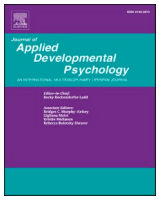
Journal of Applied Developmental Psychology 81 (2022) 101439

Contents lists available at ScienceDirect

Journal of Applied Developmental Psychology

journal homepage: www.elsevier.com/locate/jappdp



Contingent interaction with a television character promotes children's science learning and engagement

Ying Xu [a,b,*], Valery Vigil [b], Andres S. Bustamante [b], Mark Warschauer [b]

[a] School of Education, University of Michigan, United States of America
[b] School of Education, University of California Irvine, United States of America

ARTICLE INFO

Keywords:
Artificial intelligence
Contingent interaction
Screen media
Science learning
Engagement

ABSTRACT

While educational television or video programs are important and accessible learning resources for young children, the lack of contingent interaction afforded within this type of programming may limit how much children learn from them. In this project, we leveraged natural language processing technologies to enable contingent interaction between children and a television character, with the goal of enhancing children's learning and engagement. Our randomized controlled study involving 77 four to six year old children suggested that incorporating contingent interaction within video programs helped children better understand the science concepts introduced in the video. Children who had contingent interaction also showed a higher level of vocalization and more positive affect during video watching and developed a more positive perception of the media character, though they spent less time looking at the screen. These findings shed light on the potential of contingent interaction with on-screen media characters enabled by artificial intelligence to promote learning and engagement.

Children between the age of three to six spend, on average, about two hours daily watching television or videos (Rideout & Robb, 2020). Such high levels of media consumption have sparked research and debate on the consequences on learning in early childhood (for reviews, see Gunter & Gunter, 2019; Lillard & Peterson, 2011; Richert, Robb, & Smith, 2011). Some studies point to the learning affordances of video programming designed to be educational (Crawley et al., 2002; Mares & Pan, 2013). Such programs provide young children with learning resources that may not be available through their other daily activities, and this is particularly true for science-related topics (e.g., deep sea animals, volcanic eruptions, or outer space). In contrast, other studies focus on the apparent limitations of video programming and the constraints this format imposes on children's learning, particularly as compared to learning through interpersonal interaction (Jing & Kirkorian, 2020; Strouse & Samson, 2021). This difference between how children learn from video versus from live interpersonal interaction stems from video programming traditionally being a one-way transmission medium that does not allow for the kind of contingent interaction known to be vital for children's engagement and learning (Lauricella, Gola, & Calvert, 2011; Richert et al., 2011). Yet, as we discuss below, how much children can learn from video watching can be enhanced by allowing children to have contingent interaction while watching media, and studies have consistently found that such interaction can improve children's learning outcomes (Roseberry, Hirsh-Pasek, & Golinkoff, 2014; Troseth, Saylor, & Archer, 2006).

The recent development of artificial intelligence (AI) enables machines to simulate interpersonal interaction, in particular, communications using natural spoken language. Many young children already frequently "talk to" AI-enabled devices or toys available in their home, such as smart speakers (e.g., Garg & Sengupta, 2020; Xu & Warschauer, 2020a), robots (e.g., Michaelis & Mutlu, 2018), and Internet-connected stuffed toys (e.g., Druga, Williams, Park, & Breazeal, 2018; McReynolds et al., 2017). Children enjoy interaction with these devices and even ascribe them social properties, such as being able to feel emotions or form friendship bonds (Lovato, Piper, & Wartella, 2019; Xu & Warschauer, 2020b). Our study interrogates whether this kind of AI-mediated communication can be integrated into children's video programming to ameliorate video deficits. To this end, we allowed children to interact with an AI-enabled on-screen media character in an animated video program, and examined the impacts on children's learning from and engagement with the program as well as their perception of the AI-enabled media character.

* Corresponding author at: School of Education, University of Michigan, United States of America.
E-mail address: yxying@umich.edu (Y. Xu).

https://doi.org/10.1016/j.appdev.2022.101439
Received 7 July 2021; Received in revised form 14 June 2022; Accepted 14 June 2022
Available online 25 June 2022

Contingent interaction and learning from video programs

While studies have shown that young children can learn literacy, numeracy, facts, and socioemotional skills from educational video programs, it is also evident that video watching is not as effective for young children's learning as authentic, face-to-face interpersonal interaction (Roseberry et al., 2014; Troseth et al., 2006). Overall, such "video deficits" exist across children aged zero to six years, and the current literature suggests that such deficits peak around the first half of a child's second year (e.g., 12–15 months) and may slowly diminish with age (Barr, 2010; Strouse & Samson, 2021). Most often, scholars examining video deficits utilize children's third birthday (36 months) as a point of reference for when the phenomenon is substantially reduced for many tasks (Jing & Kirkorian, 2020; Roseberry et al., 2014). However, some scholars have noted sizable video deficits for children as old as five years of age (Barr, 2010). Some studies suggest that the lack of *contingent interaction* found in video programming may be a primary factor in such video deficits. In these studies, contingent interaction generally involves a conversation partner engaging in a meaningfully responsive way with a child viewer (Rochat, 2001). These conversation partners were either on-screen actors/characters or another person co-viewing the video with the child.

Studies focusing on interaction with on-screen actors/characters mostly involve a live person interacting with children in a format similar to video chats (Myers, LeWitt, Gallo, & Maselli, 2017; Roseberry et al., 2014; Troseth et al., 2006). For example, Troseth et al. (2006) found that children were better able to solve a practical problem when they received information through live video chat with a person, as compared to when they watched a pre-recorded video presenting the same information. In another study, live video chats enabled children's acquisition of novel vocabulary equally as well as face-to-face interaction (Roseberry et al., 2014). Similarly, Gaudreau et al. (2020) found that children who were read a story by an adult on video could learn vocabulary at a level comparable to children participating in face-to-face story time. In addition, a smaller number of studies examined the learning benefits of interacting with an animated character instead of a person on screen (Calvert et al., 2019; Hyde, Kiesler, Hodgins, & Carter, 2014). Most of these studies used a "Wizard of Oz" technique wherein an experimenter controlled the responses of the animated character. Using this technique, Calvert et al. (2019) found that children learned math concepts from a video significantly better when the video's main character asked children questions and replied in a timely and responsive manner as compared to watching the same video without this interaction. In fact, many commercial television shows aim to create the illusion of contingent interaction for the viewer by implementing a "pseudo-interaction" technique where a human actor or animated media character invites children to respond to their questions, pauses for a fixed amount of time, and then provides a generic response. This pseudo-interaction technique, however, does not appear to be more effective in promoting learning or engagement than videos without pseudo-interaction (Roseberry et al., 2014). This reinforces the importance of contingency and responsiveness in interactions for supporting children's learning from videos.

Another line of research examines how social interaction with human co-viewers, primarily parents, can support children's learning from screen media. Children consistently learn better when conversing with parents during television watching compared to watching alone, with the benefits including increased language development and understanding of the educational content (Ewin, Reupert, McLean, & Ewin, 2020; Strouse, O'Doherty, & Troseth, 2013; Strouse, Troseth, O'Doherty, & Saylor, 2018). For example, Strouse et al. (2013) found that when a co-viewing parent posed questions and gave responsive feedback, children learned more vocabulary words than when they watched television on their own or when their parents only asked questions but did not provide feedback. In reality, however, children frequently watch video programs without a parent available (Anderson & Hanson, 2017), and even parents who are present do not always actively engage in conversation with their children around the video program, missing the opportunity to provide meaningful scaffolding that promotes engagement and learning (Wang, 2014). This is partly because traditional children's video programming is not designed to encourage meaningful parent-child interaction during the video watching.

Engagement and perceptions

Contingent interaction during video watching, with human partners or on-screen characters, also affects children's engagement with the media content. Research has examined children's visual attention, affect, and vocalizations as indicators of engagement. For example, Strouse et al. (2018) found that when an on-screen actor responded contingently to children through a live video feed, children looked at the television screen more frequently than those viewing a pre-recorded video containing the same information. The same study found that children were more likely to respond to prompts from the contingent on-screen character than those in the pre-recorded video condition. Calvert et al.'s (2019) study also confirmed the advantages of contingent interaction in eliciting children's relevant vocalizations. Similarly, when comparing responsive and unresponsive co-viewers, children look to responsive co-viewers more often (Myers, Crawford, Murphy, Aka-Ezoua, & Felix, 2018), hinting at a triadic interaction (person-person-object) that can support social engagement for young children (Hobson, 2005). Lee, Heeter, and LaRose (2010) found that participants reported higher enjoyment when watching an interactive video narrative where they could navigate the protagonist's choice-making, than a linear non-interactive version of the same video.

