Abstract

Is infants' word learning boosted by non-human social agents? An on-screen virtual agent taught infants word-object associations, in a set-up where the presence of contingent and referential cues could be manipulated using gaze contingency. Twelve-month-old Japanese-learning children (n = 36) looked significantly more to the correct object when it was labeled after exposure to a contingent and referential display than a non-contingent and non-referential display. These results show that communicative cues can augment learning even for a non-human agent, a finding highly relevant for our understanding of the mechanisms through which the social environment supports language acquisition, and for research on the use of interactive screen media.

Keywords: Social cues, temporal contingency, word learning, first language acquisition, gaze-contingent eye-tracking, virtual agent

Communicative cues in absence of a human interaction partner enhance 12-month-old infants' word learning

Theories of early language acquisition have long emphasized the role of social interaction for learning (e.g., Snow, 1972; Tomasello, 2003; Vygotsky, 1962), and a wide range of studies report a positive relation between toddlers' language outcomes and the number of social cues provided by their caregivers (Altvater-Mackensen & Grossmann, 2015; Gros-Louis, West, & King, 2014; Hirsh-Pasek, Adamson, et al., 2015; Tamis-LeMonda, Bornstein, & Baumwell, 2001). While these and other findings document a consistent and general advantage of environments rich in social interactions, the essential components and mechanisms by which these environments lead to a learning advantage remain elusive. One fundamental question towards answering this unresolved issue is to ask what constitutes a social agent for a toddler, and how this affects their learning.

One insightful approach has been the investigation of the extent to which toddlers can learn from screen media, a setting that is stripped of many aspects of a social interaction. Studies comparing learning from a live teacher to learning from matched videos of this teacher have often shown poorer learning for toddlers under 2.5 years of age from the latter, an observation commonly described as the *video deficit* effects (Anderson & Pempek, 2005). Typically, this line of research has compared infants' learning after live exposure to a closely matched video exposure, and found better performance after the former for various aspects of language learning such as phonetic acquisition (e.g., Kuhl, Tsao, & Liu, 2003) and word learning (e.g., Krcmar, Grela, & Lin, 2007), but also in other domains such as imitation (e.g., Barr & Hayne, 1999) or object retrieval (e.g., Troseth, Saylor, & Archer, 2006).

The same line of research has, however, also demonstrated that, even without the physical presence of the interaction partner, enriching a video condition with social cues can enhance learning. For instance, toddlers learn words better if their mother appears on screen

rather than an unknown experimenter (Krcmar, 2010), or when they observed a reciprocal social interaction on the screen before learning (O'Doherty, Troseth, Shimpi, Goldenberg, Akhtar, & Saylor, 2011). Furthermore, replacing a prerecorded video with a live video setting, thus adding real time contingency, also enhances learning. Toddlers learned novel verbs equally well from live video as from live interaction, but not from yoked video (Roseberry, Hirsh-Pasek, & Golinkoff, 2014). Similarly, a study exposing toddlers to live video or yoked video for a week found better word learning in the live video group for toddlers 22-25 months of age (but not younger; Myers, LeWitt, Gallo, & Maselli, 2016).

These results demonstrate that enriching an agent encountered in an on-screen situation with socially meaningful cues can enhance learning success. Does this mean that the cues themselves, even without the "looks" of a human interaction partner, play a crucial role in supporting learning? Infants' sensitivity to social-communicative cues in the absence of a human-like interaction partner has been documented in studies on the role of contingent responsiveness in eliciting social-like reactions. Twelve-month-old toddlers have been found to follow the gaze of a bear-like object as long as it either had facial features, had previously beeped and blinked contingently upon infants' gaze, or both, but not if it lacked both a face and contingent responsiveness (Johnson, Slaughter, & Carey, 1998). A similar behavior was reported in eight-month-olds, who followed the turning direction of an amorphous object on screen only if that object had previously moved contingently on infants' gaze (Deligianni, Senju, Gergely, & Csibra, 2011). Thus, infants react socially (in particular, they follow gaze) when confronted with a non-human counterpart, as long as it exhibits some cues that, in humans, would signal communicative abilities. It is unclear whether such cues, once disentangled from the actual human interaction partner, are still interpreted as actual social-communicative cues. Consider, for instance, contingent responsiveness: Although it can rather unanimously be regarded as a social cue in the context of a caregiver-child interaction,

what about a contingently reacting, non-social screen display, such as that encountered on a smartphone or tablet screen? Infants are sensitive to such on-screen contingencies, for instance triggered by a gaze-contingent display, early on (Wang, Bolhuis, Rothkopf, Kolling, & Knopf, & Triesch, 2012). Whether these cues elicit gaze following because they are regarded as social-communicative cues or not is thus an open question, and one that we will not aim to answer in the present study.

Instead, the present study assesses whether, given that infants gaze-follow non-human agents enriched with such cues, would they also more readily learn word-object associations from such agents as compared to less social agents? This is a central question in the quest for understanding the mechanisms through which social situations in general lead to better learning. The present study aims to address it by using an innovative experimental design. Using gaze-contingent eye-tracking, we implement a scenario in which a non-human virtual agent (teacher) teaches novel word-object associations to an infant. The scenario is either enriched with two types of cues, contingency and reference, or lacks those. We will continue referring to them as social-communicative cues in the context of the present study, all the while acknowledging the above-mentioned caveat that this characterization might not adequately describe their nature once they are isolated from human interaction partners.

