# Artificial Intelligence Enhances Children's Science Learning from Television Shows

Ying Xu[1], Kunlei He[2], Julian Levine[2], Daniel Ritchie[2], Zexuan Pan[1], Andres Bustamante[2], and Mark Warschauer[2]

[1]University of Michigan, Ann Arbor

[2]University of California, Irvine

**Manuscript under review by *Journal of Educational Psychology***

## Author Note

The authors declare no conflict of interests.

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.

Correspondence concerning this article should be addressed to Ying Xu, Marsal Family School of Education, University of Michigan, Ann Arbor, MI, 48104. E-mail: yxying@umich.edu

# Abstract

Educational television programs are important learning resources for young children, especially those from under-resourced households. These programs' potential can be amplified if children are given the opportunity to meaningfully interact with media characters during their video watching. In this project, we partnered with PBS KIDS to develop interactive science-focused videos in which the main character, powered by artificial intelligence (AI), engaged in dialogic interactions with children by asking them questions and providing responsive feedback. The children who watched the interactive videos performed better on a science post-test than children who watched the broadcast version of the video (without any interaction) or a pseudo-interactive version (in which the media character asks children the same questions and gives generic feedback after a fixed amount of time). The AI character's responsiveness positively influenced both the quantity and quality of children's verbalizations during video watching, compared to the pseudo-interactive condition. This paper sheds light on the feasibility and effectiveness of using conversational technologies to support active learning in children through educational programs.

*Keywords:* artificial intelligence, video watching, dialogue, science learning

## Educational Impact and Implications Statement

Television and video programs are important educational resources for young children. Our research indicates that conversational artificial intelligence can enhance the interactivity of these shows by allowing children to dialogue with the media character, thus enhancing children's learning. This underscores the significance of responsive interactivity in media, suggesting that future educational video resources could benefit from integrating conversational technologies to promote active science learning through television shows.

## Artificial Intelligence Enhances Children's Science Learning from Television Shows

Television or video watching has become a staple of modern childhood. A national study by Rideout and Robb (2020) revealed that U.S. children aged three to six watch an average of 1.5 hours of video content daily, which includes educational programs intended to promote children's school-readiness skills. High-quality educational programs can provide valuable learning experiences to a large audience, particularly those with limited access to other educational resources (Choi, 2021). Recognizing the potential of these programs, substantial research has been carried out towards understanding the optimal design features and strategies that enhance children's learning experiences (for a review, see Richert et al., 2011).

Television programs have traditionally been a one-way transmission, which does not inherently offer opportunities for children to actively interact with the educational content as interactive apps do (Jing & Kirkorian, 2020a; Strouse & Samson, 2021). However, this potential limitation can be mitigated when children watch these programs alongside a knowledgeable partner, such as a parent, who engages them in dialogic interaction. This dialogic interaction—characterized by asking questions and providing responsive feedback based on the children's responses—has been shown to effectively improve children's learning from video content (Anderson & Hanson, 2017; Ewin et al., 2021; Richert et al., 2011; St. Peters et al., 1991). However, such interactions often depend on the caregiver's availability, which may not always align with children's screen time (Haines et al., 2013; Stevens & Takeuchi, 2011; Strouse et al., 2013)

An alternative approach for providing dialogic interactions may be through embedding questions within the video content. Many television programs, such as Dora the Explorer, have adopted a "pseudo-interactive" design, where media characters prompt children to answer questions, pause for a set duration, and then offer a generic response. However, this form of pseudo-interaction may not be optimal, as it lacks the capability to offer responsive feedback to children. In fact, several studies using a “Wizard of Oz” technique— where researchers manipulate media characters’ responses to individual children — have demonstrated that the most effective learning outcomes are achieved when a character combines questioning with responsive feedback, outperforming approaches that involve only questioning or do not include questioning at all (Calvert et al., 2020; Peebles et al., 2018).

With the advances in artificial intelligence (AI), dialogic interactions between child viewers and media characters in video programming has now become possible. Supported by speech recognition technologies and natural language processing, young children already frequently talk to AI-enabled devices or toys available in their homes—such as smart speakers, social robots, or internet-connected stuffed animals (Cassell, 2022; Druga et al., 2018; Lovato et al., 2019)—and many of these applications were designed with specific learning or developmental goals in mind. The research has found that, although conversational AI is not yet fully optimized for conducting free-form, creative conversations with children, it is capable of engaging in structured dialogue with clear learning objective, typically by asking children questions, listening in, and providing responsive feedback (for a review, see Xu, 2023). This possibility has inspired our research to examine the feasibility and effectiveness of integrating this kind of conversational AI into children’s educational programs to enhance learning through dialogic interactivity.

In this preregistered article, we partnered with PBS KIDS, the largest public media services provider in the U.S., to incorporate conversational AI into a science show, Elinor Wonders Why. In the resulting interactive videos, the main character, Elinor, asks children questions and provides responsive feedback as children watch the episodes. We examined the effect of interactive videos on children's science learning and verbal engagement. To this end, we employed a between-subjects randomized design to compare the outcomes of children who watched interactive videos with those who viewed the same episodes in two alternative formats: the standard non-interactive version, as broadcast on PBS KIDS, and a pseudo-interactive version, where Elinor poses questions, pauses, and offers generic responses similar to Dora the Explorer. The findings of this study could potentially extend our understanding of how dialogic interactivity with AI can support learning within the context of educational programs, as well as offer practical implications for leveraging AI to enhance children's learning experiences through these programs.

## Literature Review

### Children's Learning From Educational Video Programs

Educational video programs are important learning resources for children outside of school. These programs are designed to support children's learning in domains encompassed in school readiness, such as literacy, numeracy, science, socioemotional skills. A well-known initiative that provides funding for the development of such educational programs for children aged two to eight is the U.S. Department of Education's Ready to Learn (RTL) initiative (Wartella, et al., 2016) which has been awarded several millions of dollars annually. This initiative has led to the production of numerous well-received series, including "Elinor Wonders Why," an animated science series that is the focus of this study.

Educational programs are typically designed with specific curriculum objectives in mind. For instance, the "Elinor Wonders Why" series is grounded in the PBS KIDS Science Learning

Framework (Christensen et al., 2019) informed by the Next Generation Science Standards (NGSS; National Research Council, 2013). Therefore, this series is guided by a two-fold learning objective: to support children's understanding of scientific concepts and to facilitate their application of these concepts in problem-solving across various contexts. Furthermore, "Elinor Wonders Why," like many other educational programs, embeds these curriculum objectives in narrative storylines, with the goal of making the content more appealing to children as well as to leverage the narrative structure to highlight and make educational concepts, such as cause and effect, more prominent and concrete (Fisch, 2000; Jing & Kirkorian, 2020b).

Generally, research indicates that educational programs can be effective in enhancing children's knowledge and skills in areas aligned with the intended learning objectives. For example, Hurwitz (2019) conducted a meta-analysis to examine the effects of high-quality educational programs on children's literacy outcomes, which drew upon data from 45 independent interventions, each involving a randomized control trial where half of the children were given access to the program and the other half continued with regular activities or received alternative lessons. This analysis revealed that the positive impact of these programs is equivalent to approximately one-and-a-half months of additional literacy learning beyond the typical growth. This finding was corroborated by another, more recent meta-analysis, which suggested a small-to-medium effect ($r = .30$) of watching television on children's vocabulary learning (Jing et al., 2023). In addition to these two meta-analyses focused on literacy and language development, a robust body of studies has also documented positive learning benefits resulting from watching educational programs designed to promote a broad range of subject domains, such as math (e.g., Kostyrka-Allchorne et al., 2017; McCarthy et al., 2018) and science (e.g., Bonus et al., 2023; Hsueh et al., 2017).

**Dialogue Interactivity as Scaffolding of Children's Learning from Videos**

Researchers have long explored how learning occurs through educational television and how to better support children in this process. Theories have been established to suggest that dialogic interactivity—where children engage in *question* answering and receiving responsive *feedback or scaffolding*—is effective in improving their learning from educational programs (Mayer, 2002; 2003). This approach is grounded in the sociocultural theories that posit learning is a socially-mediated process in which children acquire knowledge and skills through guided interactions with a more knowledgeable partners. Such external scaffolding expands what a child can comprehend on their own to a deeper level of potential level of development, termed as "the zone of proximal development" as conceptualized by Vygotsky (1978, *p*.86). Over time, children internalize this learning, which enables them to progress to subsequent levels of development.

In the context of dialogic interactivity, questioning primes children to selectively attend to relevant information presented, thereby likely enhancing their retention of this information. While this type of priming can support children's comprehension of information presented across media formats (Kendeou et al., 2008), it might uniquely support comprehension of video programs, as combination of linguistic and visual input may present a challenge for children to effectively select the most relevant information and integrate these information sources to form a coherent understanding (Kendeou et al., 2020). Moreover, questioning also activates children's pre-existing knowledge relevant to the topic and facilitates the integration of new information with this existing knowledge base. This process supports children to make inferences beyond what is presented and apply what they learn from videos in various contexts.

The benefits of questioning are amplified by responsive feedback, which typically includes hints and explanations (Graesser et al., 2009). Based on a classic model of feedback processing, while questioning elicits children's current state of understanding, the feedback received by children prompted them to re-evaluate their initial responses, facilitated their identification of errors, and subsequently led them to adjust their responses to align with the expected direction (Bangert-Drowns et al., 1991). A systematic analysis of the effects of responsive feedback for children aged three to eleven years confirmed overwhelmingly positive effects of feedback on improving children's task-level performance, including problem-solving tasks (Fyfe et al., 2023). The principles underlying this interactivity, which support general comprehension, can also be extended to domain-specific learning, such as in science. For instance, studies demonstrated that interactive strategies used in a classroom instruction or interventions not only enhanced students' language comprehension, but also students' content area knowledge in science (France, 2021; Kim et al., 2021; Kim et al., 2023).

Empirical evidence in the context of television watching robustly supports the theoretical framework outlined above. These studies compared children's comprehension of television programs with or without the availability of dialogic interactivity, either coming from a live co-viewing partner (e.g., Roseberry et al., 2014; Strouse et al., 2013), usually parents, or coming from the questions and feedback embedded in the media content themselves (e.g., Calvert et al., 2020; Xu et al., 2022; Yang et al., 2022). Findings from these studies indicate that the combination of questioning and responsive feedback yields better learning outcomes, surpassing those with only questioning or no questioning at all. For instance, Strouse and colleagues (2013) developed an intervention in which parents were trained to ask preschool-aged children questions and provide feedback while watching word-learning videos. The study found that children learned most effectively when parents' questions were followed by responsive feedback that includes corrections to their interpretations. This approach was more effective than when parents asked the same questions but provided generic feedback that did not correct misconceptions or encourage further elaboration. Notably, both these groups outperformed a third group---in which children watched the videos without any parental dialogue---in terms of improvement in their word learning. The benefits of parent questioning and feedback were replicated by another study using a "Wizard of Oz" technique wherein an experimenter controlled the responses of the animated character. Calvert et al. (2020) found that preschool-aged children's math learning outcomes from a video were significantly better when the video's main character, Dora, asked children questions and replied in a timely and responsive manner as compared to just asking questions without providing corrective feedback. Across these studies, there is converging evidence that children's learning from educational programs can be enhanced by dialogic interactivity, whether it comes from a human or a media character, which lays the groundwork for the present study to investigate the potential of artificial intelligence in supporting such interactive learning experiences.

