learning of the novel words with our two analytic approaches. In the *window analysis*, we found a significant intercept term for the in-person condition, indicating above-chance word recognition (b = 0.427, SE = 0.164, t(93)= 2.60, p = .011, d = 0.44). The models in which the video chat or virtual agent groups served as the baseline did not show significant intercepts (video chat: b = 0.127, SE = 0.164, t(93) = 0.77, p = .441, d = 0.18; virtual agent: b = -0.240, SE = 0.164, t(93) = -1.46, p = .148, d = -0.22).

As to group differences, we found a significant difference between the in-person and virtual agent groups, suggesting that toddlers learned better in the former condition (b = -0.667, SE = 0.232, t(93) = -2.87, p = .005, d = 0.66). In contrast, the video chat group did not significantly differ from the in-person group (b = -0.300, SE = 0.232, t(93) = -1.29, p = .199, d = 0.35) or the virtual agent group (b = 0.366, SE = 0.232, t(93) = 1.58, p = .118, d = 0.40), indicating toddlers' learning performance was not substantially different when there was a real person present on video chat from either of the other conditions.

The *GCA analysis* paralleled the results reported in the window analysis. It revealed a significant intercept term for the in-person condition, indicating above-chance word recognition (b = 0.738, SE = 0.255, t = 2.89, p = .004). Releveling the baseline group to the video chat or virtual agent conditions did not show a significant intercept term (video chat: b = 0.322, SE = 0.255, t = 1.26, p = .206; virtual agent: b = -0.279, SE = 0.254, t = -1.10, p = 0.272). There was again a significant difference between the in-person and virtual agent

conditions (b = -1.018, SE = 0.360, t = -2.82, p = .005), but no significant difference between the video chat and in-person (b = -0.416, SE = 0.361, t = -1.15, p = .249) or virtual agent (b = 0.601, SE = 0.360, t = 1.67, p = .095) conditions. The effect of linear time terms on the difference between in-person and video chat was significant, suggesting differences in the slope of word recognition responses (b = 1.283, SE = 0.574, t = 2.24, p = .025). No other effects were significant (see Supplementary Information B in OSF repository for full model output). Both of these analyses thus reveal an interesting pattern: Toddlers only show above-chance evidence of learning the novel word-object associations in the in-person group, pointing to a unique advantage of live, in-person interactions for word learning. When comparing conditions directly, the virtual agent group shows significantly lower performance than the in-person group. The children's performance in the video chat condition does not significantly differ from that in the other conditions, which is obvious when observing the widely overlapping distribution of individual performance in Figure 3.