In the enriched condition, both the agent and the object-to-be-taught react contingently upon the toddler's gaze, and the agent also exhibits referential cues, while these cues are absent in the control condition. The present design enables us to assess to what extent the addition of two communicative cues, contingency and reference, can enhance the learning of novel word-object associations from a virtual, non-human agent.

While previous studies on learning words or word-object associations from screens with added social cues have mostly assessed toddlers older than 12 months of age, studies on gaze

following from contingent social cues have worked with infants younger than 12 months. The present study studied infants at 12 months of age, an age when word recognition starts to rapidly increase (Bergelson & Swingley, 2012), and well beyond the age where they can learn words from screens in the lab given salient animation of the target object, either in form of temporal synchronization of sound and object (Gogate, & Bahrick, 1998) or of infant-initiated target object movement and naming (Shukla, White, & Aslin, 2011). Given that infants are known to succeed at tasks indicative of full word learning, such as retention (Horst & Samuelson, 2008) or disambiguation (e.g., Bion, Borovsky, & Fernald, 2013) only at a later age, our design included what can be conceived of as the most basic step of word learning, namely the cross-modal mapping of a wordform to a visual target. We will come back to this issue in the Discussion.

In sum, the current study assesses whether a screen display that is contingent and referential but not human can support the early acquisition of word-object associations.

Methods

Participants

Thirty-six normally developing, monolingual Japanese-learning infants from the Tokyo region were included (18 female, mean age = 362 days, range 351-379 days). Due to the novelty of our paradigm, there was no straightforward way to select previous studies for power calculation and thus we used a rule of thumb. We acknowledge that this is less than ideal, and now conducted power calculations based on related studies (Deligianni et al., 2011; Roseberry et al., 2014; Woodward, Markman, & Fitzsimmons, 1994) concluding that our sample size decision was reasonable (see Appendix A). Eleven additional infants were tested but excluded from analysis due to crying (3), fussiness (5), or calibration and equipment

problems (3). This exclusion rate of less than 25% is not at all uncommon for infant experiments (see, for instance, a meta-analysis by Bergmann et al., 2018). Nevertheless, it would be insightful to assess whether the excluded participants were representative of the sample as a whole, which we were able to do to only a limited extent. Three of the excluded participants were excluded based on equipment error, which is independent of the participants themselves. As to the remainder, there were no systematic differences in age of excluded (mean age = 360 days, range 352-378 days) compared to included (see above) participants. The study was approved by the (name masked for anonymous review) Ethics Committees. Infants were recruited from the laboratory participant pool. Caregivers signed an informed consent prior to their inclusion in the study and received a book voucher after participation.

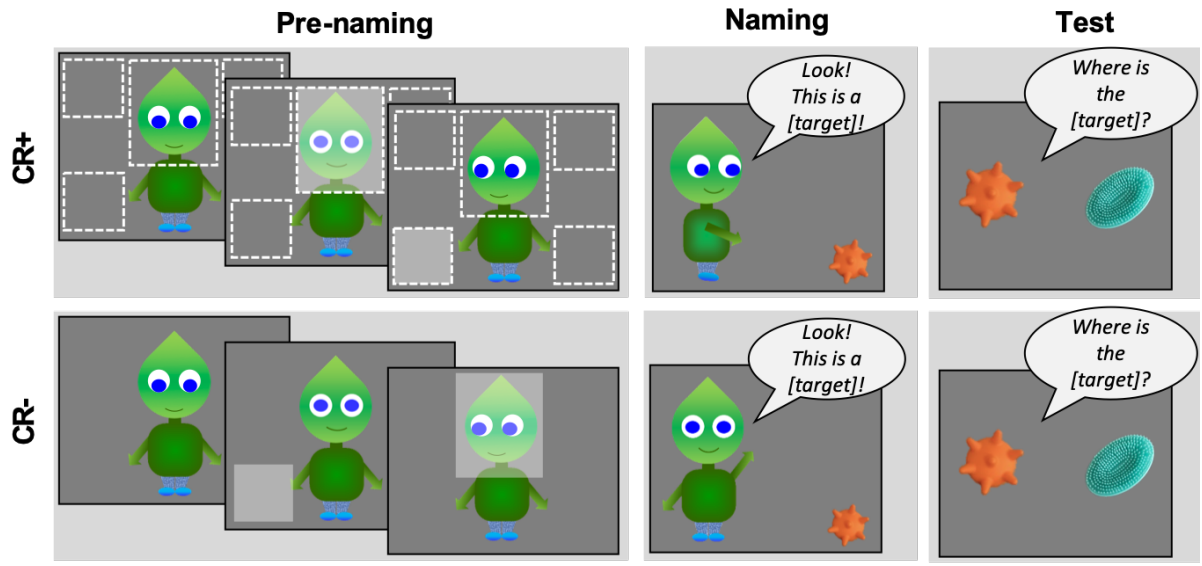

Figure 1. Design by condition and experimental phase. CR+: Contingent and referential condition. CR-: Non-contingent and non-referential condition, In the CR+ condition, dashed rectangles depict areas of interests for which the screen was gaze-contingent. Shading exemplifies infant gaze in a particular trial.