Young children's perceptions of media characters are also influenced by whether a character is capable of contingent interaction. For example, children notice a character's inability to understand their responses and ultimately give generic feedback, and prefer to interact with a character that is capable of contingent conversation (Carter, Hyde, & Hodgins, 2017). By contrast, engaging in conversations without contingent feedback may cause children to doubt the character's reliability as a source of information (Breazeal et al., 2016). Several other studies also suggest that children develop social bonds with a character from contingent interactions (Brunick, Putnam, McGarry, Richards, & Calvert, 2016; Calvert et al., 2019; Gray, Reardon, & Kotler, 2017). For example, Brunick et al. (2016) suggested that a contingent interaction with a character could foster children's attachment (e.g., feeling of comfort and safety) with that character and also encourage children to perceive the character as more "real" and relevant to the real world. Furthermore, Calvert et al. (2019) found that children's social bonds with an animated contingent character, as measured by attachment and friendship, were positively linked to how much children learned from the video. Taken together, these results highlight the prospects of embedding contingent interaction within children's media to heighten engagement and foster positive perceptions of the characters.

Interaction with AI-enabled devices

Intelligent systems that support natural speech interaction may be especially valuable for young children, whose lack of proficiency in literacy and fine motor skills cause them difficulty in productively interacting in digital environments via other modalities, such as writing or touching specific places on a screen. Speech-based intelligent systems enable complex dialogue that mimics human-to-human conversation. While there are differing opinions on whether machine-mediated communication is "social" in nature (see, e.g., Roseberry et al., 2014), it is clear that state-of-the-art systems are capable of enabling temporally and semantically contingent dialogue with children (see, Xu, Branham, Collins, Deng and Warschauer, 2021 for a review).

A growing number of studies suggest that contingent interaction with speech-based intelligent systems can result in learning. For

example, Kory and Breazeal (2014) developed a robot learning companion embedded in a stuffed doll that taught children about exotic animals by engaging children in dialogue. The robot described the animal (e.g., "I like how it's white with such big antlers!") and intermittently asked children questions to allow verbal input (e.g., "Did you know it can go for weeks without drinking water?"). Children were able to memorize the information the robot told them. Another robot played a food-selection game with children and talked about that food item in French, showing that game-like conversation helped children learn French words (Freed, 2012). Moreover, interacting with intelligent systems can result in learning outcomes comparable to interacting with a human partner. For example, Xu and colleagues developed smart speakers that could narrate a story while engaging children in relevant conversation. They found that dialogue with the smart speaker benefited children's comprehension equally as well as dialogue with a human partner (Xu, Aubele, et al., 2021; Xu, Wang, Collins, Lee, & Warschauer, 2021).

Other studies examine children's perceptions of intelligent systems or characters in these systems. Perception measures generally focus on children's attachment, trust, and perceived capabilities of the intelligent systems. Danovitch (2019) suggested that children tend to trust digital systems as informants. Similarly, Di Dio et al. (2020) found that children who played a game with an interactive robot established trusting relationships with the robot that were similar to the relationships formed by a second group of children who played the same game with a human partner. Noles, Danovitch, and Shafto (2015) suggested that in their study, children under five years of age prefer to seek information from an interface with human embodiment (human frame) rather than an interface represented as a search window on a computer screen (technological frame). In another study, children anthropomorphized the smart speaker they conversed with and regarded it as sociable and smart (Xu & Warschauer, 2020b). Children's generally positive perception of smart speaker agents appeared tied to the systems' capability of comprehending speech input and providing contingent feedback. However, smart speakers' occasional failure to respond appropriately may weaken children's confidence in the system's ability to be responsive (Cheng, Yen, Chen, Chen, & Hiniker, 2018; Yarosh et al., 2018).

Dialogic interaction for scaffolding science learning

In this study, we examine preschool-aged children's science learning from watching an animated video with AI-assisted contingent interaction. We focus on science learning for two primary reasons. First, science learning resources are generally lacking for this age group despite the importance of science, and many preschool programs do not offer a formal science curriculum (Bustamante, Greenfield, & Nayfeld, 2018; Tu, 2006). As such, many young children acquire science knowledge and skills through informal activities, such as watching video programs. Second, science learning is a complex process. Preschool-aged children need scaffolding provided by knowledgeable mentors to fully comprehend scientific concepts (Fleer, 1992; Hsin & Wu, 2011). One key characteristic of "scaffolded science inquiry" is that adults take a supportive, rather than leading, role in children's learning activities (Weisberg, Hirsh-Pasek, Golinkoff, Kittredge, & Klahr, 2016). This way, children retain agency to direct their exploration and remain motivated and engaged while they discover science principles. Scaffolded science inquiry primes children to develop scientific thinking skills as well as an understanding of science concepts (for a review, see Weisberg, Hirsh-Pasek, & Golinkoff, 2013).

Dialogic interaction is central to scaffolded science inquiry, as language is the vehicle for thinking and learning. During dialogic interaction, children are exposed to scientific language, and verbal engagement and exchanges that lead to active, mindful, and reflective learning experiences (Alper, Masek, Hirsh-Pasek, & Golinkoff, 2018). For example, Ferrara, Hirsh-Pasek, Newcombe, Golinkoff, and Lam (2011) demonstrated that parents could scaffold children's development of spatial skills by using spatial language in dialogue with children during a building blocks activity. Parents' spatial language, in turn, increases children's use of spatial language and parent-child interactions result in children's learning of spatial concepts (Gunderson, Ramirez, Beilock, & Levine, 2012; Pruden, Levine, & Huttenlocher, 2011).

The current study

This study aims to examine the impact of contingent interaction with artificially intelligent media characters in video watching on children's learning. To this end, we developed a "*conversational video*" based on an animated science show. This conversational video leveraged natural language processing technologies to allow children to answer questions asked by the show's main character and receive contingent feedback. We conducted a randomized controlled trial in which children from 4 to 6 years old were assigned to watch a science video with or without contingent interaction with a conversational agent. Our study was guided by the following research questions: compared to a traditional video without interactive features; 1) does a conversational video improve children's learning of the science concepts?; 2) does a conversational video increase children's engagement as measured by vocalization, affect, and visual attention?; and 3) does a conversational video impact children's perceptions of the media character? We formed the following hypotheses based on the literature reviewed above:

H1. Contingent interaction with the intelligent characters will enhance children's science learning.

H2. Contingent interaction with the intelligent characters will enhance children's active engagement. Specifically, children who have the contingent interaction opportunities will produce more relevant vocalizations (H2a), show more positive affect (H2b), and pay more attention to the video (H2c).

H3. Contingent interaction with the intelligent characters will help children form a more positive perception of the media character.

Method

Study design

We utilized a two-condition experimental design with participants assigned to either the ***experimental condition*** in which children had contingent interaction with the show's main character as they watched the episode or the ***control condition*** in which children watched the same episode but without the opportunity for contingent interaction with the character.

In both conditions, children participated in two sessions that were scheduled one week apart. In the first session, children's English language proficiency was assessed using a computer-based assessment (i.e., Quick Interactive Language Screener; Golinkoff et al., 2017). In the second session, children watched one episode of the science show with or without contingent interaction. Children were then asked questions to assess how much they learned from the show and whether they had a positive perception of the show's main character. Both of the study sessions were carried out remotely due to the COVID-19 Pandemic; children participated in the study from their homes. Children used a laptop provided by our research team to watch the conversational video and used their own devices to communicate with the experimenter via video conferencing. No unstable connections or delays in video conferencing occurred during the study. Since the study was designed to have children watch the video independently, we expected similar outcomes if the study had been conducted in-person. Each study session was video recorded in its entirety.

Stimuli

The video children watched in this study was based on one 11-min episode of a PBS KIDS animated television show, *Elinor Wonders Why*. In this episode, Elinor, the main character, explores the viscosity of different liquids and learns about the concepts of concentration and dilution by observing how bees make honey. Children in the control group watched the episode in the same format that would be aired on PBS KIDS, and those in the experimental group watched the episode in the conversational format that incorporated AI-assisted interaction described below. The show had not been released to the public at the time of the experiment, so no participants had prior exposure to the show.

In the conversational video, the main character Elinor asked children questions during the episode and provided responsive feedback. Elinor asked nine questions, three of which were small talk moments, and the remaining six were content-based questions. The small talk happens right before the episode starts and aims to build rapport with the child while also evoking children's curiosity about the episode content. For example, Elinor opens up the conversation by asking children whether they like honey and how it feels when they touch it. The content-based questions were spread throughout the video and sought to clarify children's understanding of scientific facts or ask children to apply what they have learned in the video to help Elinor solve a similar problem. For example, Elinor asks children how bees turn nectar into honey and asks children what they should do to dilute ketchup to make it come out of a bottle faster. The conversational video lasted about 15 min, which was 4 min longer than the standard version of the episode.

The underlying natural language processing model analyzed children's responses using the Google cloud service (Sabharwal & Agrawal, 2020) and performed end-to-end language processing that classifies speech utterances into semantic intents (i.e., categorization of the intended meaning). Given that children can respond to a particular question in a variety of ways, we trained the agent to associate more than one semantic intent with each conversational opportunity. These intents were created based on predicted responses formulated by the research team, as well as children's actual utterances during field testing. We also included a fallback intent that was triggered when a child's utterance did not match any of the predefined intents or when the child did not respond to the question at all. When a fallback intent was triggered, Elinor scaffolded the conversation by rephrasing her original question using more accessible language (e.g., changing from an open-ended question to a multiple-choice question). If the child's response to the scaffolded question triggered the fallback intent again, Elinor then provided neutral feedback and explained the correct answer to the child. Fig. 1 displays a sample dialogue moment between a child and Elinor.