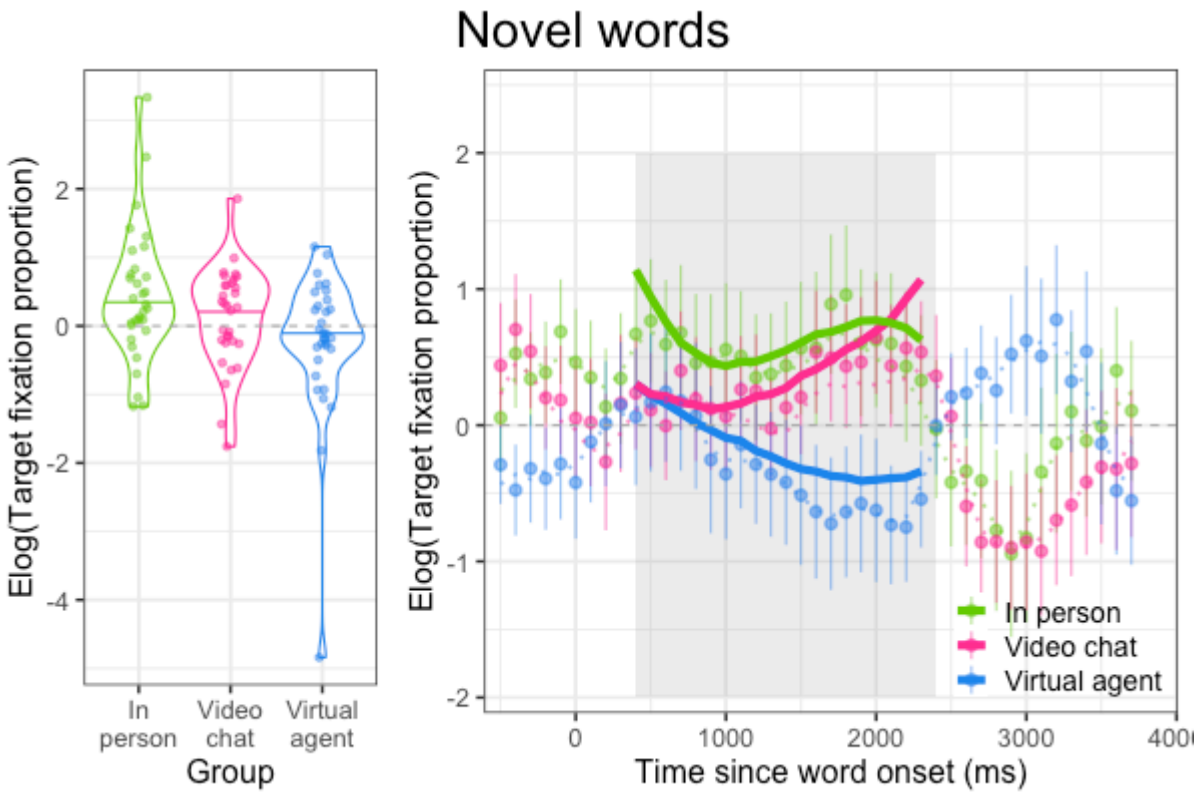

Figure 3. Empirical logit of mean target fixation proportion to correct word-object association in test phase for novel words. Density plots on the left side indicate density estimates over mean target fixation proportion in 400-2400 ms time-window of analysis. Points represent means by participant, solid horizontal line represents median. Dashed line indicates chance level. Plot on the right represents time-course of looks to target around the time-window of analysis (shaded in grey). Gaze proportions were binned into 100 ms units. Dashed line indicates chance level. Circles and error bars represent the observed mean and ±1SE of the mean over each time bin. Solid lines represent model fits derived from GCA model reported in the main text.

4. Discussion

The present study assessed differences in 16-month-old toddlers' learning of novel word-object associations under three different conditions: They were either taught during in-person interaction, by a person through video chat, or by a virtual, interactive on-screen agent. Crucially, in all three experimental groups, the interactions were entirely scripted and as well matched as possible, allowing for a rather direct comparison of these three formats. Our results revealed that toddlers only showed above-chance word learning from live, in-person exposure. The difference between learning from live exposure and virtual agent was significant, such that toddlers learned better from in person exposure than from the virtual agent. However, word learning in the video chat condition did not significantly differ from both in-person and virtual agent conditions, suggesting that learning results are overlapping across these conditions. We integrate these results to previous work in Table 1.

Table 1: Summary of studies on word learning across different conditions. Note that we only include studies that feature at least one contingent screen condition. Age indicates the infant age group; human whether the teacher was a human; agency whether the child got to interact with the teacher directly; contingency whether the teacher was contingent on the child's looking; 3D whether the teacher was present in person; and learning whether significant learning was observed for the novel words. See Introduction for a summary of the papers and full reference.

Study	Condition	Age (months)	Human	Agency	Contingency	3D	Learning
Roseberry 2014	in person	36	yes	yes	yes	yes	yes
	video chat	36	yes	yes	yes	no	yes
	yoked	36	yes	no	no	no	no
Myers 2016	video chat	12-21	yes	yes?	yes?	no	no
	yoked	12-21	yes	no	no	no	no
	video chat	22 & 25	yes	yes?	yes?	no	yes
	yoked	22 & 25	yes	no	no	no	no
Troseth 2018	video chat	24 & 30	yes	no?	yes?	no	yes
	yoked	24 & 30	yes	no	no	no	no
Kirkorian 2016	specific contingent	24	yes	no	yes	no	yes
	general contingent	24	yes	no	partial	no	no
	non-contingent	24	yes	no	no	no	no
Tsuji 2020	contingent	12	no	no	yes	no	yes
	yoked	12	no	no	no	no	no
Present	in person	16	yes	no	yes	yes	yes
	video chat	16	yes	no	yes	no	no
	contingent	16	no	no	yes	no	no