Research Design

The experiment consisted of two within-participant blocks, one of which contained all trials of the contingent and referential (CR+) condition. The other block contained all trials of the non-contingent non-referential (CR-) condition. Block order was counterbalanced across infants. In each block, infants were taught three novel word-object associations.

The experiment started with a display showing the virtual agent waving, while central speakers played a greeting phrase ("Konnichiwa! Kyō-wa issho ni asobō ne". *Hello! Would you like to play with me?*) Then, within each block, infants were exposed to two familiarization trials, nine exposure trials (three trials for each of the three novel word-object associations), and six test trials (two trials for each novel word-object association), described next.

The familiarization trials were intended to familiarize infants with the contingent reactivity of agent and objects in the CR+ condition, or the absence thereof in the CR- condition. In each of the two trials, first the picture of a novel object was presented on the right or left side of the screen. This object was distinct from the objects named in the next phase and would only appear during familiarization. In the CR+ condition, the object would slowly inflate and deflate if the infant looked at it and stop moving if the infant looked away. In the CR- condition, the object would also inflate and deflate, but this movement was preprogrammed and not contingent on infants' gaze. We used gaze data collected in a pilot study with infants of the same age exposed to the same CR+ condition as a basis to match length and amount of movement during all phases of the control condition (for details, see Appendix B). The display of the object was followed by a display of the virtual agent, who looked up and smiled each time the infant looked at it and looked down if not in the CR+ condition. In the CR- condition, these movements were again preprogrammed (Figure 1).

The exposure trials displayed the virtual agent in the center of the screen. In each trial, one of the three novel objects exposed in this block was positioned to either the lower right or lower left side of the agent. Trials were subdivided into a *pre-naming* phase, in which the infants had time to visually explore the display and again experience the contingency or absence thereof, and a *naming* phase, in which the teacher named the novel objects. The *pre-naming* phase of each exposure trial started with the virtual agent eyes facing down and body and arms in neutral position. In the CR+ condition, the position of the eyes and mouth of the virtual agent changed if the infant looked at her face so she would appear to look up and smile. If the infant looked at the object, it would slowly inflate and deflate while the agent would direct her gaze toward it. The display moved on to the naming phase when the phase had lasted 6000 ms or the infant had looked back at the virtual agent for 10 times or more (the latter only happened in 15 of 324 trials across infants, with an average of 3 times (*SD* = 1.46)). In the CR- condition, the virtual agent and object would show the same types and amount of movement as in the CR+ condition, but not contingent on infants' gaze.

The *naming* phase started with the teacher looking up and smiling for 500 ms before a carrier sentence naming the respective object started playing. In the CR+ condition, the teacher showed referential cues during naming in a display where she was looking at, pointing at, and turning her torso toward the object. The object continued to inflate and deflate if the infant looked at it. In the CR- condition, the teacher showed movements that were again matched in timing and quantity to the CR+ condition, but were not contingent. This time, however, the quality of movement was different: Instead of referring to the object, the teacher looked at the infant and raised one of her arms. The object inflated and deflated in a non-contingent, but matched fashion.

The exposure trials were followed by the test trials for that condition. Each test trial displayed a combination of two of the previously learned objects side by side on the screen

(Figure 1). Thus, infants had been exposed equally to each of these objects previously. After two seconds, infants heard a sentence in which one of the objects was named (looking-while-listening procedure, Fernald et al., 2008).

Preceding each trial, an attention getter (the picture of a flower) appeared centrally on screen, and the trial was initiated by the experimenter once the infant's gaze was fixated on it.

Stimuli

The six novel word-object associations shown during the experiment were organised into paired triplets, which were taught in the CR+ and CR- conditions, respectively. Which triplet was taught in which condition was counterbalanced across infants.

The pictures of novel objects were photos of six real, unfamiliar inanimate objects (Appendix C). An additional novel object in two color variants was chosen for the two silent familiarization trials. The names for the target objects [de:ʑo, kippo, ku:be, monɕa, sappu, ɕingʲo] were phonotactically legal Japanese non-words. They were constructed as disyllables with heavy-light syllable weight, a structure frequently occurring in Japanese infant-directed words (Mazuka, Igarashi, & Nishikawa, 2006). Frequency of occurrence of the constituting syllables and of the whole string were matched. The non-words were embedded in carrier sentences (see Appendix D). In each trial of the exposure phase, the target non-word appeared three times (e.g. "Kore wa ku:be da yo. Ku:be. Ku:be". *This is a ku:be. A ku:be. A ku:be*.). In each trial of the test phase, the object was named once (e.g.,"A, ku:be da. Wakaru ka na?" *Oh, there's the [Target]! Can you find it?*). Non-words, their carrier sentences, as well as a short greeting phrase were recorded by a female native speaker of Japanese in infant-directed register.

The virtual teacher present on screen during the exposure phase was designed to have human-like facial and body features including eyes, a mouth, a torso, and extremities that would allow her to exhibit referential and contingency cues (see Figure 1).

Procedure

Infants were sitting on a caregiver's lap in a sound-attenuated room facing the screen of a Tobii XL eye-tracker. The experimenter was hidden behind a wall in the same room. Both caregiver and experimenter wore headphones with masking music. Infants' gaze was calibrated with Tobii Studio's infant-friendly 5-point calibration. Their gaze was recorded with a sampling rate of 120 Hz, and the experiment was administered using E-Prime 2.0.