The visuals of each conversation moment were designed to be connected seamlessly with the rest of the episode. These seamless connections create an immersive and interactive viewing experience for children as if Elinor is actually talking with them within the television show. During the conversation moments, Elinor uses social cues, such as facial expressions, lip-flap, eye gaze, and body movements, in order to make the entire dialogue process feel more natural and encourage children to talk more in response to the prompts. Fig. 2 displays four still frames of Elinor talking to children during conversation moments.

Participants

Eighty-two children were recruited from communities in the Western USA. We reached out to a local community organization that was willing to distribute our study information to its listserv that hundreds of local families with young children subscribed to. We also contacted potential participants from a list of families who had previously participated in one of our previous studies. Children we recruited reflected a wide range of racial/ethnic, socioeconomic, and language backgrounds. Among the 82 children we recruited, 77 children completed both study sessions. The five children who did not complete the study were excluded because they were too distracted ($n = 3$), they were unwilling to stay seated ($n = 1$), or their parents gave them answers to the posttest questions ($n = 1$). Thus our final sample included 77 children, with 21 four-year-olds, 28 five-year-olds, and 28 six-year-olds. Forty-nine children were Latino (63.6%), 12 were White (15.6%), and 16 were Asian or mixed race (20.8%). Fifty-five of the children (71.4%) spoke predominantly English at home, whereas 22 spoke other languages at home, including Spanish, Chinese, and Japanese (28.6%). Forty-nine were girls (63.6%). Twenty-eight children were reported to have more than one monthly experience talking with smart speakers or other voice assistants on smartphones, and the rest of the forty-nine children never or rarely had such experience. Table 1 presents the participant information. Children's background information across the two experimental conditions was balanced.

Note that our sample size was relatively small due to the practical constraints imposed by the COVID-19 pandemic. A power analysis was conducted after we enrolled eighty participants. We planned on using linear regression with covariates to examine the condition difference with the goal of increasing the precision of estimates. Our covariates, including children's age and English language proficiency, are linearly correlated with the posttest outcomes and account for 30% to 50% of the variance in our science learning outcome measure. The power analysis suggested that when using linear regression with covariates, a sample size of 80 in two conditions would provide 80% statistical power to detect an effect size between 0.4 and 0.5 (Cohen's d) at $p < 0.05$. This minimum detectable effect size falls within the range of our previous experimental study on conversational agents ($d = 0.4$, Xu et al., 2021).

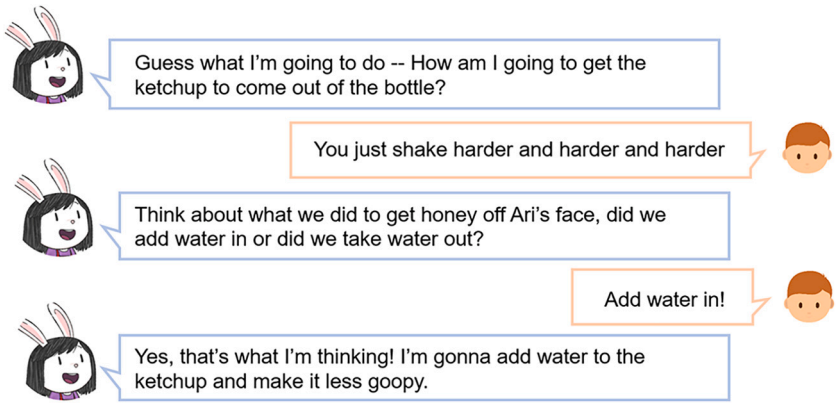


Fig. 1. Sample Dialogue Moment Between a Child and Elinor.

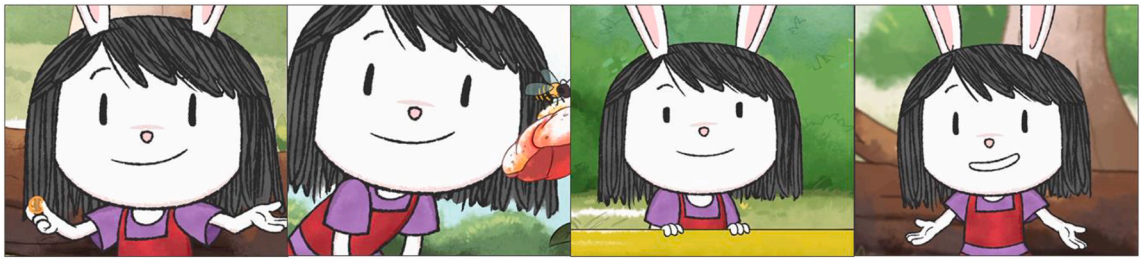


Fig. 2. Still Frames of Elinor Talking to Children During Conversation Moments.

Table 1
Background Information by Condition.

	Full Sample	Experimental Group	Control Group	Difference
QUILS	64.97 (30.78)	66.47 (30.99)	63.43(30.90)	$t(77) = 0.43$ $p = 0.67$
Female	63.64%	66.67%	60.53%	$\chi^2(1) = 0.10$, $p = 0.74$
Age in month	66.12 (9.72)	67.68 (9.24)	64.68 (10.08)	$t(77) = 1.37$ $p = 0.17$
Race/Ethnicity				$\chi^2(2) = 1.59$, $p = 0.45$
White	15.58%	25.51%	10.52%	
Latino	63.64%	61.54%	65.79%	
Others	20.78%	17.95%	23.68%	
Predominant Home Language				$\chi^2(1) = 0.11$, $p = 0.74$
English	71.43%	74.36%	68.42%	
Other	28.57%	25.64%	31.58%	
Prior CT Usage				$\chi^2(1) = 0.00$, $p = 0.95$
Heavy users	36.36%	38.46%	34.21%	
Non-heavy users	63.64%	61.54%	65.79%	
Mother's Education				$\chi^2(2) = 3.26$, $p = 0.20$
Less than high school	14.29%	20.51%	7.89%	
Above high school	20.78%	15.38%	26.32%	
Above Bachelor's degree	64.94%	64.10%	65.79%	
Typical TV Time During Weekdays				$\chi^2(2) = 0.49$, $p = 0.78$
<30 min	41.56%	43.59%	39.47%	
30–60 min	49.35%	46.15%	52.63%	
>60 min	9.09%	10.26%	7.89%	
N	77	39	38	

Note. Heavy prior use of conversational agents was defined as more than monthly. QUILS = Quick Interactive Language Screener; CT = conversational technologies.

Measures

Baseline language skills

Children's English proficiency was assessed via the computer-based Quick Interactive Language Screener (QUILS). This assessment consisted of 48 culturally neutral items that evaluated children's vocabulary, syntax, and language processes, with internal consistency among children ages 3 to 6 of 0.93 (Golinkoff et al., 2017). In addition, we asked parents to report their child's predominant language at home and subsequently categorized them into either English dominant or non-English dominant.

Science learning

To assess children's episode-specific science learning outcomes, the research team developed a questionnaire that was aligned with the Next Generation Science Standard (NGSS) and the US Department of Education's Ready to Learn Science Framework. These items were different from Elinor's questions in the episode. The questionnaire was reviewed by an advisory board of science curriculum consultants from the Ready to Learn program and other prominent experts on young children's science learning. For ten questions, we first asked children to freely recall the answer, and if children were not able to answer correctly, we provided them with three options to choose from. Two points were given to children for a correct answer without prompting, one point was given for a correct answer with prompting, and a score of zero was given for an incorrect answer. For two questions, children were asked to provide explanations of their answers. We scored their explanations from 0 to 4 points. A score of 0 indicated a completely incorrect answer; a score of 1 indicated an answer was incorrect but contained some correct ideas related to the episode; a score of 2 indicated an answer was largely correct but some language was inaccurate; a score of 3 indicated a correct answer with accurate language; and a score of 4 indicated a correct answer with additional accurate information supporting the answer. We calculated a total score by summing the points across all items, with a possible range from 0 to 28. The Cronbach's alpha of the learning outcome items was 0.81.

Engagement

Children's engagement during video watching was coded from the video captured from the computer camera during the sessions. Videos were divided into 5-s segments and each segment was coded by trained researchers (Willoughby, Evans, & Nowak, 2015; Zhou & Yadav, 2017). We coded three items: vocalizations, affective expressions, and visual attention. These three items were commonly coded in studies on children's use of screen media (see, e.g., Xu, Yau and Reich, 2021; Zhou & Yadav, 2017). For each time segment, we coded whether each item was present or not present. We then calculated the proportion of time segments each item was present. Eight children (three in the experimental and five in the control group) were excluded from the engagement analysis for failing to successfully record the video watching session, or because the child's face was significantly out of frame during the session recording. Thus, the analytic sample of this set of engagement outcomes consisted of 69 children.