The results for the in-person group are in line with a multitude of studies emphasizing the advantages of live, real world exposure over screen-mediated exposure on language

learning. It is noteworthy that these results ensued in our study even though the experimenter was required to restrict both her interactions and language to a predetermined script that was closely matched to the screen conditions. In addition, in the current experiment the live group arguably had the hardest task in that they had to transfer word-object associations learned in the real world onto the 2-D world of a screen during the test phase. What remains so special about in-person exposure, even under conditions where most aspects of a broader social context are removed? One characteristic that makes in-person exposure special is that toddlers already have accumulated ample experience with live persons that turn out to be relevant social agents and thus potential teachers. Therefore, a period aimed at experiencing the teacher as a social agent preceding the teaching situation is not necessary for in-person teaching, because the toddlers' expectation is already in place. Moreover, the in-person teacher might have been more salient by virtue of appearing in the physical world, including larger in size and having more salient 3-D movement. For all these reasons, all other factors matched, in-person exposure, compared to on-screen exposure, might more easily lead to learning success.

One potential difference between the in-person and on-screen groups was the fact that the experimenter was able to see the toddler. Thus, even in a scripted interaction, it was not possible to control any spontaneous facial or gestural reactions of the experimenter in response to the toddler. Furthermore, while the experimenter was instructed to respond to the toddler in a standardized fashion, ensuring the precise timing and manner of reaction in response to the toddler's behavior was hard to control completely. These problems were mostly addressed by the design of the video chat condition, in which the experimenter did not see the toddler directly, but only had access to gaze location information. Even so, we observe overlapping distributions of learning outcomes for in person as compared to live video chat.

Notice, however, that our video chat group differed notably from previous studies which found successful learning in the video chat (Roseberry et al., 2014; Myers et al., 2017) not only because we did not allow the experimenter to interact with the child outside of the very set script (which was limited to contingent smiling and gaze following), but also in that the teacher herself had access to reduced information during the teaching phase. This should have led to poor performance in the video chat condition, which was more impoverished than that previous work.

Indeed, although performance in the video chat group was statistically not different from in-person interaction, comparison against chance revealed no significant word recognition effect in the video chat group. As discussed further above, and as suggested in the previous literature (Troseth et al., 2006; 2008), due to factors like the relative unfamiliarity of interacting with a social agent via a screen as well as a video deficit effect, toddlers in the video chat group might need additional cues to the interactive and socially relevant nature of the situation in order to make the most of the learning opportunities. Toddlers in the virtual agent group learned significantly worse compared to toddlers in the in-person group, while their performance overlapped with that of the video chat group. One reason why this difference might matter again lies in toddlers' experience: Their expectation of a real person in the 3-D world being a social agent might be higher than their expectation of a person or cartoon character on screen being one. Considering that the virtual agent condition was the only group significantly differing from the in-person group, even more social context might be necessary in order for toddlers to learn from the on-screen character. Another factor that might explain the lack of overlap between the in-person and virtual agent conditions might be the lack of varied, and thus more natural responses that were present when the toddler saw a real experimenter in person or on video chat. Although we circumvented the possibility of spontaneous interaction between toddler and experimenter in the video chat by preventing the

experimenter from seeing the actual toddler, the experimenter's reactions to screen prompts still likely showed a certain degree of temporal and gestural variation. In contrast, the virtual agent reacted always in the same way in response to children's gaze. From early on, infants are sensitive to the natural variation in environmental contingencies and start preferring agents that display imperfect contingencies from those that display perfect ones (Striano, Henning, & Stahl, 2016). That said, we also want to bear in mind a previous study with Japanese 12-month-old infants, which revealed above-chance learning of novel word-object associations from a virtual agent under similar conditions (Tsuji et al., 2020). Due to differences in lab, eye-tracking device, and language and cultural background across that study and ours, it is difficult to pinpoint a reason for these differences.

It is, however, conceivable that age plays an important role. The degree to which the video deficit affects learning has been reported to change with age, with 6-month-old infants often learning equally well from 2-D and 3-D sources, the video deficit effect peaking at around 15 months of age, and lasting until around 36 months of age (see Barr, 2010). Studies focusing on word learning have so far focused on older, 36-month-old children (Roseberry et al., 2014), or only found significantly better learning from video chat compared to passive video for toddlers above, but not below 22 months of age (i.e., not at 12–16 months or 17–21 months; Myers et al., 2017). Therefore, further study is necessary to assess differences in word learning from screen media across the first three years of life.