In addition, for bilingual children, it is crucial to consider the relationships between the two languages during dialogic interactions. Cummins (1979) posits that skills, knowledge, and metalinguistic awareness acquired in one language can facilitate the acquisition and understanding of another language. This theoretical framework suggests that the cognitive and linguistic skills developed in a child's home language (e.g., Spanish) are not isolated but are transferable to another language (e.g., English), and there are common underlying proficiencies that might exist, thereby enhancing learning outcomes in bilingual or multilingual contexts. Indeed, empirical evidence also suggests that children with stronger Spanish reading proficiency predict their later growth in English reading comprehension, in addition to their initial English oral proficiency (e.g., Relyea & Amendum, 2020). Thus, this highlights the need to consider children's proficiencies in both their home language and English as pivotal factors to account for in studies incorporating dialogic interactivity to support children’s learning, as these factors might together contribute to how much children can learn in these programs.

### Children's Verbal Engagement During Dialogic Interactivity

A crucial mechanism in dialogic interactivity that enhances learning is children’s verbal engagement (Guthrie & Klauda, 2013). Research has suggested that dialogic interactions, either with caregivers or embedded in the media, can enhance both the quantity and quality of children’s verbal engagement; such enhanced engagement is positively correlated with learning outcomes (for a review, see Lepola, 2023). In terms of quantity, relevant metrics include the percentage of questions to which children respond and the length of their responses. For instance, Troseth et al. (2020) found that, when parents were prompted to utilize dialogic questioning strategies, children were more cognitively engaged as they talked more and used more diverse vocabulary. Similarly, Zhou and Yadav (2017) found that children who were asked questions during story book reading generated more comments relevant to the stories than those who listened to the story without questioning. The increased quantity of children’s verbal response was found to positively correlate with children’s learning outcomes (Calvert et al., 2020; Yang et al., 2022). In terms of quality, metrics include the relevance of children’s responses to the topic and their accuracy. For instance, Troseth et al. (2020) found that when an electronic book was designed with embedded questions, children’s comments were more likely to be on-topic compared to the spontaneous comments made by another group of children who read the same book without the questions. Another, Xu et al. (2021), observed that during dialogic reading sessions with a digital voice assistant, preschool-aged children who provided higher-quality responses—characterized by topical relevance and accuracy—demonstrated better story comprehension.

Furthermore, there is robust evidence suggesting that children's verbal engagement is influenced by the responsiveness of conversation partners, namely the ability of parents or teachers to provide *feedback* that is responsive and adaptive to children's responses or even non-responses after a question is asked (for a review, see Rowe & Snow, 2020). For example, a study conducted in a preschool dramatic play setting found that the more a teacher is able to continue and extend children's contributions, the more frequently children respond, leading to children generating more utterances that are both directly relevant and tangential to the play (Meacham et al., 2016). Similar findings were reported in the context of children's interactions with cartoon characters during video watching. Carter and colleagues (2017) conducted an experiment in which one group of children watched the original "Mickey Mouse Clubhouse," where Mickey asked the children simple questions and waited for 10 seconds before praising them, regardless of their responses. This was compared to a Wizard-of-Oz condition in which Mickey's praises were secretly initiated by an experimenter immediately after a child finished their response. The study found that children answered only 35% of Mickey's questions in the original condition, whereas the response rate went up to 73% if Mickey responded with the right timing. Overall, these studies suggest that children are aware of the conversational partner's ability to be responsive, which could impact their communicative behaviors.

## Artificial Intelligence for Personalization of Learning

The rapid advancement in natural language processing and speech technologies has enabled AI to simulate the learning benefits of dialogic interactivity. AI could provide additional opportunities of dialogue with children, especially when a human conversation partner is unavailable. Furthermore, the implementation of automatic speech recognition in many current AI systems enables children to use speech, rather than text, to engage in the dialogue. This feature is particularly beneficial for young children who may not have fully developed literacy or fine motor skills, making it difficult for them to navigate digital media through conventional methods like typing or tapping.

Building on these technological advancements, studies have begun incorporating strategies that are grounded in human-to-human dialogic interactivity into intelligent systems, specifically focusing on questioning and feedback. A classic example for this line of work was the AutoTutor, a pedagogical agent, initially developed in the 1990s, that holds conversations with students in natural language and "simulates the dialogue moves of human tutors as well as ideal pedagogical strategies" (p. 124, Graesser et al., 2017). The specific dialogue moves employed by AutoTutor involve engaging students with questions, followed by supportive strategies which include prompts to encourage students to elaborate, hints to simplify questions for easier understanding, and explanations that help students calibrate their responses towards an anticipated answer (D'Mello & Graesser, 2023). Other more recent studies also utilized a similar principle. For instance, Ruan et al. (2020) developed a system with a chat function that guides elementary school students in learning about math problem-solving. In this system, the AI agent prompts students to brainstorm potential solutions and then uses scaffolding to guide them towards solving problems more effectively. Other studies in the area of language comprehension have also suggested the effectiveness of AI-assisted questioning and feedback in improving students' comprehension. (McNamara et al., 2023; McCarthy et al., 2020)

Studies have found that providing students with opportunities to interact with these intelligent tutors enhances children's learning (for reviews, see Kulik & Fletcher, 2016; Ma et al., 2014; Steenbergen-Hu & Cooper, 2013). In some studies, the benefits derived from the intelligent tutor are comparable to those provided by tutoring from a trained human tutor (e.g., VanLehn, 2011; Wang et al., 2023; Ward et al., 2013; Xu et al., 2022). For example, Xu and colleagues developed a system where children aged three to six received comprehension questions and responsive feedback during storybook reading, either from an intelligent tutor or a human reading partner (Xu et al., 2022). Their study found that children's comprehension of the story improved regardless of whether the conversation was with an AI or a human, compared to those children who did not receive any questioning during story listening.

The use of AI to simulate dialogic interactivity between human tutors and children is largely enabled by AI's capability to recognize and interpret children's responses in natural language. Traditionally, interactivity in digital learning systems has been confined to button-clicking, where children answer multiple-choice questions and receive feedback on the correctness of their specific selections. This approach, while offering added benefits (e.g., McMaster et al., 2023), also misses opportunities to use open-ended questions where children can articulate their own thought processes and receive more targeted scaffolding (Hamel et al., 2021). Given the technological advances, modern intelligent tutoring systems can better understand the variety of children's responses, predict correct answers, and identify common misconceptions. Specifically, AI tutors were trained to compare children's responses against expected answers and known misconceptions, which mirrors the approach of human tutors

(D’Mello & Graesser, 2023). AI tutors then provide feedback aimed at steering students towards the correct answers through a series of dialogic interactions.

It is important to note that the majority of previous research on intelligent tutors has been concentrated on more structured lessons aimed at older students from upper elementary school and beyond, where the tutor’s objective is to facilitate students in completing tasks similar to assessments (for reviews, see Dai, et al., 2022; Feng et al., 2021). In such cases, dialogic interactivity may be more conceptual, focusing on the abstract principles of the subjects. In contrast, our study introduces AI-assisted dialogue within the framework of narrative-based educational programs, which situates the discussion of scientific concepts within a more immediate and relatable context. This approach may be particularly suited to the younger demographic our research aims to engage.

## The Current Study

This study aims to examine the impact of dialogic interactions with AI-powered media characters in video watching on children’s learning and engagement. To this end, we developed “interactive videos” based on an animated science show. These conversational videos leveraged AI to allow children to answer questions asked by the show’s main character and receive responsive feedback. We conducted a randomized controlled trial in which children from four to seven years old were assigned to watch two science episodes in three different formats: interactive, pseudo-interactive, and non-interactive. The interactive format incorporated both key components of dialogic interactivity: questioning and responsive feedback. The pseudo-interactive format retained the same questioning component, but the feedback was generic. The non-interactive format lacked both of these components. Almost all children participating in the study were self-identified as Hispanic, an ethnic group that spends a significant amount of time consuming video content (Rideout & Robb, 2020) and values educational opportunities afforded by video programs (Kalinowski et al., 2020).

Our study was guided by the following two research questions:

RQ1: Does watching interactive videos and having dialogue with the media character improve children's science learning outcomes, as compared to watching videos in a pseudo-interactive or non-interactive format?

H1: Given the documented benefits of dialogic interactivity, we hypothesize that children who watch interactive videos featuring both questioning and responsive feedback will perform better than those who watch pseudo-interactive videos given the lack of responsiveness in feedback. This performance is expected to surpass that of children who watch non-interactive videos, which lack any form of questioning or feedback. Furthermore, we hypothesize that interactive videos enhance learning by increasing not only the recall of information but also the ability to transfer learning to other scenarios.

RQ2: Does watching interactive videos increase children's verbal engagement, as measured by children's quality and quantity of their responses to the character's questions, as compared to watching the pseudo-interactive format?

H2: Drawing on literature regarding the impact of conversational partner responsiveness on children's verbal engagement, we hypothesized that children would be more likely to respond to the character's questions in the interactive condition, thus resulting in a greater quantity of responses in terms of both rate and length. Additionally, we hypothesized that the responsive feedback provided by interactive videos would offer scaffolding to children, thus increasing the quality of the responses in terms of topic relevance and accuracy.

RQ3: What is the correlation between science learning outcomes and verbal engagement across the interactive and pseudo-interactive conditions?

H3: Given that verbal engagement has been identified as a key mechanism for improving learning outcomes, we hypothesized that across both conditions, children's verbal engagement, both quantity and quality, would be positively correlated with children's science learning outcomes.

# Method

## Participants

A total of 275 children, aged four to seven, were recruited from a charter school with a predominantly Latine population in the Western USA. Out of these, 29 began the study but did not finish all sessions, hence they were excluded from our analysis. Eight of these children were recruited but never tested, so they were never allocated to an experimental condition. The remaining 21 children discontinued after beginning at least one study session. Of these, 7 were in the interactive condition, 9 in the pseudo-interactive condition, and 5 in the non-interactive condition. The primary reasons for discontinuation were the school's summer closure and absences due to illness. Thus, our final sample included 246 children who completed all study sessions. The average age of these children was 6.21 years (min = 3.58 years, max = 7.66 years), and 95.47% of the sample were self-identified as Hispanic or Latine. Table 1 presents the demographics of the children included in this study.

The randomization followed the established procedure where each recruited individual was assigned to a condition using a random number generator, with each number corresponding to one experimental condition (Kang et al., 2008). Among the 246 children, 81 watched the interactive video, 79 watched the pseudo-interactive video, and the remaining 86 watched the non-interactive version. A balance check confirmed that baseline characteristics were equivalent across the three groups (see Table 1).

## The Interactive Videos

The development of the interactive videos was based on a PBS KIDS science show, “Elinor Wonders Why,” targeting children of preschool and early elementary age. Each 10-minute episode typically begins with Elinor encountering a science problem, and her discoveries unfold within a storyline that includes other characters, events, and settings. This experiment utilized interactive versions of two episodes of “Elinor Wonders Why” of approximately 11 minutes each. The first episode demonstrates the principles of aerodynamics through Elinor experimenting with ways to build a fast cardboard car to win a race. The second episode focuses on the phenomenon of reptile molting as Elinor observes how and why snakes shed their skin.