Vocalizations. Children's vocalizations during each 5-s time segment of the video watching were transcribed and coded for comments relevant to the video content. Note that these vocalizations may be spontaneous or prompted by the agent in the video. Thirty percent of the time segments were double coded, and the Cohen's Kappa interrater reliability was 0.85 for narrative-relevant vocalization.

Affective expressions. Affective expressions were indicated by the presence or absence of children's positive expressions during each 5-s segment. Positive expression was recorded if the child showed at least one of the following expressive displays during the segment: smiling, cheering, clapping, dancing, jumping in excitement, laughing audibly, singing, showing eagerness, giggling, raising cheeks, pulling up lip corners, crinkling eyes, showing affection, smirking, speaking in a warm emotional tone, or using terms of endearment (Bai, Repetti, & Sperling, 2016). The interrater reliability was 0.74 among the 30% of time segments that were double coded.

Visual attention. Attention was coded as children's complete *visual* attention to the screen during the 5-s segment. If children's eyes were gazing at the screen during the entire time segment, their visual attention was coded as present. If children shifted their eye orientation away from the screen at any point, their visual attention was coded as absent. The interrater reliability was 0.81 among the 30% of time segments that were double coded.

Perceptions

To assess children's perceptions of the main character, Elinor, we used a survey questionnaire modified from Richards and Calvert (2017). The survey assesses children's perceptions of media characters along three aspects: the child's attachment to the character, whether the character could serve as a role model, and the perceived responsiveness of the character. To measure children's attachment to Elinor, we asked two questions such as “Do you want to make friends with Elinor?” To measure Elinor's ability to serve as a role model, we asked two questions such as “Is Elinor good at solving problems?” And to measure Elinor's perceived responsiveness we asked three questions such as “Can Elinor understand what you say?” We used a Smiley-o-meter to gather responses on a 0–2 scale (Read & MacFarlane, 2006). The smallest face has a null expression and corresponds to a score of 0 (“Not at all”), a bigger face with a slightly positive expression corresponds to a score of 1 (“Kind of”), and the biggest face with a highly positive expression corresponds to a score of 2 (“Very much”). All three faces were in the same color (i.e., yellow) to reduce possible bias introduced by color. Cronbach's alpha is 0.71 among these items.

Results

Children's interaction with the media character

This section reports descriptive statistics on children's interaction with the media character to help contextualize the comparison between the experimental and control groups. As shown in Table 2, overall, the 39 children in the experimental group responded to 81% of the questions asked by Elinor (e.g., approximately seven out of the nine questions a child received). The average length of children's responses was 2.5 words, and 75% of their responses were a direct answer to Elinor's question. In terms of correctness, among the six content-based questions with correct answers, children on average correctly answered three of them. There was a strong correlation between children's in-episode responses and their post-viewing science learning outcomes. The Pearson correlation of posttest science learning outcomes with response rate was 0.42 ($p = 0.01$), with response length was 0.39 ($p = 0.01$), and with relevancy was 0.48 ($p = 0.00$). The strongest Pearson correlation was with correctness at 0.71 ($p < 0.001$).

There appeared to be a significant difference in the response pattern across age, language proficiency, and children's predominant home language use; younger children, those who had a lower score in the language assessment, and those who did not primarily speak English at home were less likely to provide a verbal response, and, when they did verbally respond, their responses were shorter and less relevant to the question. For example, 4-year-old children responded to 68% of the questions with an average response length of 1.7 words, while 6-year-olds responded to 93% of the questions with an average response length of 3.1 words. Similarly, four-year-olds were able to directly and appropriately respond to 62% of the questions, while six-year-olds did so for 86%.

The agent was able to satisfactorily decipher and interpret children's responses and initiate feedback with appropriate timing, thereby ensuring the semantic and temporal contingency of the interaction. The speech-to-text translation accuracy rate was 79%, meaning that the agent correctly translated about eight out of every ten words spoken by a child. The intent classification accuracy rate was higher at 86%, meaning that the agent was able to accurately map children's responses to the correct intent and give correct feedback, on average, almost eight of the children's nine responses. The intent classification accuracy rate was higher because the language processing model used semantic-based understanding, and thus errors in specific words did not necessarily influence the overall meaning of a given response. For example, the meaning is similar between a child's actual response “the water can blow out” and the response “the water can go out” as translated by the machine. The accuracy rates on speech-to-text translation and intent classification were higher among older children and those with higher

Table 2
Children's Interactions With the Media Character by Age and English Proficiency.

	Child Responses				Character Performance	
Sample	Response Rate	Length	Relevant Response	Correctness	Speech-to-Text Accuracy	Intent Accuracy
Overall	0.81 (0.25)	2.47(1.83)	0.75 (0.29)	0.49 (0.26)	0.79 (0.22)	0.86 (0.13)
By Age						
4-year-olds	0.68 (0.26)	1.71 (1.15)	0.62 (0.36)	0.29 (0.33)	0.72 (0.25)	0.80 (0.18)
5-year-olds	0.75 (0.30)	2.30 (1.48)	0.72 (0.32)	0.53 (0.27)	0.76 (0.23)	0.85 (0.11)
6-year-olds	0.93 (0.09)	3.09 (2.28)	0.86 (0.11)	0.58 (0.19)	0.85 (0.20)	0.88 (0.10)
By English Proficiency						
Low	0.72 (0.29)	1.72 (1.09)	0.60 (0.34)	0.31 (0.31)	0.71 (0.27)	0.80 (0.17)
Medium	0.80 (0.26)	3.01 (2.60)	0.75 (0.23)	0.48 (0.26)	0.79 (0.23)	0.86 (0.14)
High	0.89 (0.18)	2.63 (1.25)	0.90 (0.18)	0.55 (0.20)	0.86 (0.18)	0.89 (0.10)
By Predominant Home Language Use						
Non-English	0.74 (0.28)	1.78 (0.91)	0.68 (0.33)	0.44 (0.31)	0.75 (0.26)	0.83 (0.16)
English	0.82 (0.23)	2.72 (2.01)	0.78 (0.26)	0.51 (0.22)	0.82 (0.20)	0.88 (0.11)

Note. Standard deviations are in parentheses. Response Rate = the percentage of Elinor's questions children provided with a verbal response; Length = the average number of words in children's verbal responses; Relevant Responses = the percentage of Elinor's question children provided with an on-topic, direct verbal response; Speech-to-Text Accuracy = the agent underlying Elinor' word-by-word accuracy rate of translating children's verbal speech to text; Intent Accuracy = the agent's accuracy rate of classifying children's verbal speech to intent category.

Table 3
Outcome Measures by Condition.

	Full sample	Experimental Group	Control Group	Cohen's *d*
Science Learning				
Mean (*SD*)	14.80 (6.02)	16.00 (5.26)	13.5 (6.54)	0.42†
Median [Min, Max]	15.0 [2.00, 26.00]	17.0 [4.00, 26.00]	14.5 [2.00, 25.00]	$p = 0.07$
Vocalization				
Mean (*SD*)	0.04 (0.04)	0.07 (0.03)	0.01 (0.01)	1.58***
Median [Min, Max]	0.04 [0, 0.18]	0.07 [0.02, 0.18]	0 [0, 0.04]	$p < 0.001$
Positive Expression				
Mean (*SD*)	0.06 (0.08)	0.09 (0.09)	0.04 (0.06)	0.44**
Median [Min, Max]	0.03 [0, 0.42]	0.05 [0, 0.42]	0.02 [0, 0.28]	$p = 0.006$
Visual Attention				
Mean (*SD*)	0.80 (0.16)	0.74 (0.17)	0.87 (0.12)	−0.82***
Median [Min, Max]	0.83 [0.19, 1.00]	0.77 [0.19, 0.99]	0.90 [0.46, 1.00]	$p < 0.001$
Perception of the Media Character				
Mean (*SD*)	9.30 (3.29)	10.1 (3.09)	8.44 (3.32)	0.50*
Median [Min, Max]	9.00 [0, 14.00]	10.0 [3.00, 14.00]	9.00 [0, 14.00]	$p = 0.04$
Attachment				
Mean (*SD*)	2.81 (1.19)	2.82 (1.23)	2.81 (1.17)	0.01
Median [Min, Max]	3.00 [0, 4.00]	3.00 [1.00, 4.00]	3.00 [0, 4.00]	$p = 0.97$
Role Model				
Mean (*SD*)	3.11 (0.96)	3.18 (0.896)	3.03 (1.03)	0.16
Median [Min, Max]	3.00 [0, 4.00]	3.00 [1.00, 4.00]	3.00 [0, 4.00]	$p = 0.49$
Responsiveness				
Mean (*SD*)	3.41 (1.93)	4.00 (1.61)	2.78 (2.07)	0.63**
Median [Min, Max]	4.0 [0, 6.00]	4.00 [0, 6.00]	3.00 [0, 6.00]	$p = 0.006$

*** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. † $p < 0.1$.

language proficiency (see Table 2). However, in all cases, the intent classification errors occurred when the agent classified a child's valid response as "fallback" and provided feedback that was generic but not inappropriate. The agent's performance deciphering and interpreting children's responses was consistent with the state-of-the-art natural language processing models reported in other studies focusing on children in this age range (Dietz et al., 2021).