Our interpretations of the results for the two on-screen groups, video chat and virtual agent, resonate with a larger literature demonstrating that learning from screens can be facilitated by enriching the situation with various kinds of social cues, including but not limited to contingent responsiveness. For instance, toddlers learn words better in a setting where they see their own mother on screen instead of an unknown experimenter (Krcmar, 2010), or when they observe a reciprocal social interaction on the screen before being

exposed to a learning phase (O'Doherty, Troseth, Shimpi, Goldenberg, Akhtar, & Saylor, 2011). The same holds for off-screen teaching situations with minimal cues to social agency. For instance, infants gaze-followed a bear-like artificial agent when it displayed contingent responsiveness or face-like features, but not if he lacked both (Johnson, Slaughter, & Carey, 1998). In the same vein, a contingently-reacting artificial agent was only gaze-followed when an experimenter had previously interacted with him socially, thus by either conversing or hand-clapping, but not if the experimenter was clapping with sticks, a response not expected in a communicative mode (Beier & Carey, 2014). A social context that reduces learning success does not need to consist of a general reduction in social cues, but might also be manifested by unexpected social behavior displayed by the interaction partner. For instance, infants and young children prefer looking at or choosing an object presented by a speaker of their native language compared to an object presented by a speaker of a non-native language (Kinzler, Dupoux, & Spelke, 2012; Marno et al., 2016). Learning from screens can be enhanced by manipulating not only the screen content itself, but also the environment: Co-viewing has been shown to increase learning success from screen content under certain conditions (e.g., Sims & Colunga, 2013; Strouse & Troseth, 2014), which can be attributed to including the screen content into the broader social context of toddlers' environments.

In our view, rather than thinking in terms of a dichotomy between screens and in-person interaction, the most productive way to move forward involves locating the results of the present study as well as recommendations on the conditions under which toddlers can learn from digital media in a broader context of cues to social agency. If we take a natural social interaction between a caregiver and her child as the standard situation that informs a child of social agency, any teaching event that deviates from this pattern might reduce the efficacy of the teaching act. In this context, we would conceptualize learning from a screen as a deviation similar in quality as, for instance, learning from an unknown person or learning

from a person displaying unexpected behavior. To clarify, this is not to say that the magnitude of the deviation need be the same. Instead, we propose that both kinds of teaching events can be viewed as a deviation from the standard teaching event, and that the consequences on learning based on these deviations can be explained by the same mechanisms, which we propose are based in the toddler's priors and the ensuing predictability of a situation as involving a social agent. Some of these priors are present early on in infant development, while others might develop based on experience with social teaching situations. For instance, infants are drawn to many aspects of a real interaction partner early on, such as face-like configurations, voices, or temporal contingencies. These early preferences might help draw infants' attention to social situations in the absence of specific experience. Infants' experience over the first few months of life then leads to perceptual attunement and more specific preferences; thus, human faces or the imperfect contingencies typical of natural social interactions. The standard social situation depicted above is likely the one a given toddler has encountered the most frequently in her life experience with social agents, and thus is the one she will consider the most likely cue to social agency and a subsequent teaching event.

These considerations lead to a view where educational digital devices need to contain a certain amount of resemblance to natural social teaching situations in order for toddlers to efficiently learn from them, and that this amount might change over development. Indeed, a recent meta-analysis has reported that the quality of screen media content shows positive associations with language skills (Madigan, McArthur, Anhorn, Eirich, & Christakis, 2020). As illustrated by some of the research cited above, this resemblance can be achieved multiple ways, for instance by known persons, human-like appearance, contingent responsiveness, or co-viewing. The present study's contribution to this debate consists in assessing 16-month-old toddlers' word learning from screen media in the absence of a prior or concomitant rich

interaction context. We conclude that contingent responsiveness and even human-ness alone is not sufficient in this setup, since toddlers only learned when taught by a person in the same room. Thus, toddlers this age may require more cues to social agency in order to learn from an interactive digital teacher. The comparison to results from Tsuji et al. (2020) further suggests that these minimal criteria show developmental and/or cultural changes. We propose that future studies assess the conditions under which screen media can be efficient learning tools systematically from the perspectives of salience and predictability of the teaching situation.