Elinor’s dialogic interactions with children in these interactive videos resemble the structures of the standard questioning and responsive feedback that we reviewed above. The questioning and feedback were designed to fall within a child’s zone of proximal development (Vygotsky, 1978), and were developed based on the Next Generation Science Standards (NGSS; National Research Council, 2013) and vetted by PBS KIDS’ science learning advisory board. There were approximately 10 conversational moments throughout each 11-minute episode. The basic structure of the dialogue is depicted in Figure 1.

In the remainder of this section, we illustrate how Elinor posed questions and provided responsive feedback using one of the conversational moments as an example. This conversation moment occurred during the episode on aerodynamics, after Elinor experimented with different ways to improve her cardboard car so that it could go just as fast as her friend's race car. She engaged the children by asking them to make observations, "We made a really fast car! It's so fun to try different things and figure out what works and what doesn't. So, how did we make our car go faster?"

Once the child's response came in, the AI first used automatic speech recognition to transcribe children's spoken words into text, and then used natural language processing to categorize the children's response into one of several predefined answer categories. These categories were developed both deductively and inductively based on children's actual responses to the question during pilot testing. For example, the answer categories for this question included change in *shape*, *color*, or *engine*.

Elinor then provided feedback based on the categories of children's answers. Specifically, in this context, the changing shape was considered the correct answer, as this was the only experiment Elinor conducted in the episode that resulted in the car with a streamlined shape going faster. When a child correctly identified changing the car's shape as the key factor, Elinor acknowledged their answer and elaborated to deepen their understanding, explaining, "Exactly! We changed the shape of our car to make it streamlined—pointy in the front and round in the middle, which makes it go faster."

However, if a child's answer to this initial question was incorrect, indicated uncertainty (e.g., "I don't know"), Elinor asked a follow-up questions that employed one or two scaffolding strategies—offering multiple options and contextual clues—to guide the children towards the correct answer. The "multiple options" strategies involve simplifying the open-ended question by presenting several choices for the children to select from. For instance, in Elinor's follow-up question reminded children of the various changes made to the car during the episode with, "We tried changing the color, adding a cupholder, and making our car's shape streamlined. Which of these changes made the car go faster?" Elinor's follow-up questions might also contain contextual clues. For example, when questioning why a snake shed its skin and a child struggled to answer correctly, Elinor referenced an analogous scenario in the episode, where one of Elinor's friends felt tight wearing a shirt that fit when they were younger. Both scaffolding strategies (i.e., providing multiple options and contextual clues) offered children a second opportunity to reconstruct their responses. If the child provided a correct response to this follow-up question, Elinor would affirm the answer, further reinforcing the concept. If the child still did not provide the correct answer, Elinor would then explain the correct answer to model how a more knowledgeable peer might approach the question.

At times, children might opt to remain silent instead of verbally responding to a question. In such instances, Elinor employs a gentle and encouraging approach to motivate the child to share their thoughts. She uses phrases like, "I'm really curious about what you think," "It's fun to make observations together," and "I wonder what you think." These expressions are designed not only to create a supportive experience but also to signal to the child that thinking together is a shared and enjoyable process. However, if a child continues to remain silent after this encouragement, Elinor would explain the question further and then move on, ensuring that the child does not become frustrated while maintaining the flow of the episode watching.

Additional examples of how Elinor handled situations when children provided incorrect or off-topic responses, or did not respond, are shown in Appendix A.

**Performance of the Interactive Videos**. Based on data collected from this study, the AI-powered interactive character, Elinor, demonstrated satisfactory performance in deciphering children's speech and accurately classifying children's responses into the predefined answer categories. To assess the agent's performance, we compared its classification of each response's category with that of a trained human researcher. The agent's classification of response categories was automatically logged during the experiment and later retrieved for this analysis. A trained human researcher, blinded to the agent's classifications, independently read children's responses and determined the category for each response. We then compared these two sets of classifications—one by the AI agent and the other by the human researcher—by calculating their inter-rater reliability. The inter-rater reliability, measured using Cohen's Kappa, was calculated at 0.88, indicating excellent agreement (McHugh, 2012). This performance did not vary significantly depending on children's age ($r = .11$, $p = .37$), baseline English ($r = .03$, $p = .82$) and Spanish proficiency ($r = .03$, $p = .47$), or baseline science knowledge ($r = .11$, $p = .37$). This level of accuracy was consistent with other studies involving children in a similar age range (Dietz et al., 2021; Xu & Warschauer, 2020).

## Study Design and Procedure

The study was divided into four sessions. The first two sessions involved the assessment of baseline characteristics, and in the last two sessions, children watched an episode in each session and then completed posttests.

In the first session, children were assessed on their baseline science knowledge using an adaptive computer-based program, which lasted 30 minutes (Lens on Science, Zucker et al., 2016). This measure was included as children's prior knowledge of the target educational content is recognized as an important factor predicting their learning and engagement with educational programs. Children with a deeper understanding of a subject are better able to utilize their knowledge base to draw inferences and concentrate on relevant information (Fisch, 2000).

In the second session, children’s oral language abilities were measured using a bilingual language assessment, the Bilingual English Spanish Oral Screener (Bilingual English Spanish Oral Screener [BESOS], Peña et al., 2010). This baseline measure was included as language ability plays a key role in learning from educational programs, which predominantly deliver information verbally (Kendeou, 2020). We assessed proficiency in both languages, as all participants in our study were bilingual in English and Spanish. It is theorized that children's general comprehension skills acquired in their home language (Spanish) can transfer to English (Cummins, 1979). Nevertheless, baseline English proficiency is predicted to remain a more significant predictor of learning outcomes, particularly because the educational content was delivered in English.

During the third and fourth sessions, children watched one episode, respectively, in their designated format. Immediately after watching the episode, they completed a post-test to assess their learning from that episode. The order of the two episodes was counterbalanced across the three conditions. Each of these two sessions lasted about 40 minutes. An experimenter was present in the room during the episode watching; they were instructed to sit in the corner of the room and not in sight of children when they faced the screen. The experimenter did not intervene with the video watching unless a child left the seat, which happened among six participants. In this case, the experimenter gently reminded the child that the episode was continuing and asked them to sit down, which was effective and sufficient to bring all children back on task.

## Measures

### *Baseline Measures*

Our study collected two baseline measures that were suggested in the previous literature as relevant to children’s learning from educational programs and from dialogic interactivity.

**Science Knowledge**. To capture children’s prior science knowledge, we used Lens on Science (Greenfield & Penfield, 2013), a computer-adaptive assessment based on item-response theory (IRT) developed according to the National Research Council’s Framework for K-12 Science Education (National Research Council, 2012). Lens on Science draws from a large item bank to assess children’s knowledge and skills across three science areas: life science, earth/space science, and physical/energy science. Typically, 35-40 items were administered in total. Rasch reliability for the instrument, based on a sample of 1,753 students, was estimated at 0.87 (Greenfield et al., 2015).

**Language Proficiency**. We administered the Bilingual English-Spanish Oral Screener (BESOS) to assess the language proficiency of children in both English and Spanish. The BESOS comprises four subtests of Spanish semantics, English semantics, Spanish morphosyntax, and English morphosyntax. Each test contains 10–18 items that are sensitive to development. According to this assessment, two separate scores were computed: one for the child’s proficiency in English and another for their proficiency in Spanish. Based on normative data from a sample of 349 typically-developing first graders (Bedore et al., 2023), preliminary Cronbach’s alpha reliability estimates for the four subtests are .816 for Spanish semantics, .648 for English semantics, .890 for Spanish morphosyntax, and .840 for English morphosyntax, respectively.

### *Science Learning Outcomes*

We developed questionnaires to assess children's learning from the episodes. The questions, different from the conversational prompts Elinor asked during the episode, were also aligned with the NGSS (National Research Council, 2013) and the US Department of Education's Ready to Learn Science Framework (Wartella, et al., 2016). The assessments consisted of a total of 23 items, which were aimed to capture children's ability to recall the educational content introduced in the episodes as well as their ability to transfer the learned concepts to new scenarios. For the aerodynamics episode, the assessment consisted of a total of 9 items, divided into 3 items focused on recall and 6 items on transfer. Similarly, the assessment for the reptile molting episode included 16 items, with 5 dedicated to assessing recall and 11 for transfer.

In terms of the format and scoring of the items, for 16 out of the total of 23 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 seven other questions, children were asked to provide explanations of their answers. We scored their explanations from 0 to 2 points. A score of 0 indicated a completely incorrect answer; a score of 0.5 indicated an answer was incorrect but contained some correct ideas related to the episode; a score of 1 indicated an answer was largely correct but some language was inaccurate; a score of 1.5 indicated a correct answer with accurate language; and a score of 2 indicated a correct answer with additional accurate information supporting the answer. The last question was picture identification with a maximum score of 1 point. All questions were read aloud to children, and children responded verbally. We calculated a total score by summing the points across all items, with the maximum possible score being 45 points. For all questions, we treated answers equally, focusing on the accuracy of the content. It is important to note that responses in Spanish were very rare.

The scoring was conducted by three trained researchers by reviewing video recordings of the posttest sessions. The inter-rater reliability, calculated by intra-class correlation (ICC) given the numeric nature of the score, was 0.92 on average across all items, indicating a strong level of agreement. The Cronbach's alpha internal consistency of this assessment was 0.83. The assessment score was highly correlated with children's age ($r = .63, p < .001$), baseline science knowledge ($r = .76, p < .001$), and English proficiency ($r = .71, p < .001$), suggesting excellent external validity.

### ***Verbal Engagement***

Children's responses to each of the questions from the media character were analyzed to assess their verbal engagement. For children assigned to the interactive and pseudo-interactive conditions, verbal responses to Elinor's questions during the episode viewing were transcribed by research assistants. Note that while children's responses in the interactive videos were initially transcribed by AI during the testing sessions, we relied on transcriptions performed by research assistants for our analysis to ensure accuracy. These responses were then coded for quantity and quality based on metrics used by Xu et al. (2021) and Westerveld & Roberts (2017). The unit of coding was each child's response. It is important to note that a significant majority of the responses we received were in English, accounting for 98.94% of the total. Spanish responses made up a small fraction, only 0.95%, and an even smaller percentage, 0.01%, consisted of answers that mixed both languages. Responses in both English and Spanish were evaluated on the same basis; for instance, we considered responses that were relevant and correct in either language as relevant and accurate for our study.

**Response Quantity**. In terms of quantity, we coded **response rate** and **response length**. For the response rate, we used a binary coding scheme whereby a 0 was assigned when verbal speech input was present, and a 1 was assigned for non-responses or responses that included only filler words (e.g., *um*). For response length, the total number of words was counted in each response, including repetitive words. Filler words, however, were excluded from the word count. Two coders conducted the coding, achieving a 100% agreement rate for these two metrics.

**Response Quality**. To assess quality of verbal engagement, we coded **relevance** and **accuracy**. Relevance was scored on a 3-point scale, with a score of 0 denoting an irrelevant response, 1 a response relevant to episode content but not to the pertinent question, and 2 a response directly relevant to the question. Accuracy was scored dichotomously with 0 for inaccurate responses and 1 for accurate responses. Each child response was coded according to this scheme by two research assistants who were provided with a detailed scoring rubric and completed a training set for each episode before scoring the remaining transcripts. The Cohen’s kappa inter-rater reliability between the two coders was 0.91 for scoring relevance and 0.95 for scoring accuracy. Discrepancies in the scoring were then resolved by Authors 2 and 3. The two authors reviewed each response that was coded differently by the two coders. They referenced the scoring rubrics for guidance and engaged in detailed discussions about the nature of the discrepancies. This process continued until they reached a consensus on the appropriate coding for each response.