Learning by video condition

Hypothesis 1 posited that children who had contingent interaction with the media character would have a higher science learning assessment score than the children who were in the broadcast video condition. The maximum score of the science learning assessment was 28 points, and children in our sample got an average score of 14.8 points (standard deviation of 6.0), slightly over half of the full score (see Table 3). As shown in Table 3, children in the experimental group outperformed those in the control group by 2.5 points, which equates to correctly answering one more question (out of 12 questions) via free recall (Cohen's d ES = 0.41, $p = 0.07$). Furthermore, a regression analysis adjusting for children's age in months, English proficiency, and prior agent usage suggested that children who watched the conversational video scored 0.33 standard deviations higher than did those who watched the broadcast video ($p = 0.05$). These findings support H1.

Post Hoc analysis of heterogeneous effects

Given that the analysis above suggested the conversational video has a significant effect on children's learning, we further analyzed whether this effect varied by children's age, English language skills, or home language use. We focused on these three variables given that prior research suggests that children's language skills and cognitive development (using age as a proxy) correlated with how children interact with and learn from digital media. Indeed, in our sample, children's age, English language skills, or home language use are significantly correlated with both posttest science learning outcomes and in-episode engagement (See Appendix B).

In terms of the direction of the heterogeneous effects, we did not have a priori hypotheses for this set of analyses, and the literature suggests two different directions for the potential interaction effect. On one hand, dialogic interaction, and the language and comprehension scaffolding they provide, are especially useful for children who are younger or who have lower levels of English language skill or home use (Reese, Leyva, Sparks, & Grolnick, 2010). As such, we may expect that dialogue with the media character would benefit younger children or those who have lower English skills or use more greatly than it would other children. On the other hand, research has also suggested that the effects of dialogic interaction depend partially on children's actual participation. In fact, we did find that older children and those with higher English proficiency or who spoke predominantly English participated in the dialogue more actively (i.e., responded to a higher percentage of questions). Thus we may also expect that older children and those with higher English proficiency or home use benefit more greatly from their interaction with the character than do other children.

Our results showed that neither age nor English proficiency were significant moderators (Table 4, Model 2 & Model 3). The effect size for home language use as the moderator was substantial; speakers of languages other than English at home benefited from the conversational videos 0.58 SD more than those children who spoke English at home. However, this effect was not statistically significant ($p = 0.07$; Table 4, Model 1).

Table 4
Regression Analyses of Interaction Effects on Science Learning.

	Model 1	Model 2	Model 3
Conversational Video	0.14 (0.19)	**0.31**† (0.16)	**0.30**† (0.16)
Non-English	−0.11 (0.27)	**−0.41*** (0.20)	**−0.41*** (0.20)
QUILS	**0.41***** (0.09)	**0.37**** (0.12)	**0.39***** (0.10)
Age	**0.30***** **(0.09)**	**0.30***** **(0.09)**	**0.22**† (0.12)
Heavy prior CT usage	0.27 (0.17)	0.24 (0.17)	0.25 (0.17)
Convo Video*Non-English	**0.58**† (0.33)		
Convo Video *QUILS		0.03 (0.16)	
Convo Video *Age			0.16 (0.16)
Intercept	−0.35*** (0.16)	−0.41* (0.16)	0.42* (0.16)
R^2	0.55	0.54	0.54

Note. Heavy prior usage was defined as more than monthly. QUILS = Quick Interactive Language Screener; CT = conversational technologies.
*** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. † $p < 0.1$.

Engagement by video condition

Hypothesis 2 was that interaction with the media character would enhance children's engagement as measured by their vocalizations, positive expressions, and visual attention during video watching. In terms of vocalizations, children in the conversational video condition made comments about 7% of the time, while those in the control group rarely did so (1% of the time, Cohen's d ES = 1.58, $p < 0.001$). Children who watched the conversational video also showed more instances of positive expressions (9% of the time) as compared to the children in the control group who showed positive expressions about 4% of the time (Cohen's d ES = 0.44, $p < 0.01$). While children in both conditions looked at the screen for the majority of the video watching, children in the control condition remained visually oriented to the screen longer (87% of the time) than did those in the experimental condition (74% of the time, Cohen's d ES = -0.82, $p < 0.001$). This seemed due to children gazing away from the screen when they were formulating responses to questions.

Regression models adjusting for children's age, language proficiency, and prior experiences with conversational technologies suggested that interacting with the media character during video watching increased children's relevant vocalizations by 1.61 *SD* ($p < 0.001$) and increased their positive expressions by 0.70 *SD* ($p = 0.003$). However, interaction with the character reduced the time children looked at the screen by 0.93 *SD* ($p < 0.001$). Given the relative rarity of the vocalization and expression items, we also conducted a robustness check using Tobit regression models which are designed to estimate linear relationships between variables when floor or ceiling effects are present (McBee, 2010). The results were consistent and confirmed that the conversational video resulted in significantly higher levels of relevant vocalization and positive expressions.

Overall, these analyses supported H2a and H2b regarding the effects that interacting with media characters has on vocalization and expressions. However, this analysis failed to support H2c regarding the interaction's effect on visual attention, in fact, we found a significant difference in the opposite direction.

Post Hoc analysis of engagement without interaction moments

The analyses above suggested that children's engagement differed significantly based on their video watching conditions. We conducted further analyses to understand whether the differences in engagement occurred primarily within the conversational moments. To do so, we created a subset of the data by excluding the conversational moments. We then examined the frequency of vocalizations, positive expressions, and visual attention (Table 5) and reran our regression analysis on all engagement outcome variables. We found that children in the control and experimental groups had similar levels of positive expression and visual attention in non-conversational moments, while children in the experimental group had slightly more vocalization ($p = 0.08$).

Table 5
Engagement Measures by Condition Excluding Conversational Moments.

	Experimental Group	Control Group	Cohen's *d*
Relevant Vocalization			
Mean (*SD*)	0.02 (0.02)	0.01 (0.01)	**0.50†**
Median [Min, Max]	0 [0, 0.05]	0 [0, 0.04]	
Positive Expression			
Mean (*SD*)	0.06 (0.07)	0.04 (0.06)	0.28
Median [Min, Max]	0.03 [0, 0.34]	0.02 [0, 0.28]	
Visual Attention			
Mean (*SD*)	0.85 (0.13)	0.87 (0.12)	0.15
Median [Min, Max]	0.89 [0.42, 0.99]	0.90 [0.46, 1.00]	

*** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. † $p < 0.1$.

Perceptions of the media character by video condition

Hypothesis 3 posited that contingent interaction with the media character would help children establish a more positive perception of the character. Overall, children's responses to our survey indicated that children in both the experimental and control group held a moderate to positive perception of Elinor as measured on the three-point Smiley-o-meter scale (average of 1.3 on a scale of 0 to 2). As shown in Table 3, the aggregated score of all seven items for children who watched the conversational video was 1.7 points higher than that of children in the control group (10.1 compared to 8.4 out of a maximum of 14; Cohen's d ES = 0.50, $p < 0.05$). When broken down into the three subscales (i.e., attachment, role model, responsiveness), children who watched the conversational video and those in the control group reported similar levels of attachment and role model identification with Elinor. However, children in the conversational video condition perceived Elinor to be more responsive as compared to children in the control condition (Cohen's d ES = 0.63, $p < 0.01$).

We then employed regression analyses for the overall perception scores and the three subscales, controlling for children's age, language proficiency, and prior experience with conversational technologies. For the overall score, children who watched the conversational video reported a significantly higher perception score (*beta* = 0.50, $p = 0.04$). Analyzing the individual subscales revealed that, while there was not a significant difference between the two groups on the attachment score (*SD* = 0.04, $p = 0.86$) or role model score (*SD* = 0.15, $p = 0.56$), children who watched the conversational video were significantly more likely to perceive Elinor as capable of responsiveness (e.g., able to hear, understand, and solve a problem; *SD* = 0.64, $p = 0.007$). These findings supported H3 regarding the interaction's effects on children's overall perception.

Discussion

Children spend an average of two hours watching television and video programs every day. However, non-interactive programs based on one-way media do not effectively capitalize on how children learn. While laboratory studies have confirmed the usefulness of allowing children to interact contingently with on-screen characters, these characters are controlled in real-time by a researcher makes this approach impractical on a larger scale. This study explored artificial intelligence as a potential solution to this dilemma. We conducted an experiment to examine the impact of interaction with an intelligent media character on children's learning (RQ1), engagement (RQ2), and perceptions of an animated media character (RQ3).