Acknowledgements

We thank Michel Dutat and Vireack Ul for their assistance in setting up the study, and Luce Legros for assistance with recruiting and testing infants.

Funding: This research was supported by grants from the Fondation Fyssen, and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 659553, the Agence Nationale pour la Recherche [ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017], and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award.

References

Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist, 48*(5), 505-522. doi: https://doi.org/10.1177/0002764204271506

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457-474. doi: 10.1016/j.jml.2007.09.002

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review, 30*(2), 128-154. doi: 10.1016/j.dr.2010.03.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01.

Beier, J. S., & Carey, S. (2014). Contingency is not enough: Social context guides third-party attributions of intentional agency. *Developmental Psychology, 50*(3), 889–902. doi: 10.1037/a0034171

Chassiakos, Y. L. R., Radesky, J., Christakis, D., Moreno, M. A., & Cross, C. (2016). Children and adolescents and digital media. *Pediatrics, 138*(5), e20162593. doi: 10.1542/peds.2016-2593

Deligianni, F., Senju, A., Gergely, G., & Csibra, G. (2011). Automated gaze-contingent objects elicit orientation following in 8-month-old infants. *Developmental Psychology, 47*(6), 1499–503. doi: 10.1037/a0025659

DeLoache, J. S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., … O'Doherty, K. (2010). Do babies learn from baby media? *Psychological Science, 21*(11), 1570–4. doi: 10.1177/0956797610384145

Dink, J., & Ferguson, B. (2018). *eyetrackingR.* R package version 0.1.8. http://www.eyetracking-R.com.

Dunst, C. J., Gorman, E., & Hamby, D. W. (2010). Effects of Adult Verbal and Vocal Contingent Responsiveness on Increases in Infant Vocalizations. *Center for Early Literacy Learning, 3*(1), 1-11. Retrieved from: http://www.earlyliteracylearning.org/cellreviews/cellreviews_v3_n1.pdf

Elsabbagh, M., Hohenberger, A., Campos, R., Van Herwegen, J., Serres, J., de Schonen, S., … Karmiloff-Smith, A. (2013). Narrowing Perceptual Sensitivity to the Native Language in Infancy: Exogenous Influences on Developmental Timing. *Behavioral Sciences, 3*(1), 120–132. doi: 10.3390/bs3010120

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental psycholinguistics: On-line methods in children's language processing, 44*, 184-218.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America, 100*(13), 8030–5. doi: 10.1073/pnas.1332441100

Goldstein, M. H., & Schwade, J. A. (2008). Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning. *Psychological Science, 19*(5), 515–523. doi: 10.1111/j.1467-9280.2008.02117.x

Hakuno, Y., Omori, T., Yamamoto, J. I., & Minagawa, Y. (2017). Social interaction facilitates word learning in preverbal infants: Word–object mapping and word segmentation. I*nfant Behavior and Development, 48*, 65-77. doi: 10.1016/j.infbeh.2017.05.012

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science, 1*(2), 233-238. doi: 10.1111/1467-7687.00036

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2012). 'Native'objects and collaborators: Infants' object choices and acts of giving reflect favor for native over foreign speakers. *Journal of Cognition and Development, 13*(1), 67-81. doi: 10.1080/15248372.2011.567200

Kirkorian, H. L., Choi, K., & Pempek, T. A. (2016). Toddlers' Word Learning From Contingent and Noncontingent Video on Touch Screens. *Child Development, 87*(2), 405–413. doi: 10.1111/cdev.12508

Krcmar, M. (2010). Can social meaningfulness and repeat exposure help infants and toddlers overcome the video deficit? *Media Psychology, 13*(1), 31-53. doi: 10.1080/15213260903562917

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology, 10*(1), 41-63. doi: 10.1080/15213260701300931

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America, 100*(15), 9096–101. doi: 10.1073/pnas.1532872100

Madigan, S., McArthur, B. A., Anhorn, C., Eirich, R., & Christakis, D. A. (2020). Associations Between Screen Use and Child Language Skills: A Systematic Review and Meta-analysis. *JAMA pediatrics 174*(7), 665-675. doi: 10.1001/jamapediatrics.2020.0327

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language, 57*(2), 252-272. doi: 10.1016/j.jml.2007.03.005

Marno, H., Guellai, B., Vidal, Y., Franzoi, J., Nespor, M., & Mehler, J. (2016). Infants' selectively pay attention to the information they receive from a native speaker of their language. *Frontiers in Psychology, 7*, 1150. doi: 10.3389/fpsyg.2016.01150

McClure, E. R., Chentsova-Dutton, Y. E., Barr, R. F., Holochwost, S. J., & Parrott, W. G. (2015). "Facetime doesn't count": video chat as an exception to media restrictions for infants and toddlers. *International Journal of Child-Computer Interaction, 6*, 1-6. doi: 10.1016/j.ijcci.2016.02.002

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: Chapman and Hall/CRC Press.