Children in the interactive condition, as depicted in Figure 1, may have received follow-up questions if they answered the initial questions incorrectly. Thus, when analyzing children's response relevance and accuracy, we differentiated between children's initial and follow-up responses in the interactive condition.

In our analysis of response quality (i.e., relevance and accuracy), we faced a decision on how to handle instances where children provided no verbal response. Opting to maintain the integrity of our randomized group assignments and better reflect children's performance in answering Elinor's questions, we chose not to exclude these non-responses. Instead, when children did not provide a verbal response, we coded them as 0 for both relevance and accuracy. We have also provided the descriptive statistics for an analysis that excludes non-responses in the Appendix B, offering insight into the quality of children's responses when they chose to respond.

## Analytic Approach

The first and second research question aimed to estimate the effects of video watching formats (i.e., interactive, pseudo-interactive, non-interactive) on children's science learning and verbal engagement outcomes. To achieve this, regression analyses were utilized as the primary statistical method. Data were analyzed with each child treated as an individual unit of analysis. The key independent variable of interest was the experimental condition assigned to each child. Specifically, the reference group consisted of children assigned to the "interactive video watching" group, which allowed us to estimate the effect of interactive video compared to the other two video watching formats. To enhance the precision of our estimates, we included covariates for children's age, baseline language skills, and baseline science knowledge.

The third research question aimed to describe the correlation between science learning and verbal engagement outcomes. Pearson correlation coefficients were first calculated for each verbal engagement metric and science learning posttest, for interactive and pseudo-interactive conditions respectively. Regression models were also employed to control for potential confounding covariates (i.e., age, baseline language skills, and baseline science knowledge). In the regression model, interaction terms between each of the verbal engagement predictors and the experimental condition were added in order to examine the differences in the relationships between engagement and learning across the two conditions.

### Transparency and Openness Statement

We report how we carried out the data collection, all data exclusions, all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards (Kazak, 2018). This study's design and its analysis were pre-registered (https://osf.io/j7upd). All data, analysis code, and research materials are available the Open Science Framework and can be accessed at [doi provided upon publication]. Data were analyzed using R, version 4.0.0 (R Core Team, 2020) and the packages ggplot2, version 3.2.1 (Wickham, 2016), lme4, version 1.1.35.1 (Bates et al., 2009), lavaan, version 0.6.17 (Rosseel, 2012).

## Results

### Effect on Children's Science Learning

Our first research question examined the impact of this intervention on children's science learning. A researcher-developed assessment measured knowledge of concepts introduced in both episodes of the show using questions that were different from the ones asked by Elinor during the video watching.

Descriptive statistics on children's science learning are shown in Table 2. Across the three experimental conditions, children who watched the interactive version of the video exhibited the highest overall performance in the post-viewing science assessment, with a mean score of 28.41 out of 45 possible points (*SD* = 8.52). This was followed by children who viewed the pseudo-interactive version, achieving a mean score of 26.87 (*SD* = 7.59), and those in the non-interactive condition, with a mean score of 25.06 (*SD* = 8.61). This trend—where interactive video viewers outperformed their counterparts in both pseudo-interactive and non-interactive conditions—was consistent across both recall and transfer question types.

Regression analyses including children's age, baseline science knowledge, and language proficiency as covariates were performed. Four children with post-test scores that were two standard deviations below the mean were excluded from the analysis (Aguinis et al., 2013), with two in the interactive condition and two in the pseudo-interactive condition (see Figure 2A). As displayed in Table 3 and Figure 2B, the findings indicated that children in the interactive condition scored 0.32 standard deviations higher than those in the non-interactive condition ($p < .001$) and 0.16 standard deviations higher than those in the pseudo-interactive condition ($p = .046$), with the interactive condition serving as the reference group. In addition, in order to compare the effects between non-interactive and pseudo-interactive conditions, we changed the reference group to children in the non-interactive condition in our regression model (Gordon, 2015). We found that children in the pseudo-interactive condition also performed better than those in the non-interactive condition by 0.17 standard deviations ($p = .04$). A robustness check was conducted, which involved including the four children with outlier post-test scores. The resulting coefficients remained consistent; however, the difference between the pseudo-interactive and interactive conditions became marginally significant ($p = .06$) due to the increased standard error.

When further breaking down the post-test scales into recall and transfer, a similar trend was observed in both learning dimensions in terms of the magnitudes of the effect. However, the levels of significance decreased, likely due to increased standard errors associated with the reduced number of items after separating the scale. For the recall dimension, those in the interactive condition outperformed their counterparts in the non-interactive condition by 0.33 standard deviations ($p$ = .001) and those in the pseudo-interactive condition by 0.19 standard deviations ($p$ = .056). For the transfer dimension, it was observed that children in the interactive condition excelled over their peers in the non-interactive condition by 0.36 standard deviations ($p < 0.001$). Moreover, they scored 0.15 standard deviations higher compared to those in the pseudo-interactive condition, although this difference was not statistically significant ($p$ = 0.13).

### *Exploratory Analysis on Heterogeneous Effects by Child Characteristics*

We conducted additional exploratory analyses to examine the potential influence of children's baseline characteristics on the effects of interactive videos. Specifically, we investigated whether factors such as age, science knowledge, and English proficiency played a role. These variables were chosen based on their significance as predictors of science post-test outcomes in the previous main analysis, as well as the existing literature on using dialogue to support children's learning (Mol et al., 2008). Three separate regression models were employed (Table 4), and there were no significant interaction effects between assigned condition and age, science knowledge, or English proficiency.

## Effect on Children's Verbal Engagement

We also examined the effect of interactive videos on children's verbal engagement by examining how children responded to the questions posed by Elinor. Given that children in the non-interactive condition did not receive questions at all, they were not included in this analysis. Thus, the analytic sample consisted of 160 children who were in the interactive (*n* = 81) or pseudo-interactive condition (*n* = 79).

**Response Quantity**. Descriptive statistics are displayed in Table 5. Children exhibited a higher likelihood of responding to questions posed by the interactive character.  Children in the pseudo-interactive condition responded to 68.80% of the questions. For those in the interactive condition, the initial response rate averaged 78.24%, increasing to 80.80% in the follow-up responses. Regarding response length, children in the pseudo-interactive condition had an average response length of 2.58 words. Similarly, in the interactive condition, the initial responses from children also averaged 2.58 words, yet there was a slight increase to 2.98 words among the follow-up questions.

Regression analysis was performed including children's age, science knowledge, English and Spanish proficiency as covariates. The findings, detailed in Table 6, reveal that children in the interactive condition exhibited a response rate that was 0.38 standard deviations higher than those in the pseudo-interactive condition when making their initial attempts to answer questions ($p < .05$). This discrepancy widened further to 0.50 standard deviations when accounting for the additional scaffolding and corrective feedback given to children in the interactive condition for incorrect initial responses ($p < .01$). However, in terms of the length, no significant differences in either initial or follow-up response were found between the two conditions.

***Exploratory Analysis of Change in Children's Response Rate.*** The above analysis suggests that children in the interactive condition exhibited a higher overall response rate compared to those in the pseudo-interactive condition. However, it would be revealing to examine how their response rate changed over the course of the video watching.

Children in both the interactive and pseudo-interactive conditions were given the same instructions before watching the videos. They were informed that Elinor would ask them questions during the video and they were supposed to respond to those questions. Therefore, we anticipated that any differences in the level of verbal engagement among children in the two conditions would mainly be attributed to their perception of Elinor’s contingency or lack thereof, with the exception of the initial question that children were presented with.

To confirm this hypothesis, we compared children’s response rate to the initial question and found that 72.05% of children in the interactive condition responded to the first question compared to 80.52% of children in the pseudo-interactive condition. This difference was not statistically significant through a two-sample proportion test ($t$ = 1.01, $p$ = .30). The decision to focus on children’s initial attempts in both conditions aimed to ensure a fair comparison of children's inclinations to respond across conditions, as in the pseudo-interactive condition, the children did not receive the nudge that prompted responses, unlike in the interactive condition.

To examine how children’s response rate changes over time, we employed a multilevel logistic regression model to estimate the likelihood of children’s responses to each question. This model included the question number, as an indicator of progression through the video, and the video condition—either interactive or pseudo-interactive—as key variables. An interaction term between question number and video condition was also included to examine whether the change of children’s response rate differed depending on whether they were in the interactive or pseudo-interactive condition. A random effect for individual child participants was included to allow each child to have a different baseline likelihood of responding.

The regression results are presented in Table 7. The estimate for the effect of question number was 0.049 (*SE* = .011, *p* < .001). The estimate for the effect of the pseudo-interactive condition compared to the interactive condition, was -0.618 (*SE* = 0.397, *p* = .157), yet this is not statistically significant. However, the estimate for the interaction effect between question number and the video watching condition was significant ($\beta$ = -0.054, *p* < .001), suggesting that the positive effect of question number on the log odds of responding was 0.054 lower in the pseudo-interactive condition than the interactive condition. In other words, the estimated odds of a child verbally responding in the interactive condition increased by 5.0% with each question (*OR* = 1.050, *p* < .001). In contrast, in the pseudo-interactive condition, for each additional question, the estimated odds of a child responding decreased by 0.5% (*OR* = 0.995, *p* < .001). Overall, this analysis indicated that as the video progressed, children in the pseudo-interactive condition became less likely to respond to the questions, while children in the interactive condition became increasingly more inclined to respond.

**Response Quality**. Descriptive statistics of response quality are presented in Table 8. In terms of relevance, children in the pseudo-interactive condition received an average relevance score of 1.21 for their responses to Elinor. For those in the interactive condition, initial relevance scores averaged 1.34, increasing to 1.43 in the guided responses. Regarding accuracy, the pseudo-interactive group answered 40.76% of Elinor's questions correctly, while the interactive group's accuracy was 45.09% initially, improving to 53.59% in the guided responses.

We employed the same regression models as response quantity controlling for children's age, science knowledge, English and Spanish proficiency. Results are displayed in Table 9. When examining children's initial attempts to respond to questions, children in the interactive condition demonstrated a slight advantage in both response accuracy and relevance over those in the pseudo-interactive condition, though these differences were not statistically significant. Considering the scaffolding and corrective feedback provided to children in the interactive condition when their initial responses were incorrect, analysis of their subsequent responses showed significantly better performance. In terms of response accuracy, children in the interactive condition performed 0.46 standard deviations better than those in the pseudo-interactive condition ($p < .001$), and for response relevance, children in the interactive condition outperformed their counterparts by 0.40 standard divisions ($p < .01$).

## Correlations between Verbal Engagement and Science Learning

Thus far, our findings revealed that incorporating an intelligent character to interact with children during video watching indeed influences their verbal participation and, importantly, their science learning outcomes from the video. Yet it is also important to explore the relationship between verbal engagement and science learning outcomes—in particular, whether the relationship is influenced by the provision of contingent interaction. To provide evidence on this aspect, we examined the correlations between the four metrics of children's verbal engagement (i.e., response rate, response length, relevance, accuracy) and their science learning outcomes (i.e., score on post-test questions).