Our first research question examined whether interacting with a media character supports children's learning. We found that children who had contingent interaction with Elinor, the main character in *Elinor Wonders Why*, scored better on a science assessment of the concepts introduced in the show compared to children who watched a non-interactive version of the video. This is consistent with other studies that showed the advantages of learning from live on-screen human actors as compared to pre-recorded video (Roseberry et al., 2014; Troseth et al., 2006). Our finding also echoed Calvert et al. (2019) that suggested on-screen animated characters increased children's learning outcomes if these characters were designed to interact with children. However, while our study provided important evidence regarding the benefits of children's contingent interaction with media characters as compared to non-interactive characters, it did not directly speak to whether such contingent interaction would be more beneficial than pseudo-interactions, as we did not include a comparison group where Elinor asked questions and gave generic feedback after a fixed amount of time. Indeed, at least one study has suggested that carefully designed pseudo-interactive videos incorporating participatory cues could be just as effective as fully contingent videos in helping children learn vocabulary and comprehend stories (Gaudreau et al., 2020). To answer this question more fully, we are currently conducting another experiment study by

adding a pseudo-interactive comparison group.

Further, our post hoc analysis suggested no significant differences in learning benefits of the conversational video depending on children's age or English language proficiency. However, our data were inconclusive as to whether contingent interaction with Elinor provided extra benefit to children who did not primarily speak English at home. The interaction between learning and speaking a language other than English at home had a substantial effect size (0.62 SD) but did not reach statistical significance at the 0.05 level. Further analysis indicated a number of other differences between those who did and did not speak English at home. Children who did not speak English at home had a significantly lower English proficiency score than their English speaking peers (39 vs. 85 as measured by QUILS). They were also more likely to have a mother with a lower level of education. For children who predominantly spoke English at home, 76% of mothers had a Bachelor's degree or above and only 5% had less than a high school education. For children who did not predominantly speak English at home, only 33% of mothers had at least a Bachelor's degree, and 38% did not graduate high school. We consider our data inconclusive as to whether a moderation effect exists for home language use, and, if so, what combination of language and socioeconomic factors contribute to it. A larger study would be required to tease this out further.

Our second research question looked at children's engagement during video watching. Not surprisingly, we found that children in the conversational video group generated significantly higher levels of vocalization, suggesting that Elinor's questions effectively elicited children's verbal responses. Children in the conversational video group were also more likely to express positive affect (e.g., smile, laugh). Within the conversational video group, about half of children's positive expressions (56%) happened during the "conversational moments," in particular, when Elinor agreed with a child's accurate response (e.g., Elinor said "That is a great idea!" or "Yeah I think so too!"). This echoes research suggesting that media characters' contingent responses heighten children's enjoyment of video programs (Xu & Warschauer, 2020c). Finally, children's interaction with Elinor did not increase their visual attention to the screen during the video watching. Instead, children in the conversational video condition looked at the screen less frequently than did those in the control group. However, we noticed that most of the visually distracting moments occurred when children listened and responded to a question. During these times, children tended to look toward a nearby family member or simply raise their head as they contemplated their answers. Although counterintuitive, this is consistent with research on children's visual attention to a storybook during dialogic reading with adults and conversational agents (Authors, under review). Children tend to look at learning materials (e.g., storybooks, devices) less frequently when they are interacting directly with someone else. Indeed, the equivalent visual attention outside the conversational moments confirmed this hypothesis. As such, while visual attention is commonly used as a proxy for engagement, these findings imply that a lack of visual attention during interactive moments may not correspond to a lack of engagement. Taken together, these findings suggest that children's contingent interaction with a media character increases their vocalizations and positive affect during video watching but not their visual attention.

Our third research question focused on children's perceptions of the media character. Overall, children who watched the conversational video reported a higher level of positive perceptions of Elinor. In particular, children who watched the conversational video were more likely to perceive Elinor as capable of responsiveness. However, children in the interactive condition did not differ from children in the control group in their attachment to Elinor or their perception of Elinor as a role model. This might be explained by children's short exposure to Elinor during one video watching session. While some research suggests that children can begin to form a connection with characters during such short exposures, this type of bonding is likely strengthened by repeated exposures. Thus, we may expect that the added value of contingent interaction would be more readily apparent the more children interact with Elinor.

Overall, our study suggests contingent interaction with an intelligent media character during video watching had positive effects on children's learning, engagement, and perceptions of the character. It is reasonable to speculate that these positive effects could be the result of cognitive and/or social mechanisms at play during contingent interaction. In terms of cognition, dialogue with Elinor engaged children in science-related discussions that could clarify their misconceptions, reinforce their correct understanding, and encourage them to verbalize their thought processes. Alternatively, children's contingent interaction with Elinor may cause them to view Elinor as socially relevant, thus leading them to trust and value the information conveyed in the video. However, the question as to whether people of any age can establish truly *social* relationships with even the most sophisticated machines is controversial. It is unclear whether social elements are actually involved in children's interaction with Elinor. While our study shows some positive effects on children's learning, engagement, and perceptions, further investigation into the nature of children's interaction with artificially intelligent agents and the mechanisms at play is warranted.

In addition, our study focused on young children's science learning given the relative complexity of learning abstract science concepts (e.g., dilution, concentration, etc.) and the established research on scaffolded science inquiry on which we could base our conversational video design. However, we believe that contingent interactions with conversational agents could be used to further other learning domains, such as other STEM subjects, language and literacy, and social-emotional skills. Numerous studies discussed in our literature review section have suggested that providing children with contingent dialogue during video watching supported their vocabulary learning, and Peebles, Bonus, and Mares (2018) also found that it had positive effects on children's socioemotional understanding of television narratives. Nevertheless, although we believe adding AI-based interactions to videos could be useful for learning in other domains, we also believe that the design of such videos should be grounded in domain-specific teaching knowledge (e.g., NGSS for science in our study) to maximize their benefits.

Practical implications

Our study has practical implications for the media industry interested in early childhood education. Young children's media consumption has steadily increased over time (Rideout & Robb, 2020) and there is no reason to believe that that trend will reverse. As such, it is increasingly important to develop high-quality programming that can add value to screen time and make it available to children at a larger scale. Our study presents a case that AI technologies can be used to facilitate this goal. Children now spend more than two-thirds of their video-watching time using Internet-connected devices (Rideout & Robb, 2020), thus making it feasible to incorporate speech recognition and natural language technologies into children's video programs. Indeed, major media producers, such as PBS KIDS and Sesame Workshop, have recognized this development and have envisioned the potential of incorporating AI into video programs believing that contingent conversation with on-screen media characters will improve children's

learning from video programs (Brunick et al., 2016; Calvert, 2021). More importantly, companies such as Google, Amazon, and IBM have released development tools that allow developers to more easily build conversational applications that incorporate speech recognition and natural language processing. The conversational video in our study relied on a development tool released by Google that had a satisfactory level of accuracy when processing children's speech. We believe that the number of tools available to developers and their quality will continue to grow as AI technology continues to advance. Taken together, our study demonstrates that such an approach is technically feasible, but also effective for promoting children's science learning and engagement, thus providing evidence that can spur major media producers to further innovation in this direction.

Our study has particular relevance to minority children and those from a lower socioeconomic background, and this is one of the reasons why we recruited the majority of our participants from a low-income Latino community. Latino children, on average, spend more than twice as much time per day watching videos or television as White children (Rideout & Robb, 2020). Though this is often portrayed negatively, recent research suggests that this discrepancy is in part due to Latino families' appreciation of the educational opportunities of television and video for their children (Kalinowski, Xu, & Salen, 2021). Moreover, these children also have limited access to early science instruction, both at school and at home. Though Latino parents want to spend time with their children and help their education, they often lack the confidence in how to best help them, particularly around science (Silander et al., 2018). As such, providing minority children with high-quality, accessible digital content could help ameliorate the gap in early science learning.

Limitations and future directions

There are several ways this line of research could be extended.

First, the children in our study watched the conversational video only one time. Although this relatively short exposure improved children's learning of science content specific to the episode, the longer-term impacts of this kind of media usage are unclear. Future research should incorporate a longitudinal design that would provide children with sustained access to this type of conversational video. This could illuminate whether children's ongoing contingent interaction with a media character can enhance their general science knowledge and skills and ability to apply them to solve other science problems. In addition, research should examine the role that novelty and familiarity play in such interaction. It is possible that as children become more familiar with this type of interactive video, they will be more comfortable actively participating, thus leading to increased learning benefits. However, it is also possible that the positive effects of interaction with intelligent characters may decrease over time as some of the novelty wears off.

Second, the conversational video used in this study could not perfectly interpret children's responses (intent classification accuracy at 86%). It is unclear if the character's occasional feedback errors may have dampened the effectiveness of the conversational video. As such, future research may want to compare this fully automated conversational video with a condition in which the character feedback is controlled by a behind-the-scenes experimenter to ensure accuracy (i.e., Wizard of Oz). Further, while our conversational video was designed to enable interaction that is both temporally and semantically contingent, it is worth exploring the respective effects of these two types of contingency. Indeed, in Carter et al.'s (2017) study, three- to five-year-old children were more likely to verbally engage with programs that wait for their response (i.e., temporal contingency) than those that do not, but providing responsive feedback (i.e., semantic contingency) above and beyond temporal contingency did not further enhance children's response rates. It would be interesting to extend Carter et al.'s (2017) initial study to the context of science learning through animated programs.