Miller, J. L., & Lossia, A. K. (2013). Prelinguistic infants' communicative system: Role of caregiver social feedback. *First Language, 33*(5), 524–544. doi: 10.1177/0142723713503147

Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2016). Baby FaceTime: Can toddlers learn from online video chat? *Developmental Science, 20*(4), e12430. doi: 10.1111/desc.12430

O'Doherty, K., Troseth, G. L., Shimpi, P. M., Goldenberg, E., Akhtar, N., & Saylor, M. M. (2011). Third-party social interaction and word learning from video. *Child Development, 82*(3), 902-915. doi: 10.1111/j.1467-8624.2011.01579.x

Psychology Software Tools, Inc. *[E-Prime 2.0].* (2016). Retrieved from https://www.pstnet.com.

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna. Retrieved from https://www.r-project.org/

Rideout, V., Saphir, M., Tsang, V., & Bozdech, B. (2013). Zero to eight: Children's media use in America. *Common Sense Media.* Retrieved from http://www.commonsensemedia.org/research/zero-to-eight-childrens-media-use-in-america-2013.

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child Development, 85*(3), 956–70. doi: 10.1111/cdev.12166

Shuler, C. (2012). *iLearn II: An Analysis of the Education Category of the iTunes App Store*. New York, NY: The Joan Ganz Cooney Center at Sesame Workshop. Retrieved from

SR Research Ltd. (2005). *EyeLink 1000 User Manual (version 1.5.0)*. Retrieved from http://sr-research.jp/support/EyeLink%201000%20User%20Manual%201.5.0.pdf

Sims, C., & Colunga, E. (2013). Parent-Child Screen Media Co-Viewing: Influences on Toddlers' Word Learning and Retention. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35,* 1324-1329. doi: https://escholarship.org/uc/item/1m4019fp

Striano, T., Henning, A., & Stahl, D. (2005). Sensitivity to social contingencies between 1 and 3 months of age. *Developmental Science, 8*(6), 509-518. doi: 10.1111/j.1467-7687.2005.00442.x

Striano, T., & Reid, V. M. (2006). Social cognition in the first year. *Trends in Cognitive Sciences, 10*(10), 471-476. doi: 10.1016/j.tics.2006.08.006

Strouse, G. A., & Troseth, G. L. (2014). Supporting toddlers' transfer of word learning from video. *Cognitive Development, 30*, 47-64. doi: 10.1016/j.cogdev.2014.01.002

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal Responsiveness and Children's Achievement of Language Milestones. *Child Development, 72*(3), 748–767. doi: 10.1111/1467-8624.00313

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young Children's Use of Video as a Source of Socially Relevant Information. *Child Development, 77*(3), 786–799. doi: 10.1111/j.1467-8624.2006.00903.x

Troseth, G. L., Strouse, G. A., Verdine, B. N., & Saylor, M. M. (2018). Let's chat: On-screen social responsiveness is not sufficient to support toddlers' word learning from video. *Frontiers in Psychology, 9*, 2195. doi: 10.3389/fpsyg.2018.02195

Tsuji, S., Jincho, N., Mazuka, R., & Cristia, A. (2020). Communicative cues in the absence of a human interaction partner enhance 12-month-old infants' word learning. *Journal of Experimental Child Psychology, 191*, 104740. doi: 10.1016/j.jecp.2019.104740

Wang, Q., Bolhuis, J., Rothkopf, C. A., Kolling, T., Knopf, M., & Triesch, J. (2012). Infants in Control: Rapid Anticipation of Action Outcomes in a Gaze-Contingent Paradigm. *PLoS ONE, 7*(2), e30884. doi: 10.1371/journal.pone.0030884

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York.