Data Cleaning and Preparation

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 the windows chosen in previous studies using comparable designs (e.g., Mani & Plunkett, 2007), and accounts for the fact that infants need several hundred milliseconds to initiate a gaze shift (Fernald et al., 2008). Data points with low validity were excluded from analysis as recommended by the manufacturer (Tobii Technology, Inc, 2016). All analyses were conducted in R version 3.5.3 (R Core Team, 2019) with the packages eyetrackingR version 0.1.7 (Dink & Ferguson, 2018) and lme4 version 1.1-12 (Bates, Maechler, Bolker, & Walker, 2015). Figures were made with ggplot2 version 3.1.0 (Wickham, 2016). We included test trials in which infants fixated on both pictures at one point during the trial (see, e.g., Swingley & Aslin, 2007; excluding 14.7% of trials), and where they looked to the screen for more than 25% of the time window of analysis (excluding 13.4% of remaining trials). Subsequently, we excluded from analysis a given condition for a particular infant if it was represented by less than 2 (of 6) trials. This excluded one infant entirely (that only had 1 trial

left for either condition; already reflected in Participant section), and one condition each for two infants. The main analysis is based on an average of 4.3 trials (*SD* = 1.2) in the CR+, and an average of 4.5 trials (*SD*= 1.3) left in the CR- condition.

Since other researchers might have opted for different exclusion criteria, in particular not choosing to exclude conditions with only one trial, or else choosing to exclude all data of an infant that did not have a sufficient number of trials in either condition, we ran two alternative versions of the analysis. In the first version, we included all trials that passed our trial exclusion criterion. In the second version, we excluded all infants that did not have at least two trials in both conditions. Both versions led to qualitatively the same results (see Appendix E). Note that eight additional infants only had trials of one condition left for analysis, indicating that the experiment might have been too long for sustaining their attention.

Data Analysis

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, and empirical logit transformation was used 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 model took the form [Elog ~ Condition * (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. Differences between the two conditions were assessed using model comparison with the

likelihood ratio test. In order to provide a more commonly interpretable effect size measure in addition to the likelihood ratio test, we adopt the approach proposed by Westfall, Kenny, and Judd (2014) to calculate a Cohen's d-type effect size for fixed effects in linear mixed effects models. We calculate this effect size based on a model on the average difference between conditions, thus without time terms. Since it is not only of interest whether conditions differ, but also whether each individual condition leads to above-chance word recognition, we also inspected the model intercept, with each condition serving as the comparison level in two separate analyses. Statistical significance (*p*-values) for the intercept was assessed using the normal approximation (i.e., treating the *t*-value as a *z* -value).

In order to explore potential differences in looking patterns during the exposure phase, we also examined the total looking times spent on the virtual agent and the object during the pre-naming and naming phase of the experiment.

Results

Model comparison revealed a significant difference between looking behavior in the test trials of the two conditions [$\chi^2(1) = 24.68$, $p < .001$], with the slope for condition indicating higher target fixation proportion in the CR+ than the CR- condition ($b = 0.361$, $SE = 0.073$). The effect size calculated over the average difference between conditions was d = 0.23, a rather small effect size. There was no effect of condition on the linear [$\chi^2(1) = 2.15$, $p = .142$], quadratic [$\chi^2(1) = 1.22$, $p = .269$] and cubic [$\chi^2(1) = 0.068$, $p = .794$] time terms (Figure 2).

The intercept for the model where the CR+ condition was the baseline level was significant ($b = 0.271$, $SE = 0.134$, $t = 2.03$, $p = .042$), while this was not the case when the CR- condition served as baseline level ($b = -0.090$, $SE = 0.136$, $t = -0.66$, $p = .507$). These

results suggest that infants did learn the novel word-object associations in the CR+, but not the CR- condition.