As presented in Table 10, our Pearson correlation analysis revealed a generally positive correlation between children's verbal engagement and their science learning outcomes. Regarding children's response rate, the interactive condition had a correlation of 0.53 for initial response and 0.50 for follow-up responses (both $p < .001$). In the pseudo-interactive condition, however, this correlation was not significant ($r = .19$, $p = .10$). Regarding children's response length, the interactive group showed a correlation of 0.27 ($p = .02$) for the initial responses and

0.28 (*p* = .01) for the follow-up response, similar to the pseudo-interactive group (*r* = .26, *p* = .02). Furthermore, regarding children's response relevance, the interactive condition had a correlation of 0.62 for initial response and 0.60 for follow-up responses (both *p* < .001), compared to the pseudo-interactive condition's weaker correlation of 0.29 (*p* = .01). Similarly, the interactive condition also showed significantly stronger correlations at 0.65 initially and 0.64 with follow-up questions (both *p* < .001), versus the pseudo-interactive condition's 0.45 (*p* < .001). However, response length did not appear to significantly correlate with learning outcomes in the interactive condition, yet this correlation was significant in the pseudo-interactive condition (*r* = .27, *p* = .02).

To further examine whether the correlations between verbal engagement and learning outcomes differed across interactive and pseudo-interactive conditions, we conducted a regression analysis (see Table 11). These analyses revealed that, for response rate and response relevance, regardless of the provision of follow-up questions, they were significantly stronger predictors of children's post-test science learning outcomes in the interactive conditions compared to the pseudo-interactive condition, controlling for children's age and baseline language and science knowledge (Model 1, Model 2, Model 7, and Model 8 in Table 11). However, the correlations between science learning with response length and accuracy significantly did not differ between conditions. Taken together, these data suggest that while children's verbal engagement could serve as strong predictors for their learning outcomes in general, their verbal engagement with an AI-enhanced character showed stronger correlations.

## Discussion

Children spend a substantial amount of time watching video programs. Consequently, both the research community and the media sector have made extensive efforts to understand which factors enhance children's learning from such content. These investigations have consistently shown that opportunities for children to engage in dialogic interactions around media can enhance both learning and engagement (Stevens & Takeuchi, 2011; Strouse et al., 2013). Our research is among the first to employ AI, in particular, automatic speech recognition and natural language processing, in facilitating this endeavor.

Our primary research question centered on whether children's understanding of science concepts in videos would improve when they engaged in dialogue with an interactive character while watching, as opposed to simply watching a non-interactive or pseudo-interactive version. The data confirmed our hypothesis: the best learning outcomes were seen when children had dialogic interactions—that include both questioning and responsive feedback—with the character. This was followed by the pseudo-interactive condition, where children were posed questions but received only generic feedback, and lastly, the non-interactive condition. The observed advantage of interactive videos over non-interactive videos is consistent with the widely documented findings regarding the benefits of dialogic interactivity in children's learning from videos (for a review, see Strouse & Samson, 2021) and prior research on incorporating conversational agents in television watching (Xu et al., 2021; Xu et al., 2022). In particular, the effect size of the added dialogic interactivity was moderate and is consistent with the prior studies (Xu et al., 2021; Xu et al., 2022).

Nevertheless, one may theorize that such a difference might not be due to the contingent nature of the interactions themselves but rather simply the fact that children were asked questions (and thus prompted to think about the concepts) and provided with repeated reinforcement of information (as children received explanations of the answers). Yet our next finding suggested that the character's responsiveness, as reflected in the character's ability to initiate responses at the appropriate timing and provide feedback and scaffolding, did matter. This finding contributes to the body of research that points to the importance of responsive feedback. For example, as described in the literature review section of Calvert et al. (2020), the study involving Dora the Explorer, where the character corrected children's incorrect responses, resulted in an increase in posttest math scores by 0.69 standard deviations compared to a non-responsive version. In this version, Dora asked questions, but the feedback provided did not include corrections for incorrect answers given by the children. It is likely that, as in our study, the process of receiving corrective feedback and reconstructing responses played an important role in facilitating children's learning. Indeed, as we presented in the results, children who watched the interactive episodes initially answered 44% of Elinor's questions incorrectly, which was quite comparable to those in the pseudo-interactive condition. However, given that Elinor, in the interactive condition, provides scaffolding to children to guide them to reconstruct their answers, these children were able to correct 54% of their initially incorrect answers, thus leading to a higher overall accuracy rate than those in the pseudo-interactive condition.

Moreover, these mechanisms may be enhanced by social factors, where children's motivation and engagement to learn are heightened when they interact with a responsive character. This idea aligns with theories established from other research, which indicates that children tend to trust and find information more relevant when it comes from a media character with whom they feel a connection (Calvert et al., 2014; Danovitch & Mills, 2014). Therefore, it is plausible that Elinor's responsiveness in the interactive videos could have deepened children's sense of connection with her, thereby facilitating a more effortful learning experience. Although we did not explicitly examine children's social perceptions of Elinor, our analysis of their verbal engagement with Elinor offers insights into how dialogic interactivity might leverage social connections to support learning.

Our second research question examined children's verbal engagement with both pseudo-interactive and interactive characters. Overall, we confirmed our hypothesis that the character's responsiveness influenced children's response patterns. Interestingly, the interactive character seemed to bolster children's inclination to respond, while the pseudo-interactive condition seemed to deter engagement. This observation aligns with the broader understanding that children adjust their behavior towards conversational partners, whether human (Kornhaber & Marcos, 2000) or virtual (Beneteau et al., 2019; Cheng et al., 2018), based on their interaction experiences and perceptions. It is also in line with the reciprocal determinism theory that outlines how children perceive and evaluate environmental influences (i.e., the responsiveness of the media character in this case) and regulate their own subsequent behaviors (Bandura, 1978). Indeed, an earlier study has also focused on the longitudinal pattern of children's interaction with responsive or non-responsive media characters, and revealed the same pattern as our study (Carter et al., 2017). However, a caveat when interpreting the response rate trend is that the nature of the questions posed by the character could also have significantly influenced a child's decision to respond. Data points of response rate for individual questions varied considerably, as illustrated in Figure 3. While we currently lack definitive evidence pinpointing which question types elicited more responses from children, this pattern appeared consistent with another study focusing on children's answering of questions posed by a smart-speaker (Xu & Warschauer, 2020). In this study, children were found to be more likely to respond to questions pertaining to their personal experiences, as well as to questions that asked them to make yes-or-no choices.

Additionally, regarding the comparison of children's response lengths in both conditions, it appears that although the word count of children's responses is similar, the length of responses from children in the pseudo-interactive condition was more likely due to repetition. This is because the pseudo-interactive Elinor is programmed to wait 10 seconds before responding. Consequently, many children, having finished their responses in less than 10 seconds, believed Elinor had not heard them and thus repeated themselves. This behavior mirrors the repetition strategy children commonly employed with voice assistants, such as Amazon Alexa, when they fail to register the children's speech (Cheng et al., 2018).

Our findings indicate that the quality of children's responses, in terms of relevance and accuracy, was similar between the interactive and pseudo-interactive conditions. This is not surprising as the quality of children's initial responses would likely be linked to their ability to comprehend the content from the episode itself, which should align with their baseline characteristics that were consistent across both groups (Xu et al., 2021). However, when taking into consideration the children in the interactive condition's advantage of receiving scaffolding and responsive feedback, the relevance and accuracy improved and became significantly higher than those in the pseudo-interactive condition who received only generic feedback. This improved response quality is consistent with the principles of ZPD, suggesting that the targeted, responsive scaffolding provided in the interactive condition effectively bridged the gap between the children's current understanding and their potential for higher-level comprehension and application of knowledge. In addition, these results could provide empirical evidence supporting the hypothesized "knowledge change" that may occur in children's learning process (Fyfe et al., 2023, *p*. 130), where feedback signals a gap between children's current and expected levels of comprehension and facilitates their identification of misconceptions. Moreover, as we discussed previously, the social factor might have also been in play here—Elinor's responsiveness might have strengthened the connections children might have felt, thus encouraging them to put more cognitive efforts into answering Elinor's questions. This social mechanism was discussed in another study that compared children's responses to questions asked by a conversational agent versus those asked by a person (Xu et al., 2021).

Lastly, our study has gathered evidence that children's verbal engagement—in particular, response rate, relevance, and accuracy—exhibits a more robust and positive correlation with their science learning outcomes when they interact with the interactive character than with the pseudo-interactive character. This result can be interpreted in two ways. On the one hand, it could imply that the benefits of responding to Elinor's questions on science learning were more potent in the interactive condition. This could be because the feedback children received further amplified the educational advantages stemming from their verbal engagement (Landry et al., 2012). On the other hand, it could reflect the fact that the verbal engagement of children during the interactive videos is a better indicator of their comprehension of the educational content within the videos. In this case, interactive videos could serve as a potential mechanism for capturing children's learning from video programming (Spitale et al., 2020), a task that would otherwise be challenging to evaluate without explicit assessment.

Taking everything into account, it would be helpful to contextualize all of these findings within the specific participant population we worked with. While the majority were Spanish-English bilingual children, our intervention was conducted in English, in line with the overwhelming majority of media resources available in the U.S. for children. Children in our sample displayed a wide spectrum of language dominance profiles, with varying proficiencies across different linguistic aspects across the two languages. Our study results revealed that children's English proficiency positively contributed to their performance in the posttest, but their proficiency in Spanish was not a statistically significant factor. Additionally, we found that children with limited English proficiency almost did not gain as much from the episode as reflected by our posttest assessments. In addition, when analyzing verbal engagement, it appeared that the overwhelming majority of children chose to use English to engage in dialogic interactions with Elinor, despite the fact that some of those children might be stronger in Spanish. It is likely that the nature of English availability in the episodes might have encouraged children to use English instead of Spanish during the interaction. This highlights the importance for future studies to develop media resources that better leverage the linguistic assets of minoritized children. Indeed, one recent exploratory study specifically focused on designing an AI partner that can converse in both Spanish and English with children during storybook reading (Xu et al., 2023). The support in their home language provided by the AI was found to encourage engagement among children with initially limited English proficiency. Future research may benefit from further exploring conversational AI that incorporates linguistic features aligned with the children's language proficiency profiles.

## Practical Implications

Our study has significant implications for the early childhood education media sector. As children's media consumption continues to rise, the demand for valuable, scalable, high-quality programming is critical. Our study presents evidence that AI technologies can be used to facilitate this goal. With children dedicating over two-thirds of their video-watching time to internet-connected devices, integrating speech recognition and natural language technologies into their video content is viable. Indeed, PBS KIDS has committed to distributing the interactive videos created in this project through its online platforms that reach millions of children a month (Schwartz, 2023; White, 2023). Importantly, technology companies such as Google, Amazon, and OpenAI have introduced developer tools for speech recognition and natural language processing, streamlining much of the technical complexity involved in creating interactive videos. This study also holds relevance to other stakeholders of children's learning. For educators and parents, our findings reaffirm the critical importance of engaging children in personalized conversations to scaffold learning both inside and outside the classroom. Such engagement should be maintained even as AI-enabled media provides complementary educational experiences. Additionally, it is imperative to address the implications of educational equity as reshaped by this technology. On one hand, AI-driven programming may offer scalable learning resources for children who might otherwise lack access to enriched learning opportunities with adults. On the other hand, the accessibility of these innovative educational tools could be uneven, particularly during the initial phases of deployment.