Third, the COVID-19 pandemic impacted our ability to recruit as many participants as we would have liked. Our sample size of eighty only allowed us to detect relatively large effect sizes in our regression analyses when examining condition effects. Furthermore, our sample size did not provide sufficient power to test the heterogeneous effects of video watching across sub-groups. We plan on extending this study with a larger sample size following the relaxation of pandemic-related restrictions.

Lastly, future studies may want to explore the mechanisms through which contingent interactions with media characters may improve learning outcomes. Specifically, children's enhanced science learning outcomes may be associated with their heightened engagement during the video watching (Xu & Warschauer, 2020c) as well as the increased perception of the media character (Calvert et al., 2019). While the sample size of this current study was not sufficient to carry out mediation analyses, the current data suggested that children's frequency of vocalization and their perceptions of Elinor might be positively correlated with the posttest science learning outcome (see Appendix A for a Pearson correlation table among outcome measures). This provided some initial evidence for exploring the mechanism of learning in the future.

Conclusion

As the time that children spend watching video increases and the mode of watching shifts to Internet-connected devices, it is imperative to investigate how new forms of video watching may better support learning. This study leveraged natural language processing technologies to allow children to interact with the main character of a science animated program. Our findings suggest that enabling this kind of contingent interaction between child viewers and media characters can bring additional educational benefits not available through traditional, non-interactive video watching. Research and design communities should take advantage of evolving conversational technologies to maximize the educational benefit of children's screen time.

CRediT authorship contribution statement

Ying Xu: Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **Valery Vigil:** Writing – original draft, Data curation, Resources. **Andres S. Bustamante:** Writing – review & editing. **Mark Warschauer:** Writing – review & editing, Supervision.

Declaration of Competing Interest

We have no known conflict of interest to disclose.

Acknowledgements

This research is based upon work supported by the National Science Foundation under Grant No. 1906321 and No. 2115382. Production of the conversational episodes is supported by the Corporation for Public Broadcasting. Funding for Elinor Wonders Why is provided in part by a Ready To Learn grant from the U.S. Department of Education [PR/Award No. U295A150003, CFDA No. 84.295A], and by the Corporation for Public Broadcasting.

Appendix A. Pearson correlation among outcome measures

	Learning	Vocalization	Expression	Attention
Learning	–			
Vocalization	0.17	–		
Expression	−0.04	0.42***	–	
Attention	0.05	−0.60***	−0.43***	–
Perception	0.20	0.15	−0.04	−0.10

*** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. † $p < 0.1$.

Appendix B. Bivariate regression analyses between demographic variables

	Science Learning	Vocalization	Positive expression	Visual attention	Perception
QUILS	0.57 (0.10) ***	0.05 (0.12)	−0.21 (0.19)	0.35 (0.11)**	−0.02 (0.13)
Age in months	0.55 (0.10)***	0.04(0.12)	−0.07 (0.12)	0.16 (0.12)	0.16 (0.12)
Race/Ethnicity (reference group: White)					
Latino	−0.90 (0.30)**	−0.29 (0.33)	0.04 (0.33)	−0.62 (0.30)*	0.02 (0.29)
Others	−0.21 (0.39)	−0.31 (0.42)	−0.21 (0.42)	0.27 (0.39)	−0.11 (0.42)
Predominant Home Language (Reference group: English)					
Non-English	−0.97 (0.24) ***	0.06(0.27)	0.10 (0.27)	−0.42 (0.27)	0.01 (0.29)
Prior CT Usage (Reference group: non-heavy users)					
Heavy users	0.31 (0.25)	−0.11 (0.26)	−0.52 (0.25)*	−0.00 (0.26)	0.02 (0.27)
Mother's Education (Reference group: Less than high school)					
Above high school	−0.07 (0.42)	0.26 (0.35)	0.16 (0.35)	−0.56 (0.34)	0.62 (0.38)
Above Bachelor's degree	0.44 (0.34)	−0.14 (0.32)	−0.36 (0.31)	−0.55 (0.30)	0.10 (0.33)
Typical TV Time During Weekdays (Reference group: <30 min)					
30–60 min	−0.02(0.25)	−0.17 (0.26)	0.26 (0.25)	−0.72 (0.24)**	0.08 (0.27)
>60 min	−0.27 (0.54)	−0.26 (0.54)	−0.44 (0.54)	−0.59 (0.51)	0.40 (0.55)

*** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. † $p < 0.1$.

References

Alper, R. M., Masek, L. R., Hirsh-Pasek, K., & Golinkoff, R. (2018). 5."Languagizing" the early childhood classroom: supporting children's language development. In *What Teachers Need to Know About Language* (pp. 85–94). Multilingual Matters.

Anderson, D. R., & Hanson, K. G. (2017). Screen media and parent–child interactions. In *Media Exposure During Infancy and Early Childhood* (pp. 173–194). Cham: Springer.

Bai, S., Repetti, R. L., & Sperling, J. B. (2016). Children's expressions of positive emotion are sustained by smiling, touching, and playing with parents and siblings: A naturalistic observational study of family life. *Developmental Psychology, 52*(1), 88–101. https://doi.org/10.1037/a0039854

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review, 30*(2), 128–154.

Breazeal, C., Harris, P., DeSteno, D., Kory Westlund, J., Dickens, L., & Jeong, S. (2016). Young children treat robots as informants. *Topics in Cognitive Science, 8*(2), 481–491. https://doi.org/10.1111/tops.12192

Brunick, K. L., Putnam, M. M., McGarry, L. E., Richards, M. N., & Calvert, S. L. (2016). Children's future parasocial relationships with media characters: the age of intelligent characters. *Journal of Children and Media, 10*(2), 181–190.

Bustamante, A. S., Greenfield, D. B., & Nayfeld, I. (2018). Early childhood science and engineering: Engaging platforms for fostering domain-general learning skills. *Education in Science, 8*(3), 144.

Calvert, S. L. (2021). Intelligent digital beings as children's imaginary social companions. *Journal of Children and Media*, 1–6.

Calvert, S. L., Putnam, M. M., Aguiar, N. R., Ryan, R. M., Wright, C. A., Liu, Y. H. A., & Barba, E. (2019). Young children's mathematical learning from intelligent characters. *Child Development, 91*(5), 1491–1508.

Carter, E. J., Hyde, J., & Hodgins, J. K. (2017, June). Investigating the effects of interactive features for preschool television programming. In *Proceedings of the 2017 conference on interaction design and children* (pp. 97–106).

Cheng, Y., Yen, K., Chen, Y., Chen, S., & Hiniker, A. (2018, June). Why doesn't it work? Voice-driven interfaces and young children's communication repair strategies. In *Proceedings of the 17th ACM conference on interaction design and children* (pp. 337–348).

Crawley, A. M., Anderson, D. R., Santomero, A., Wilder, A., Williams, M., Evans, M. K., & Bryant, J. (2002). Do children learn how to watch television? The impact of extensive experience with Blue's Clues on preschool children's television viewing behavior. *Journal of Communication, 52*(2), 264–280.

Danovitch, J. H. (2019). Growing up with Google: How children's understanding and use of internet-based devices relates to cognitive development. *Human Behavior and Emerging Technologies, 1*(2), 81–90.

Di Dio, C., Manzi, F., Itakura, S., Kanda, T., Ishiguro, H., Massaro, D., & Marchetti, A. (2020). It does not matter who you are: fairness in pre-schoolers interacting with human and robotic partners. *International Journal of Social Robotics, 12*(5), 1045–1059.

Dietz, G., Le, J. K., Tamer, N., Han, J., Gweon, H., Murnane, E. L., & Landay, J. A. (2021). *StoryCoder: teaching computational thinking concepts through storytelling in a voice-guided app for children*.

Druga, S., Williams, R., Park, H. W., & Breazeal, C. (2018, June). How smart are the smart toys? Children and parents' agent interaction and intelligence attribution. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (pp. 231–240).

Ewin, C. A., Reupert, A. E., McLean, L. A., & Ewin, C. J. (2020). The impact of joint media engagement on parent–child interactions: A systematic review. *Human Behavior and Emerging Technologies, 3*(2), 230–254.

Ferrara, K., Hirsh-Pasek, K., Newcombe, N. S., Golinkoff, R. M., & Lam, W. S. (2011). Block talk: Spatial language during block play. *Mind, Brain, and Education, 5*(3), 143–151.

Fleer, M. (1992). Identifying teacher-child interaction which scaffolds scientific thinking in young children. *Science Education, 76*(4), 373–397.

Freed, N. A. (2012). *" This is the fluffy robot that only speaks french": language use between preschoolers, their families, and a social robot while sharing virtual toys (Doctoral dissertation, Massachusetts Institute of Technology)*.