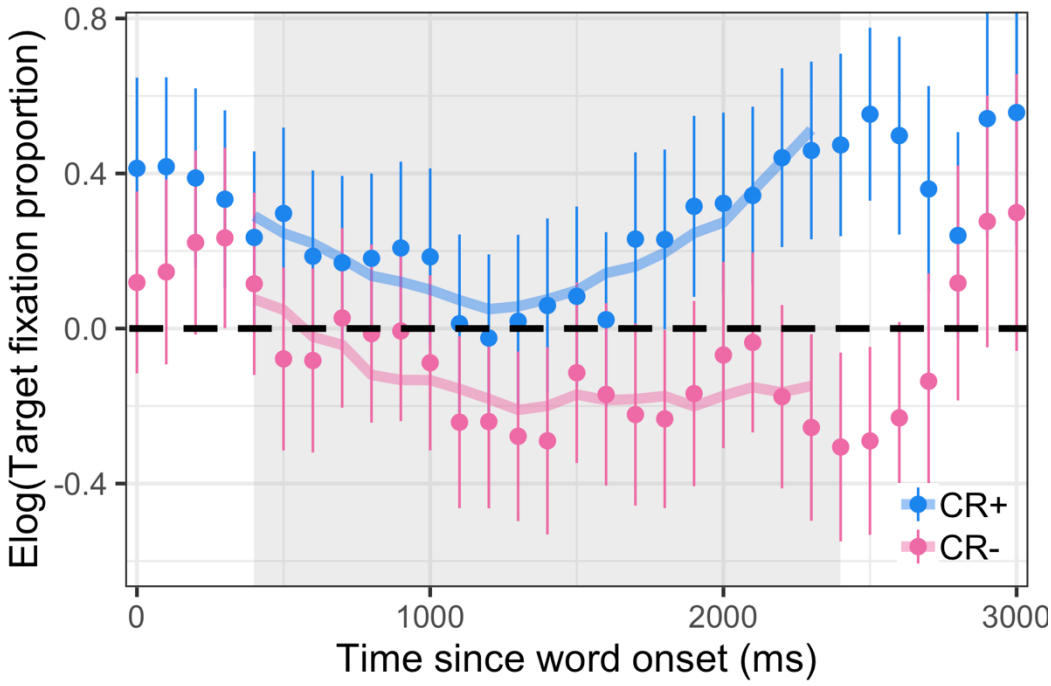

Figure 2. Time-course of infant gaze to target after target word onset in test phase. Looking times were binned into 100 ms units and underwent empirical logit transformation. Dashed line indicates chance level. Circles and error bars represent the observed mean and ±1SE of the mean over each time bin. Solid lines represent model fits derived from statistical model reported in the main text. Grey shaded area indicates analysis time-window.

Given these results, did infants also show differential looking behavior in the prenaming and naming phase of the experiment? Descriptive statistics suggest no difference by condition, with infants' percentage of time in the pre-naming phase fixating on the virtual agent being 65.7% (SD = 31.1) and on the object being 20.8% (SD = 24.9) in the CR+ condition, and 66.7% (SD = 29.8) on the virtual agent, and 18.6 % (SD = 20.6) on the object in the CR- condition. Similarly, there were no differences in percentage of time spent looking

onto different regions of interest on the screen in the naming phase between the contingent condition (virtual agent: mean = 54.7%, SD = 27.4; object: mean = 26.9%, SD = 24.8) and the control condition (virtual agent: mean = 58.4%, SD = 28.2; object: mean = 22.2%, SD = 20.5).

Discussion

The current study assessed whether 12-month-old infants' word learning from a virtual agent in a fast-mapping paradigm can be enhanced by the addition of two communicative cues, contingency and reference. Our results suggest that on-screen exposure enriched with these cues leads to better learning of novel word-object associations compared to a control condition without these cues. These findings demonstrate that enriching a non-human teacher with communicative cues does not only elicit gaze following (e.g., Deligianni et al., 2011), but can in addition support learning in infants.

What are the mechanisms through which these cues enhance word learning? The literature proposes several possibilities. Let us first consider the task, namely the cross-modal association between a novel object and its label.

This has been suggested to be one of the most basic forms of word learning, recruiting domain-general abilities to make associations between an object and its label (e.g., Yu & Smith, 2007). Under this view, communicative cues are not necessary to solve the task at hand. Instead, they might serve to increase general attention and arousal, thereby increasing the amount or quality of information processing (Kuhl, 2007, see also Posner & Rothbart, 2007), or by heightening an object's perceived salience within its environment (Hollich et al., 2000). We acknowledge that our results do not allow inferences about the necessary environments for more sophisticated forms of word learning like generalization or the acquisition of abstract concepts. For instance, being able to map a newly learned label to an

object is not equivalent to the ability to disambiguate or retain this word (e.g., Bion, et al., 2013; Horst et al., 2008). We view our findings as leading up to interesting queries on the minimal social requirements for these more complex learning tasks.

Another proposal is that a combination of contingency and reference can signal infants something meaningful about the situation or interaction partner. Within the Natural Pedagogy theory, where human communication is seen as adapted for knowledge transmission, contingency has been proposed to serve as an ostensive cue signaling to the infant that an act of knowledge transmission is following (Csibra, Gergeley, & Pisupati, 2009). Referring to and naming an object in the environment is an example of such an act. For instance, infants have been found to only follow referential head turns if preceded by ostensive cues (infant-directed speech or contingent responsiveness, Deligianni et al., 2011; Senju & Csibra, 2008). Similarly, it has been proposed that infants attribute perceptual and attentional abilities, communicative intention, and goal-directed behavior to agents (human or non-human) when they show contingent behavior (Johnson, 2003).

The looking patterns during the exposure phase do not differ between conditions, suggesting that the presence of communicative cues did not affect infants' attention to the teacher or the object. This might indicate that prolonged attention is not the mechanism through which they learned better in the socially enriched condition (see Wu & Kirkham, 2010, for a similar result for gaze following and subsequent learning). However, it is still possible that infants were more focused, or more aroused (Kuhl, 2007), during the socially enriched condition. Even if this was the case, communicative cues may still have enhanced learning by making the teacher and the teaching act meaningful. Crucially, both accounts are compatible with a notion that contingency and referential cues can boost learning even in the absence of a rich social context including a human teacher. But what would be the utility of a mechanism that enhances learning in this way, even in the absence of a human partner? One

proposal is that communicative cues, which in nature are most likely exhibited by relevant interaction partners, could serve as an early training signal for infants, aiding them to recognize which other elements of the social environment are communicatively directed to them (Deligianni et al., 2011). Indeed, infants start having expectations regarding social contingency by two months of age (Nadel, Carchon, Kervella, Marcelli, & Reserbat-Plantey, 1999) and are able to follow gaze by 3 months (Hood, Willen, & Driver, 1998), supporting a notion where such cues serve as a very early mechanism of social learning.