Furthermore, the advances of generative AI hold promises of further enhancing the quality of dialogic interactions that media characters can provide to children. In the current study, children's potential interactions with the AI-enabled character were limited to a set of predefined dialogue branches. This dialogue tree approach offers advantages, as it allows for all content to which children are exposed to be meticulously designed and validated by professionals. However, a significant drawback is the extensive time and effort required to create numerous branches to adequately cover children's likely responses, and the inability of the character to respond dynamically to child responses that fall outside the anticipated themes. Indeed, in our analysis of interactive videos, we noted that human coders were tasked with classifying children's responses into predefined categories rather than interpreting the nuances of these responses. This method may overlook opportunities for more precise, targeted scaffolding when the categories are too broad and overlook subtle linguistic nuances that could indicate children's comprehension of the specific topic. Generative AI might present an opportunity to support personalized interactions to a greater extent and enhance the quality of interactive video content on a larger scale. Nonetheless, it may still be premature to widely implement media with generative AI for children at this moment until further research is conducted to ensure its safety and educational value.

## Ethical Considerations

Several ethical considerations warrant discussion regarding the use of AI in children's media. One concern is about potential biases in underlying speech recognition and natural language processing algorithms, creating barriers for children from minoritized language backgrounds to effectively interact with these technologies. Though in this study, the media character's ability to respond appropriately and responsively did significantly vary by children's linguistic backgrounds, studies that focus more fine-grained speech recognition performance, such as on phonemic recognition, suggest that children's linguistics profile may impact speech recognition performance (Thomas et al., 2023).

Second, privacy remains a significant concern. The use of speech recognition and natural language processing in our interactive videos may require sharing children's language data with third-party services. Currently, there is a lack of robust regulation in this area. For instance, while the Children's Online Privacy Protection Rule (COPPA) was established in 1998 to protect the online safety of children under 13, it does not explicitly cover AI technologies. We urge developers to closely examine the privacy policies of their tools, especially regarding whether children's data would be retained and used for further training purposes. Additionally, it is crucial for policymakers to update and adapt regulations as AI technologies become more widespread. The privacy concern is compounded by AI's increasingly seamless integration into online media and platforms, which raises issues that children may be unknowingly interacting with AI that analyzes and adapts to their thoughts or actions. This leads to considerations regarding consent and the ethical means of engaging children with AI technologies. It is important to consider what measures can be taken to ensure that children understand and consent to these interactions in an age-appropriate manner.

Third, there is a growing interest in utilizing AI to engage children in dialogues that extend beyond STEM topics to incorporate elements such as cultural awareness, moral values, and socioemotional development. This expansion introduces more complex ethical dimensions. For instance, when AI is used to facilitate discussions about cultural heritage, it must navigate sensitively around diverse traditions and beliefs to avoid perpetuating stereotypes or biases. Yet the analysis of state-of-the-art AI models suggests that these algorithms represent Western, White mainstream norms and values (Blodgett et al., 2020; Venkit et al., 2023). This broadening of AI's application in children's media to cover these aspects of human experience, which are deeply imbued with cultural significance, requires a careful approach to ensure that AI interactions are inclusive, culturally sensitive, and ethically aligned with the diverse backgrounds of the children they serve.

## Limitations and Future Directions

There are several potential avenues for extending the scope of this research. First, a deeper exploration of learning outcomes is warranted. In our study, we assessed learning immediately following children's video sessions. However, the extent to which children retain the acquired information over time remains unknown. Furthermore, it is noteworthy that the children in our study were exposed to only two episodes on a single occasion each. As a result, our assessment of learning centered on the children's comprehension of the concepts presented within those specific episodes. Future research could consider providing children with access to interactive videos encompassing a wider spectrum of scientific concepts and investigate whether interactive videos confer similar benefits compared to pseudo-interactive or non-interactive counterparts in terms of general science knowledge and skills on a cumulative scale. Beyond cognitive and linguistic development, it is crucial to examine the impact on children's sense of belonging, experiences of loneliness, and enjoyment of the media. These factors are integral to understanding the holistic effects of AI interactions on children's well-being and development.

Second, we do not yet know how children's learning and engagement with an interactive media character would compare to watching videos and engaging in discussions with a human partner. Previous research has suggested that, in the context of storybook reading, dialogue with an AI partner led to similar comprehension benefits as with a human partner (Xu et al., 2022), albeit with reduced verbal engagement from the children (Xu et al., 2021). Nevertheless, in the context of video watching, where children have been found to establish emotional connections with media characters, it is also possible that interacting with characters children already like might foster a heightened willingness to engage (Piotrowski, 2014). In a similar vein, long-term interaction with an unfamiliar character may yield increased engagement across repeated viewings, as children gain familiarity with the character and episode content.

Third, our study recruited children from a specific geographical location, potentially increasing the homogeneity of participants in terms of race/ethnicity and linguistic backgrounds compared to a national sample. While focusing on this particular group has its advantages, given their documented longer time spent with (Rideout & Robb, 2020) and appreciation of educational television as learning resources (Kalinowski et al., 2020), it would be beneficial for future studies to replicate this study with a broader range of demographics and linguistic profiles, and to expand it to include neurodivergent children. Moreover, while the sample size was adequate for detecting the main condition effects on learning and verbal engagement, it did not allow for the detection of significant moderation effects. This limitation restricts our ability to explore whether the effects of interactive videos might differ based on children's language proficiency, science baseline knowledge, or age, which could be a promising direction for future studies.

## Conclusion

In conclusion, we found that watching an interactive version of a PBS KIDS television show benefits both children's verbal engagement and science learning. As the time children spend watching videos increases and the form of watching shifts to internet-connected devices, it is crucial to investigate how new forms of video-watching may better support learning. This study is timely, as the recent advances in AI have drawn much attention from researchers, media producers, and the general public, prompting contemplation of the role AI can play in enriching children's educational media. Our study suggests a viable mechanism for this to happen. In addition to the evidence this study contributes, our study may also serve as a call for more research to investigate this topic in depth, drawing nuanced conclusions about the specific circumstances under which children can benefit from interactive video-viewing experiences, as well as any potential developmental costs associated with such interactions. This endeavor will empower researchers to provide concrete recommendations for the effective integration of AI into children's media.

## References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2), 270-301. https://doi.org/10.1177/1094428112470848

Anderson, D. R., & Hanson, K. G. (2017). Screen media and parent–child interactions. In R. Barr & D. N. Linebarger (Eds.), *Media exposure during infancy and early childhood* (pp. 173–194). Springer International Publishing.

Bandura, A. (1978). The self system in reciprocal determinism. *American Psychologist*, 33(4), 344–358. https://doi.org/10.1037/0003-066X.33.4.344

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61(2), 213-238. https://doi.org/10.3102/00346543061002213

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Green, P. (2009). Package 'lme4'. URL http://lme4.r-forge.r-project.org.

Bedore, L. M., Peña, E. D., Collins, P., Fiestas, C., Lugo-Neris, M., & Barquin, E. (2023). Predicting literacy development and risk in Spanish-English bilingual first graders. *Child Language Teaching and Therapy*, 02656590231166923.

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., & Hiniker, A. (2019, May). Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). https://doi.org/10.1145/3290605.3300473

Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Bonus, J. A., Dore, R. A., Wilson, J. M., Freiberger, N., & Lerner, B. (2023). Of scientists and superheroes: Educational television and pretend play as preparation for science learning. *Journal of Applied Developmental Psychology*, 89, 101603. https://doi.org/10.1016/j.appdev.2023.101603

Calvert, S. L., Richards, M. N., Jordon, A., & Romer, D. (2014). Children’s parasocial relationships. Media and the well-being of children and adolescents, 187-200.

Calvert, S. L., Putnam, M. M., Aguiar, N. R., Ryan, R. M., Wright, C. A., Liu, Y. H. A., & Barba, E. (2020). Young children’s mathematical learning from intelligent characters. *Child Development*, *91*(5), 1491–1508. https://doi.org/10.1111/cdev.13341

Carter, E. J., Hyde, J., & Hodgins, J. K. (2017). Investigating the effects of interactive features for preschool television programming. *Proceedings of the 2017 Conference on Interaction Design and Children*, 97–106. https://doi.org/10.1145/3078072.3079717

Cassell, J. (2022). Socially interactive agents as peers. In B. Lugrin, C. Pelachaud, & D. Traum (Eds.), *The handbook on socially interactive agents* (1st ed., pp. 331–366). ACM. https://doi.org/10.1145/3563659.3563670

Cheng, Y., Yen, K., Chen, Y., Chen, S., & Hiniker, A. (2018). Why doesn’t it work?: Voice-driven interfaces and young children’s communication repair strategies. *Proceedings of the 17th ACM Conference on Interaction Design and Children*, 337–348. https://doi.org/10.1145/3202185.3202749

Choi, K. (2021). *Sesame Street*: Beyond 50. *Journal of Children and Media*, *15*(4), 597–603. https://doi.org/10.1080/17482798.2021.1978675

Christensen, C., Hoisington, C., Vahey, P., Hupert, N., & Pasnik, S. (2019). PBS Kids Play & Learn Science. Evaluation Report. Education Development Center, Inc. Retrieved from: https://files.eric.ed.gov/fulltext/ED599703.pdf

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49, 222-251. https://doi.org/10.3102/00346543049002222

Dai, L., Jung, M. M., Postma, M., & Louwerse, M. M. (2022). A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects. Computers & Education, 190, 104607. https://doi.org/10.1016/j.compedu.2022.104607

Danovitch, J. H., & Mills, C. M. (2014). How familiar characters influence children's judgments about information and products. *Journal of Experimental Child Psychology*, 128, 1-20. https://doi.org/10.1016/j.jecp.2014.06.001

D'Mello, S. K., & Graesser, A. (2023). Intelligent tutoring systems: How computers achieve learning gains that rival human tutors. In P. A. Schutz & K. R. Muis (Eds), *Handbook of Educational Psychology* (pp. 603-629). Routledge. https://doi.org/10.4324/9780429433726-31

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. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445039

Druga, S., Williams, R., Park, H. W., & Breazeal, C. (2018). How smart are the smart toys?: Children and parents' agent interaction and intelligence attribution. *Proceedings of the 17th ACM Conference on Interaction Design and Children*, 231–240. https://doi.org/10.1145/3202185.3202741

Ewin, C. A., Reupert, A. E., McLean, L. A., & Ewin, C. J. (2021). The impact of joint media engagement on parent–child interactions: A systematic review. *Human Behavior and Emerging Technologies*, *3*(2), 230–254. https://doi.org/10.1002/hbe2.203

Feng, S., Magana, A. J., & Kao, D. (2021, October). A systematic review of literature on the effectiveness of intelligent tutoring systems in STEM. In 2021 *IEEE Frontiers in Education Conference* (FIE) (pp. 1-9). IEEE.

Fisch, S. M. (2000). A capacity model of children's comprehension of educational content on television. *Media Psychology*, 2(1), 63-91. https://doi.org/10.1207/S1532785XMEP0201_4

France, A. (2021). Teachers using dialogue to support science learning in the primary classroom. *Research in Science Education*, *51*(3), 845–859. https://doi.org/10.1007/s11165-019-09863-3

Fyfe, E. R., Borriello, G. A., & Merrick, M. (2023). A developmental perspective on feedback: How corrective feedback influences children's literacy, mathematics, and problem solving. *Educational Psychologist*, 58(3), 130-145. https://doi.org/10.1080/00461520.2022.2108426

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. https://doi.org/10.3389/fpsyg.2020.02158

Gordon, R. A. (2015). *Regression analysis for the social sciences*. Routledge.