Garg, R., & Sengupta, S. (2020, April). Conversational technologies for in-home learning: using co-design to understand children's and parents' perspectives. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13).

Gaudreau, C., King, Y. A., Dore, R. A., Puttre, H., Nichols, D., Hirsh-Pasek, K., & Golinkoff, R. M. (2020). Preschoolers benefit equally from video chat, pseudo-contingent video, and live book reading: implications for storytime during the Coronavirus pandemic and beyond. *Frontiers in Psychology, 11*, 2158.

Golinkoff, R. M., De Villiers, J. G., Hirsh-Pasek, K., Iglesias, A., Wilson, M. S., Morini, G., & Brezack, N. (2017). *User's manual for the quick interactive language screener (QUILS): A measure of vocabulary, syntax, and language acquisition skills in young children*. Paul H: Brookes Publishing Company.

Gray, J. H., Reardon, E., & Kotler, J. A. (2017, June). Designing for parasocial relationships and learning: Linear video, interactive media, and artificial

intelligence. In *Proceedings of the 2017 Conference on Interaction Design and Children* (pp. 227–237).
Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: the role of the linear number line. *Developmental Psychology, 48*(5), 1229.
Gunter, B., & Gunter, J. (2019). *Children and television*. Routledge.
Hobson, R. P. (2005). What puts the jointness into joint attention? In N. Eilan, C. C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: Communication and other minds: Issues in philosophy and psychology* (pp. 185–204). New York, NY: Clarendon Press/Oxford University Press.
Hsin, C. T., & Wu, H. K. (2011). Using scaffolding strategies to promote young children's scientific understandings of floating and sinking. *Journal of Science Education and Technology, 20*(5), 656–666.
Hyde, J., Kiesler, S., Hodgins, J. K., & Carter, E. J. (2014, April). Conversing with children: Cartoon and video people elicit similar conversational behaviors. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1787–1796).
Jing, M., & Kirkorian, H. L. (2020). Video Deficit in children's early learning. *The International Encyclopedia of Media Psychology*, 1–8.
Kalinowski, R., Xu, Y., & Salen, K. (2021). The ecological context of preschool-aged children's selection of media content. In *Proceedings of the 2021 CHI conference on human factors in computing systems*.
Kory, J., & Breazeal, C. (2014, August). Storytelling with robots: Learning companions for preschool children's language development. In *The 23rd IEEE international symposium on robot and human interactive communication* (pp. 643–648). IEEE.
Lauricella, A. R., Gola, A. A. H., & Calvert, S. L. (2011). Toddlers' learning from socially meaningful video characters. *Media Psychology, 14*(2), 216–232.
Lee, M. S., Heeter, C., & LaRose, R. (2010). A modern Cinderella story: a comparison of viewer responses to interactive vs linear narrative in solitary and co-viewing settings. *New Media & Society, 12*(5), 779–795.
Lillard, A. S., & Peterson, J. (2011). The immediate impact of different types of television on young children's executive function. *Pediatrics, 128*(4), 644–649.
Lovato, S. B., Piper, A. M., & Wartella, E. A. (2019, June). Hey Google, do unicorns exist? Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM international conference on interaction design and children* (pp. 301–313).
Mares, M. L., & Pan, Z. (2013). Effects of Sesame Street: A meta-analysis of children's learning in 15 countries. *Journal of Applied Developmental Psychology, 34*(3), 140–151.
McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *The Gifted Child Quarterly, 54*(4), 314–320.
McReynolds, E., Hubbard, S., Lau, T., Saraf, A., Cakmak, M., & Roesner, F. (2017, May). Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 5197–5207).
Michaelis, J. E., & Mutlu, B. (2018). Reading socially: Transforming the in-home reading experience with a learning-companion robot. *Science robotics, 3*(21).
Myers, L. J., Crawford, E., Murphy, C., Aka-Ezoua, E., & Felix, C. (2018). Eyes in the room trump eyes on the screen: effects of a responsive co-viewer on toddlers' responses to and learning from video chat. *Journal of Children and Media, 12*(3), 275–294. https://doi.org/10.1080/17482798.2018.1425889
Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2017). Baby FaceTime: Can toddlers learn from online video chat? *Developmental Science, 20*(4), Article e12430.
Noles, N. S., Danovitch, J., & Shafto, P. (2015). Children's Trust in Technological and Human Informants. *In CogSci.*, 1721–1726.
Peebles, A., Bonus, J. A., & Mares, M. L. (2018). Questions+ answers+ agency: Interactive touchscreens and Children's learning from a socio-emotional TV story. *Computers in Human Behavior, 85*, 339–348.
Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? *Developmental Science, 14*(6), 1417–1430.
Read, J. C., & MacFarlane, S. (2006, June). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on interaction design and children* (pp. 81–88).
Reese, E., Leyva, D., Sparks, A., & Grolnick, W. (2010). Maternal elaborative reminiscing increases low-income children's narrative skills relative to dialogic reading. *Early Education and Development, 21*(3), 318–342.
Richards, M. N., & Calvert, S. L. (2017). Measuring young US children's parasocial relationships: Toward the creation of a child self-report survey. *Journal of Children and Media, 11*(2), 229–240.
Richert, R. A., Robb, M. B., & Smith, E. I. (2011). Media as social partners: The social nature of young children's learning from screen media. *Child Development, 82*(1), 82–95.
Rideout, V., & Robb, M. B. (2020). *The Common Sense census: Media use by kids age zero to eight, 2020*. San Francisco, CA: Common Sense Media.
Rochat, P. R. (2001). Social contingency detection and infant development. *Bulletin of the Menninger Clinic, 65*(3: Special issue), 347–360.
Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child Development, 85*(3), 956–970.
Sabharwal, N., & Agrawal, A. (2020). Introduction to Google dialogflow. In *Cognitive virtual assistants using Google Dialogflow* (pp. 13–54). Berkeley, CA: Apress.
Silander, M., Grindal, T., Hupert, N., Garcia, E., Anderson, K., Vahey, P., & Pasnik, S. (2018). *What parents talk about when they talk about learning: A national survey about young children and science*. Education Development Center, Inc.
Strouse, G. A., O'Doherty, K., & Troseth, G. L. (2013). Effective coviewing: Preschoolers' learning from video after a dialogic questioning intervention. *Developmental Psychology, 49*(12), 2368.
Strouse, G. A., & Samson, J. E. (2021). Learning from video: A meta-analysis of the video deficit in children ages 0 to 6 Years. *Child Development, 92*(1), e20–e38.
Strouse, G. A., Troseth, G. L., O'Doherty, K. D., & Saylor, M. M. (2018). Co-viewing supports toddlers' word learning from contingent and noncontingent video. *Journal of Experimental Child Psychology, 166*, 310–326.
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.
Tu, T. (2006). Preschool science environment: What is available in a preschool classroom? *Early Childhood Education Journal, 33*(4), 245–251.
Wang, M. H. (2014). *Parental scaffolding behaviours during co-viewing of television with their preschool children in Taiwan* (Doctoral dissertation. Institute of Education, University of London).
Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education, 7*(2), 104–112.
Weisberg, D. S., Hirsh-Pasek, K., Golinkoff, R. M., Kittredge, A. K., & Klahr, D. (2016). Guided play: Principles and practices. *Current Directions in Psychological Science, 25*(3), 177–182.
Willoughby, D., Evans, M. A., & Nowak, S. (2015). Do ABC eBooks boost engagement and learning in preschoolers? An experimental study comparing eBooks with paper ABC and storybook controls. *Computers & Education, 82*, 107–117.
Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y. S., & Warschauer, M. (2021). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development*. https://doi.org/10.1111/cdev.13708
Xu, Y., Branham, S., Collins, P., Deng, X., & Warschauer, M. (2021). Are current voice interfaces designed to support children's language development?. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3411764.3445271
Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers in Education, 161*, Article 104059. https://doi.org/10.1016/j.compedu.2020
Xu, Y., & Warschauer, M. (2020a). Exploring young children's engagement in joint reading with a conversational agent. In *Proceedings of the 19th ACM international conference on interaction design and children (IDC '20)*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3392063.3394417.
Xu, Y., & Warschauer, M. (2020b). What are you talking to?: Understanding children's perceptions of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. April 25-30, 2020*. Honolulu, HI: ACM. doi: 10.1145/3313831.3376416.
Xu, Y., & Warschauer, M. (2020c). " Elinor is talking to me on the screen!" integrating conversational agents into children's television programming. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–8).
Xu, Y., Yau, J. C., & Reich, S. M. (2021). Press, swipe and read: Do interactive features facilitate engagement and learning with e-Books? *Journal of Computer Assisted Learning, 37*(1), 212–225.
Yarosh, S., Thompson, S., Watson, K., Chase, A., Senthilkumar, A., Yuan, Y., & Brush, A. B. (2018, June). Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (pp. 300–312).
Zhou, N., & Yadav, A. (2017). Effects of multimedia story reading and questioning on preschoolers' vocabulary learning, story comprehension and reading engagement. *Educational Technology Research and Development, 65*(6), 1523–1545.