Our results lead the way towards disentangling the many factors that lead to a learning advantage of situations rich in social-communicative cues. Previous studies contrasting live demonstrations with pre-recorded video implemented the live contingency both semantically and socially, resulting in an interaction that was meaningful, relevant, and appropriate in content. For instance, infants were called by their name, asked questions about the toys they played with, and given affectionate feedback on their actions (Roseberry et al., 2014). Similarly, Troseth and colleagues (2006) found that teaching via closed-circuit video only augmented learning if preceded by meaningful interaction. Our results show that even a more reduced version of contingent responsiveness, at least when combined with referential cues, can already lead to increased learning success. This is not to say, however, that referentiality and temporal contingency completely isolated from other social cues would still augment learning. In fact, the literature strongly suggests that this would not be the case. For instance, contingency only leads to gaze following when combined with additional social cues, such as a human communication mode (Beier & Carey, 2014). Future work could assess the minimal conditions under which isolated social-communicative cues would augment learning, for instance by removing facial features from the virtual agent.

Our study also contributes to an important yet under-researched topic: How children learn from screens. In a time where 46% of infants under 2 years of age are reported to have

used mobile devices (Rideout, 2017), a better understanding of the impact of interactive screen media on infant learning is indispensable, but research is scarce as of now. A recent study used a touch-screen video game where a woman labeled novel objects hidden in various boxes (Kirkorian, Choi, & Pempek, 2016). Twenty-four-month-old infants in a contingent condition with specific instructions to touch a box on the screen in order to see the object showed a word-learning advantage, while infants in a more generally contingent condition ("touch the screen") or a non-contingent condition ("watch the screen", with the game advancing regardless of touching) did not. This finding demonstrates that the manipulation of temporal contingency - the display proceeding to the next screen upon touch - can lead to a word-learning advantage. Extending on this study, the present work confirms that even younger children can learn better from an interactive screen with a gaze-contingent paradigm. We do not regard the possibility for such practical applications as an opportunity to uncritically support a rise of screen media use in infants. The side effects of early interactive media use (e.g., Cheung, Bedford, Saez De Urabain, Karmiloff-Smith, & Smith, 2017) have to be taken seriously. However, given the overwhelming number of applications targeted at infants and labeled as educational, but largely untested (Hirsh-Pasek et al., 2015), we are convinced that research into sensible use of such opportunities is indispensable.

References

Altvater-Mackensen, N., & Grossmann, T. (2015). Learning to match auditory and visual speech cues: Social influences on acquisition of phonological categories. *Child Development*, *86*(2), 362-378. https://doi.org/10.1111/cdev.12320

Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist*, *48*(5), 505-522. 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. https://doi.org/10.1016/j.jml.2007.09.002

Barr, R., & Hayne, H. (1999). Developmental Changes in Imitation from Television during Infancy. *Child Development*, *70*(5), 1067–1081. https://doi.org/10.1111/1467-8624.00079

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/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. https://doi.org/10.1037/a0034171

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39-53. https://doi.org/10.1016/j.cognition.2012.08.008

Cheung, C. H. M., Bedford, R., Saez De Urabain, I. R., Karmiloff-Smith, A., & Smith, T. J. (2017). Daily touchscreen use in infants and toddlers is associated with reduced sleep and delayed sleep onset. *Scientific Reports*, *7:* 46104. http://doi.org/10.1038/srep46104

Csibra, G., Gergely, G., & Pisupati, S. (2009). Natural Pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. http://doi.org/10.1016/j.tics.2009.01.005

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. http://doi.org/10.1037/a0025659

Dink, J., & Ferguson, B. (2018). *eyetrackingR*. R package version 0.1.6. http://www.eyetracking-R.com.

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.

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, *69*(2), 133-149. https://doi.org/10.1006/jecp.1998.2438

Gros-Louis, J., West, M. J., & King, A. P. (2014). Maternal responsiveness and the development of directed vocalizing in social interactions. *Infancy*, *19*(4), 385-408. https://doi.org/10.1111/infa.12054

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., ... & Suma, K. (2015). The contribution of early communication quality to low-income children's language success. *Psychological Science*, *26*(7), 1071-1083. https://doi.org/10.1177/0956797615581493

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting Education in "Educational" Apps: Lessons From the Science of Learning. *Psychological Science in the Public Interest*, *16*(1), 3–34. https://doi.org/10.1177/1529100615569721

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, L., … Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word. *Monographs of the Society for Research in Child Development 65*(3), pp. i-vi+1-135. Retrieved from http://www.jstor.org/stable/3181533

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science, 9*(2), 131-13

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128-157. https://doi.org/10.1080/15250000701795598

Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *358*(1431), 549-559. https://doi.org/10.1098/rstb.2002.1237

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. https://doi.org/10.1111/1467-7687.00036

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. https://doi.org/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. https://doi.org/10.1080/15213260903562917

Krcmar, M., Grela, B., & Lin, K. (2007). Can Toddlers Learn Vocabulary from Television? An Experimental Approach. *Media Psychology*, *10*, 41–63. https://doi.org/10.108/15213260701300931

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science, 10*(1), 110-120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

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

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

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for Learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus. *The Institute of Electronics, Information and Communication Engineers Technical Report*, *16*, 11–15.