Graesser, A. C., D'Mello, S., & Person, N. (2009). Meta-knowledge in tutoring. In *Handbook of metacognition in education* (pp. 361-382). Routledge.

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education*, 26, 124-132. https://doi.org/10.1007/s40593-015-0086-4

Greenfield, D. B. (2015). Assessment in early childhood science education. In K. Trundle, & M. Saçkes (Eds.), Research in early childhood science education. Chapter 16 (pp. 353–380). New York, NY: Springer Publishing.

Greenfield, D. B., & Penfield, R. (2013). Lens on science: development and validation of a computer-administered, adaptive, IRT-based assessment for preschool children. Institute of Education Sciences Grant R305A090502, http://ies.ed.gov/funding/grantsearch/details.asp?ID=805

Guthrie, J. T., Klauda, S. L., & Ho, A. N. (2013). Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents. *Reading Research Quarterly*, 48(1), 9-26. https://doi.org/10.1002/rrq.035

Haines, J., O'Brien, A., McDonald, J., Goldman, R. E., Evans-Schmidt, M., Price, S., King, S., Sherry, B., & Taveras, E. M. (2013). Television viewing and televisions in bedrooms: Perceptions of racial/ethnic minority parents of young children. *Journal of Child and Family Studies*, *22*(6), 749–756. https://doi.org/10.1007/s10826-012-9629-6

Hamel, E., Joo, Y., Hong, S. Y., & Burton, A. (2021). Teacher questioning practices in early childhood science activities. *Early Childhood Education Journal*, 49, 375-384. https://doi.org/10.1007/s10643-020-01075-z

Hsueh, Y., Zhou, Z., Su, G., Lee, J., & Kitzmann, K. (2017). Science learning in early years: Effects of the Chinese television series Big Bird Looks at the World. *Global Media and China*, 2(2), 183-196. https://doi.org/10.1177/2059436417717072

Hurwitz, L. B. (2019). Getting a read on ready to learn media: A meta-analytic review of effects on literacy. *Child Development*, 90(5), 1754-1771. https://doi.org/10.1111/cdev.13043

Jing, M., Ye, T., Kirkorian, H. L., & Mares, M. L. (2023). Screen media exposure and young children's vocabulary learning and development: A meta-analysis. *Child Development.* https://doi.org/10.1111/cdev.13927

Jing, M., & Kirkorian, H. L. (2020a). Video deficit in children’s early learning. In J. Bulck (Ed.), *The international encyclopedia of media psychology* (1st ed., pp. 1–8). Wiley. https://doi.org/10.1002/9781119011071.iemp0239

Jing, M., & Kirkorian, H. L. (2020b). Teaching with televised stories: A story-focused narrative preview supports learning in young children. *Child Development*, 91(5), e1101-e1118. https://doi.org/10.1111/cdev.13385

Kalinowski, R. D., 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* (pp. 1-14). https://doi.org/10.1145/3411764.3445429

Kang, M., Ragan, B. G., & Park, J. H. (2008). Issues in outcomes research: an overview of randomization techniques for clinical trials. *Journal of Athletic Training*, 43(2), 215-221. https://doi.org/10.4085/1062-6050-43.2.215

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, 73(1), 1–2. https://doi.org/10.1037/amp0000263

Kendeou, P., Bohn-Gettler, C., White, M. J., & Van Den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, 31(3), 259-272. https://doi.org/10.1111/j.1467-9817.2008.00370.x

Kendeou, P., McMaster, K. L., Butterfuss, R., Kim, J., Bresina, B., & Wagner, K. (2020). The inferential language comprehension (iLC) framework: Supporting children's comprehension of visual narratives. *Topics in Cognitive Science*, 12(1), 256-273.

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, 113(1), 3–26. https://doi.org/10.1037/edu0000465

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2023). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension. *Journal of Educational Psychology*, 115(1), 73–98. https://doi.org/10.1037/edu0000751

Kornhaber, M., & Marcos, H. (2000). Young children's communication with mothers and fathers: Functions and contents. *British Journal of Developmental Psychology*, *18*(2), 187–210. https://doi.org/10.1348/026151000165643

Kostyrka-Allchorne, K., Cooper, N. R., & Simpson, A. (2017). The relationship between television exposure and children's cognition and behaviour: A systematic review. *Developmental Review*, 44, 19-58. https://doi.org/10.1016/j.dr.2016.12.002

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Landry, S. H., Smith, K. E., Swank, P. R., Zucker, T., Crawford, A. D., & Solari, E. F. (2012). The effects of a responsive parenting intervention on parent–child interactions during shared book reading. *Developmental Psychology*, *48*(4), 969–986. https://doi.org/10.1037/a0026400

Lepola, J., Kajamies, A., Laakkonen, E., & Collins, M. F. (2023). Opportunities to talk matter in shared reading: The mediating roles of Children's engagement and verbal participation in narrative listening comprehension. *Early Education and Development*, 1-23. https://doi.org/10.1080/10409289.2023.2188865

Lovato, S. B., Piper, A. M., & Wartella, E. A. (2019). Hey google, do unicorns exist?: Conversational agents as a path to answers to children's questions. *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 301–313. https://doi.org/10.1145/3311927.3323150

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia medica, 22(3), 276–282.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. https://doi.org/10.1037/a0037123

Mayer, R. E. (2002). Multimedia learning. In *Psychology of Learning and Motivation* (Vol. 41, pp. 85-139). Academic Press.

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, *13*(2), 125-139. https://doi.org/10.1016/S0959-4752(02)00016-6

McCarthy, E., Tiu, M., & Li, L. (2018). Learning math with curious George and the odd squad: Transmedia in the classroom. *Technology, Knowledge and Learning*, 23, 223-246. https://doi.org/10.1007/s10758-018-9361-4

McCarthy, K. S., Watanabe, M., Dai, J., & McNamara, D. S. (2020). Personalized learning in iSTART: Past modifications and future design. *Journal of Research on Technology in Education*, 52(3), 301-321. https://doi.org/10.1080/15391523.2020.1716201

McMaster, K. L., Kendeou, P., Kim, J., & Butterfuss, R. (2023). Efficacy of a technology-based early language comprehension intervention: A randomized control trial. *Journal of Learning Disabilities*, 00222194231182974. https://doi.org/10.1177/0022219423118297

McNamara, D. S., Arner, T., Butterfuss, R., Fang, Y., Watanabe, M., Newton, N., ... & Roscoe, R. D. (2023). iSTART: Adaptive comprehension strategy training and stealth literacy assessment. *International Journal of Human–Computer Interaction*, 39(11), 2239-2252. https://doi.org/10.1080/10447318.2022.2114143

Meacham, S., Vukelich, C., Han, M., & Buell, M. (2016). Teachers’ responsiveness to preschoolers’ utterances in sociodramatic play. *Early Education and Development*, 27(3), 318-335. https://doi.org/10.1080/10409289.2015.1057461

Mol, S. E., Bus, A. G., De Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, *19*(1), 7–26. https://doi.org/10.1080/10409280701838603

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Research Council. (2013). Next generation science standards: For states, by states. The National Academies Press. https://doi.org/10.17226/18290

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. https://doi.org/10.1016/j.chb.2018.03.039

Peña, E. D., Bedore, L. M., Gutiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B. A. (2010). *Bilingual English Spanish Oral Screener (BESOS): Pre-K, 1st, and 3rd grade*. Unpublished test, University of Texas at Austin, Austin, TX.

Piotrowski, J. T. (2014). Participatory cues and program familiarity predict young children's learning from educational television. *Media Psychology*, *17*(3), 311–331. https://doi.org/10.1080/15213269.2014.932288

Relyea, J. E., & Amendum, S. J. (2020). English reading growth in spanish-speaking bilingual students: Moderating effect of english proficiency on cross-linguistic influence. *Child Development*, 91(4), 1150-1165. https://doi.org/10.1111/cdev.13288

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. https://doi.org/10.1111/j.1467-8624.2010.01542.x

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.

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: interactive, linguistic, and conceptual. Journal of Child Language, 47(Special Issue 1), 5-21. https://doi.org/10.1017/S0305000919000655

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype Me! Socially contingent interactions help toddlers learn language. *Child Development*, *85*(3), 956–970. https://doi.org/10.1111/cdev.12166

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1-36.

Ruan, S., He, J., Ying, R., Burkle, J., Hakim, D., Wang, A., ... & Landay, J. A. (2020). Supporting children's math learning with feedback-augmented narrative technology. *In Proceedings of the Interaction Design and Children Conference* (pp. 567-580). https://doi.org/10.1145/3392063.3394400

Schwartz, E. H. (2023, August 8). New PBS Kids show will include generative ai conversational interactions. https://voicebot.ai/2023/08/08/new-pbs-kids-show-will-include-generative-ai-conversational-interactions/

Spitale, M., Silleresi, S., Cosentino, G., Panzeri, F., & Garzotto, F. (2020). “Whom would you like to talk with?”: Exploring conversational agents for children’s linguistic assessment. *Proceedings of the Interaction Design and Children Conference*, 262–272. https://doi.org/10.1145/3392063.3394421

St. Peters, M., Fitch, M., Huston, A. C., Wright, J. C., & Eakins, D. J. (1991). Television and families: What do young children watch with their parents?. Child Development, 62(6), 1409-1423. https://doi.org/10.1111/j.1467-8624.1991.tb01614.x

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, *105*(4), 970–987. https://doi.org/10.1037/a0032447

Stevens, R., & Takeuchi, L. (2011). The new coviewing: Designing for learning through joint media engagement. Joan Ganz Cooney Center. https://www.joanganzcooneycenter.org/wp-content/uploads/2011/12/jgc_coviewing_desktop.pdf

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), 2368–2382. https://doi.org/10.1037/a0032463

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). https://doi.org/10.1111/cdev.13429

Thomas, T., Takahesu-Tabori, A., Stoehr, A., Varady, C., & Xu, Y. (2023). Does bilingual status influence automatic speech recognition for young latino children? *Proceedings of the 14th International Symposium on Bilingualism.*

Troseth, G. L., Strouse, G. A., Flores, I., Stuckelman, Z. D., & Johnson, C. R. (2020). An enhanced eBook facilitates parent–child talk during shared reading by families of low socioeconomic status. *Early Childhood Research Quarterly*, 50, 45-58. https://doi.org/10.1016/j.ecresq.2019.02.009

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T.-H. "Kenneth," & Wilson, S. (2023). Nationality Bias in Text Generation (arXiv:2302.02463). arXiv. http://arxiv.org/abs/2302.02463

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2023). When adaptive learning is effective learning: comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, 31(2), 793-803. https://doi.org/10.1080/10494820.2020.1808794

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105(4), 1115. https://doi.org/10.1037/a0031589

Wartella, E., Luricella, A., & Blackwell, C. (2016). The Ready to Learn Program: 2010–2015 Policy Brief. *Evanston, IL: Northwestern University School of Communication*.