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

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

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., & Reserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental Science*, *2*(2), 164–173. https://doi.org/10.1111/1467-7687.00065

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. https://doi.org/10.1111/j.1467-8624.2011.01579.x

Posner, M. I., & Rothbart, M. K. (2007). Research on attention networks as a model for the integration of psychological science. *Annual Review of Psychology, 58*, 1-23. https://doi.org/10.1146/annurev.psych.58.110405.085516

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

Rideout, V. (2017). The Common Sense census: Media use by kids age zero to eight. San Francisco, CA: Common Sense Media.

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

Senju, A., & Csibra, G. (2008). Gaze Following in Human Infants Depends on Communicative Signals. *Current Biology*, *18*(9), 668–671. https://doi.org/10.1016/j.cub.2008.03.059

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences, 108*(15), 6038-6043. https://doi.org/10.1073/pnas.1017617108

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development, 43*, 549–565. https://doi.org/10.2307/1127555

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54*(2), 99-132. https://doi.org/10.1016/j.cogpsych.2006.05.001

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. https://doi.org/10.1111/1467-8624.00313

Tobii Technology, Inc. (2016). Tobii Studio User's Manual (version 3.4.5). Retrieved from https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf/?v=3.4.5

Tomasello, M. (2003). *Constructing a Language*. Cambridge, MA: Harvard University Press.

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. https://doi.org/10.1111/j.1467-8624.2006.00903.x

Vygotsky, L.S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

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. https://doi.org/10.1371/journal.pone.0030884

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020. https://doi.org/10.1037/xge0000014

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

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*(4), 553–566. https://doi.org/10.1037/0012-1649.30.4.553

Wu, Rachel, & Kirkham, Natasha Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. Journal of Experimental Child Psychology, 107(2), 118-136. http://doi.org/10.1016/j.jecp.2010.04.014

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*, 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Appendix A

Sample size decision

Since our paradigm was very novel, it was not possible to decide on the previous studies appropriate for power calculation, and thus we used a rule of thumb based on previous work's sample sizes. We acknowledge that this is less than ideal. Therefore, we now conducted power calculations based on three previous studies in the literature whose design is relevant to ours in order to get an estimate of required sample size that would have been suggested by previous work. We first summarize which data we draw on for our power calculations, and present in a Table the input and output data for our power calculations. For prospective sample size estimations, we assume a power of 80% and a significance threshold of p = .05.

A1. Verb learning from contingently reacting on-screen human teacher

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

This study was chosen, because it assesses word learning from a contingent on-screen teacher. It assessed verb learning from a live, video chat and yoked video condition in 24-30 month-olds. Since the data provided in the article do not allow effect size calculation for the difference between conditions, We here focus on the video chat condition, calculating an effect size for verb learning in this condition (looks to actions matching versus mismatching the meaning of the learned verb during test).

A2. Gaze following of on-screen contingent avatar

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. http://doi.org/10.1037/a0025659

This study was chosen since it assesses infants' reaction to contingent versus non-contingent non-human on-screen avatars. In this study, 8-months-old gaze-following to a contingently reacting versus not contingently reacting on-screen avatar is compared. We use the fixation duration measure of the comparison between conditions for effect size calculation.

A3. Word learning in 13-month-olds

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. Developmental Psychology, 30(4), 553. http://doi.org/10.1037/0012-1649.30.4.553

This study was chosen as one of the closest ones in age to the infants tested in the current study. Here, 13-month-old infants were taught one novel word-object association. We use the proportion of target choice in Study 2 to calculate effect sizes.

Our power calculation suggests that a sample size of 17 for a within-subject design and 30 per cell for a between-subject design could be sufficient (Table A1). However, given our new design, the fact that our infants were younger than in two of the studies used for power calculations, and the knowledge that novel word learning has been reported to work under some, but not all conditions at around 12 months of age, we think it is reasonable to have used the bigger sample size of n=36 for our within-participant study.

Table A1. Input data are derived from original articles (see text for details). Output data are results of power analysis. Note that n always refer to the sample size required per cell.

Input data				**Output data**		
Study	**t**	**n**	**design**	**d**	**n (within)**	**n (between)**
Roseberry 2014	7.06	12	within	2.04	4	5
Deligianni 2011	2.51	18	between	1.18	8	12
Woodward 1994	4.10	32	one-sample	0.72	17	30

Appendix B

Extraction of gaze information for matched conditions

We extracted infant gaze information onto the contingently reacting areas of interest from the CR+ condition in order to match the amount of preprogrammed movements in the CR- conditions. Infant gaze to the areas of interest elicited movement on screen in the CR+ condition as described in the main article. In order to match the amount of movement in the CR- condition, we modeled the amount of preprogrammed movement based on infant gaze in the CR+ conditions.

However, since our experiment had a within-participant design where half of the infants were tested first in the CR- condition, it was not possible to extract gaze information from the CR+ condition of the same experiment. Therefore, movement characteristics of the CR- condition were calculated from a pilot experiment with infants with the same age and background characteristics as those in the experiment (n = 32). The CR+ condition in this pilot experiment was identical to the one in the experiment, and we thus extracted infants' gaze behavior from this condition in order to program the non-contingent movement in the CR- condition. We extracted the number, delays, and lengths of gaze to the contingently reacting parts of the display for each infants. These were the virtual agent and the object during the familiarization phase, and the object during the naming phase. We first calculated the average number of looks to the virtual agent and object per trial. We then created distributions based on the means and standard deviation of the extracted gaze delays and lengths, from which we then drew the respective number of values to create a sequence of movements for each trial. We created four different sets of sequences, which were counterbalanced between infants. Table B1 shows the mean number and length of virtual agent and object movement extracted from the pilot data, as well as the actual number and length of looks to virtual agent and object during the CR+ condition.

Table B1. Length and number of looks to virtual agent and object during familiarization phase and naming phase.

	Looks during familiarization				**Looks during Naming**	
	Virtual Agent		**Object**		**Object**	
	Mean number (SD)	*Mean length in ms (SD)*	*Mean number (SD)*	*Mean length in ms (SD)*	*Mean number (SD)*	*Mean length in ms (SD)*
CR-	1.5 (1.3)	872 (660)	0.7 (1.0)	934 (1009)	1.5 (1.5)	1146 (1092)
CR+	2.3 (1.9)	842 (892)	1.1 (1.3)	464 (337)	1.9 (2.1)	637 (601)

Appendix C

Visual stimuli

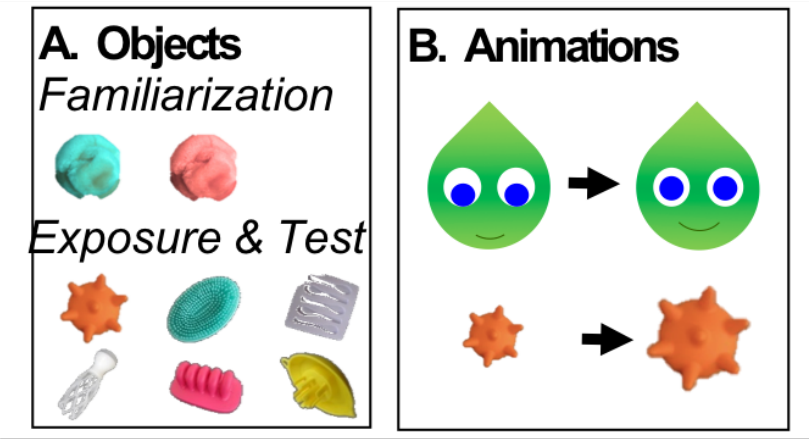

Figure C1. A. Objects used B. Schematic of animations in the CR+ condition. The virtual agent looked up and smiled if infants looked at her face. The objects inflated and deflated when infants looked at them.

Appendix D
Auditory stimuli: Carrier phrases

Greeting

こんにちは！今日は一緒に遊ぼうね。

Konnichiwa! Kyō-wa issho ni asobō ne.

"Hello! Let's play together today."

Exposure

Each of 3 targets within one condition is named twice with each of 3 carrier phrases.

1. これは［Target］だよ。［Target］。［Target］。

Kore wa ［Target］da yō. ［Target］. ［Target］.

"This is a [Target]. A [Target]. A [Target].

2. また［Target］だね。［Target］。［Target］。

Mata ［Target］da ne. ［Target］. ［Target］.

"There's the [Target] again. The [Target]. The [Target]."

3. ［Target］、面白いでしょう？［Target］。［Target］。

［Target］, omoshiroi deshō? ［Target］. ［Target］.

"The [Target] is fun, isn't it? The [Target]. The [Target]."

Test

Each of 3 targets within one condition is named twice with each of 3 carrier phrases.

1. あっ、［Target］だ。分かるかな。

A, ［Target］ da. Wakaru ka na?

"Oh, there's the [Target]! Can you find it?"

2. 見て見て、［Target］。いいね。

Mite mite, ［Target］. Ī ne.

"Look, look, the [Target] .How nice!"

Appendix E

Supplementary Analyses

This part reports on two supplementary analyses, in which different exclusion criteria were chosen. In the analysis reported in the main article, we chose to exclude based on condition: Thus, if a condition had less than ⅓ of test trials (less than two trials) left, it was excluded from analysis. To make sure this did not bias the results, we report two alternatives here.

E1: If any infant has less than ⅓ of test trials left in any condition, exclude the infant.
This analysis is based on 24 remaining infants. Model comparison revealed a significant difference between the two conditions [$\chi^2(1) = 21.89$, $p < .001$], with the slope for condition indicating higher target fixation proportion in the socially enriched than the matched condition ($b = 0.383$, $SE = 0.087$). There were no significant effects of condition on time terms.

E2: Do not exclude conditions that have less than ⅓ of trials left
This analysis keeps 1 more conditions in for 2 infants, and both conditions for 1 infants, meaning it is based on 37 infants. Model comparison revealed a significant difference between the two conditions [$\chi^2(1) = 37.46$, $p < .001$], with the slope for condition indicating higher target fixation proportion in the socially enriched than the matched condition ($b = 0.322$, $SE = 0.061$). There were no significant effects of condition on time terms.