Westerveld, M. F., & Roberts, J. M. (2017). The oral narrative comprehension and production abilities of verbal preschoolers on the autism spectrum. *Language, Speech, and Hearing Services in Schools*, 48(4), 260-272. https://doi.org/10.1044/2017_LSHSS-17-0003

White, A. (2023, August 2). PBS Kids' 'Lyla in the Loop' to feature interactive episodes with ai-assisted conversation (exclusive). https://www.hollywoodreporter.com/tv/tv-news/pbs-kids-lyla-in-the-loop-debut-2024-ai-conversation-1235547667/

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package ‘ggplot2’. Create elegant data visualisations using the grammar of graphics. Version, 2(1), 1-189.

Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y., & Warschauer, M. (2022). Dialogue with a conversational agent promotes children’s story comprehension via enhancing engagement. *Child Development*, *93*(2). https://doi.org/10.1111/cdev.13708

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 & Education*, *161*, 104059. https://doi.org/10.1016/j.compedu.2020.104059

Xu, Y., & Warschauer, M. (2020). Exploring young children’s engagement in joint reading with a conversational agent. *Proceedings of the Interaction Design and Children Conference*, 216–228. https://doi.org/10.1145/3392063.3394417

Xu, Y. (2023). Talking with machines: Can conversational technologies serve as children's social partners?. *Child Development Perspectives*, 17(1), 53-58. https://doi.org/10.1111/cdep.12475

Xu, Y., He, K., Vigil, V., Ojeda-Ramirez, S., Liu, X., Levine, J., Cervera, K., & Warschauer, M. (2023, June). “Rosita reads with my family”: Developing a bilingual conversational agent to support parent-child shared reading. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (pp. 160-172). https://doi.org/10.1145/3585088.3589354

Yang, D., Xia, C., Collins, P., & Warschauer, M. (2022). The role of bilingual discussion prompts in shared E-book reading. *Computers & Education*, 190, 104622. https://doi.org/10.1016/j.compedu.2022.104622

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, 1523-1545. https://doi.org/10.1007/s11423-017-9533-2

Zucker, T. A., Williams, J. M., Bell, E. R., Assel, M. A., Landry, S. H., Monsegue-Bailey, P., Crawford, A., & Bhavsar, V. (2016). Validation of a brief, screening measure of low-income pre-kindergarteners' science and engineering knowledge. *Early Childhood Research Quarterly*, *36*, 345–357. https://doi.org/10.1016/j.ecresq.2015.12.018

**Table1.**

Participant Demographics and Baseline Characteristics by Experimental Conditions

| | Full sample | Non-interactive | Pseudo-interactive | Interactive | ANOVA/$\chi^2$ |
|---|---|---|---|---|---|
| Age | 6.21 (0.86) | 6.11 (0.91) | 6.26 (0.75) | 6.26 (0.89) | $F$(2, 240[1])=0.80, $p$ =.45 |
| Ethnicity/Race | | | | | $\chi^2(2) = 2.05, p = .36$ |
| Hispanic or Latino | 95.47% | 92.94% | 97.40% | 96.30% | |
| Non-Hispanic or Non-Latino | 4.53% | 7.09% | 2.60% | 3.70% | |
| Baseline measures | | | | | |
| Science knowledge | 1.82 (1.08) | 1.78 (1.11) | 1.82 (1.13) | 1.88 (1.02) | $F$(2, 232[2])=0.18, $p$ =.84 |
| English proficiency | 17.68 (7.01) | 17.66 (7.34) | 17.95 (6.56) | 17.44 (7.16) | $F$(2, 235[3])=0.10, $p$ =.92 |
| Spanish proficiency | 9.78 (6.79) | 9.45(6.94) | 10.42 (6.75) | 9.51 (6.90) | $F$(2, 235[3])=0.50, $p$ =.61 |
| N | 246 | 86 | 79 | 81 | |

*Note*: Standard deviation in parentheses.

[1] Three observations were missing.

[2] Eleven observations were missing.

[3] Eight observations were missing.

**Table 2.**
Science Learning Outcomes by Experimental Conditions

| | Non-interactive | Pseudo-interactive | Interactive | ANOVA |
|---|---|---|---|---|
| Total | | | | $F(2, 243) = 3.44$, $p = .03$* |
| *M* (*SD*) | 25.06 (8.61) | 26.87 (7.59) | 28.41 (8.52) | |
| Median [min, max] | 27.12 [3, 28.08] | 28.33 [4.50, 39.92] | 29.75 [5.00, 39.67] | |
| *Learning Outcome by Episodes* | | | | |
| Episode 1 | | | | $F(2, 243) = 1.74$, $p = .18$ |
| *M* (*SD*) | 10.87 (3.69) | 11.11 (3.49) | 11.90 (3.92) | |
| Median [min, max] | 11.83 [1.00, 16.83] | 12.00 [1.83, 16.00] | 13.00 [2.00, 23.50] | |
| Episode 2 | | | | $F(2, 243) = 4.33$, $p = .01$* |
| *M* (*SD*) | 14.19 (5.59) | 15.76 (4.80) | 16.52 (5.18) | |
| Median [min, max] | 15.29 [0.00, 24.08] | 16.50 [1.5, 23.92] | 18.08 [2.75, 23.50] | |
| *Learning Outcome by Dimensions* | | | | |
| Recall Learning | | | | $F(2, 243) = 2.72$, $p = .07$ |
| *M* (*SD*) | 10.38 (3.67) | 10.89 (3.42) | 11.65 (3.42) | |
| Median [min, max] | 11 [0.50, 16.00] | 11.50 [0.50, 16.00] | 12.50 [0.50, 16.00] | |
| Transfer Learning | | | | $F(2, 243) = 3.41$, $p = .03$* |
| *M* (*SD*) | 15.63 (5.66) | 16.98 (4.93) | 17.83 (5.80) | |
| Median [min, max] | 17.12 [0.50, 23.58] | 18.08 [4.00, 24.92] | 19.50 [2.50, 26.75] | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.01$.

**Table 3.**

Regression on Science Learning Outcomes

| | Overall | Recall | Transfer |
|---|---|---|---|
| Pseudo-interactive | -0.162* | -0.188 | -0.145 |
| | (0.081) | (0.098) | (0.097) |
| Non-interactive | -0.322*** | -0.333*** | -0.355*** |
| | (0.079) | (0.096) | (0.094) |
| Age | 0.155** | 0.132* | 0.172** |
| | (0.052) | (0.063) | (0.061) |
| Science knowledge (LENS) | 0.374*** | 0.354*** | 0.376*** |
| | (0.044) | (0.053) | (0.052) |
| English proficiency | 0.039*** | 0.047*** | 0.039*** |
| | (0.007) | (0.008) | (0.008) |
| Spanish proficiency | -0.000 | -0.008 | 0.005 |
| | (0.005) | (0.006) | (0.006) |
| Intercept | -2.146*** | -2.019*** | -2.297*** |
| | (0.274) | (0.331) | (0.326) |
| Observations | 242 | 242 | 242 |
| R-squared | 0.67 | 0.59 | 0.61 |

*Note*: Standard error in parentheses. The interactive condition was set as the reference group. The dependent variables, children's science overall, recall, and transfer learning scores, were standardized, so the mean was zero and the standard deviation was one. Full information maximum likelihood (FIML) was employed to compute missing data. Four children, two in the pseudo-interactive condition, and two in the interactive condition, were excluded as outliers because their science learning scores were 1.5 times the interquartile range lower than the first quartile of their group. The 'lavaan' package in R was used to perform the analysis.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.01$.

**Table 4.**

Regression on Overall Science Learning with Interaction Terms

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Pseudo-interactive | 0.287 | -0.002 | -0.034 |
| | (0.663) | (0.175) | (0.238) |
| Non-interactive | -0.823 | -0.422* | -0.470* |
| | (0.559) | (0.165) | (0.219) |
| Age | 0.140* | 0.141** | 0.148** |
| | (0.071) | (0.052) | (0.052) |
| Science knowledge (LENS) | 0.368*** | 0.383*** | 0.372*** |
| | (0.044) | (0.070) | (0.044) |
| English proficiency | 0.039*** | 0.039*** | 0.038*** |
| | (0.007) | (0.007) | (0.010) |
| Spanish proficiency | 0.001 | -0.000 | -0.000 |
| | (0.005) | (0.005) | (0.005) |
| Age x pseudo-interactive | -0.071 | | |
| | (0.105) | | |
| Age x non-interactive | 0.081 | | |
| | (0.089) | | |
| LENS x pseudo-interactive | | -0.082 | |
| | | (0.082) | |
| LENS x non-interactive | | 0.056 | |
| | | (0.078) | |
| Eng. proficiency x pseudo-interactive | | | -0.007 |
| | | | (0.012) |
| Eng. proficiency x non-interactive | | | 0.008 |
| | | | (0.012) |
| Intercept | -2.044*** | -2.079*** | -2.083*** |
| Observations | 242 | 242 | 242 |
| R-squared | 0.68 | 0.68 | 0.68 |

*Note*: Standard error in parentheses. The interactive condition was set as the reference group. The dependent variable, children’s overall science learning scores, were standardized, so the mean was zero and the standard deviation was one. Full information maximum likelihood (FIML) was employed to compute missing data. Four children, two in the pseudo-interactive condition, and two in the interactive condition, were excluded as outliers because their science learning scores were 1.5 times the interquartile range lower than the first quartile of their group. The ‘lavaan’ package in R was used to perform the analysis.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.01$.

**Table 5.**
Descriptive Statistics of Children's Response Quantity

| | Pseudo-interactive | Interactive | ANOVA |
|---|---|---|---|
| **Response rate** | | | *F*(1, 157[1])= 5.32, *p* = .02* |
| Initial response | | | |
| *M* (SD) | 0.69 (0.46) | 0.78 (0.41) | |
| Median [min, max] | 1.00 [0, 1.00] | 1.00 [0, 1.00] | |
| Follow-up response | | | *F*(1, 157[1])= 10.17, *p* = .002** |
| *M* (SD) | | 0.81 (0.39) | |
| Median [min, max] | | 1.00 [0, 1.00] | |
| **Response length** | | | |
| Initial response | | | *F*(1, 157[1])= 0, *p* = .98 |
| *M* (SD) | 2.58 (3.66) | 2.58 (3.69) | |
| Median [min, max] | 1.00 [0, 25.00] | 1.00 [0, 45.00] | |
| Follow-up response | | | *F*(1, 157[1])= 1.73, *p* = .19 |
| *M* (SD) | | 2.98 (3.30) | |
| Median [min, max] | | 1.00 [0, 34.00] | |

*Note*: Non-responses were included for analysis

[1] Verbal engagement data were missing for one child.

Table 6
Regression on Verbal Engagement Quantity

| | Response rate | | Response length | |
|---|---|---|---|---|
| | Initial response | Follow-up response | Initial response | Follow-up response |
| Pseudo-interactive | -0.378* | -0.496** | -0.046 | -0.258 |
| | (0.155) | (0.154) | (0.161) | (0.160) |
| Age | 0.136 | 0.151 | -0.048 | -0.029 |
| | (0.123) | (0.122) | (0.128) | (0.128) |
| Science knowledge (LENS) | -0.043 | -0.046 | -0.088 | -0.086 |
| | (0.106) | (0.105) | (0.110) | (0.110) |
| English proficiency | 0.053** | 0.050** | 0.054** | 0.052** |
| | (0.016) | (0.016) | (0.017) | (0.017) |
| Spanish proficiency | -0.017 | -0.016 | -0.004 | -0.005 |
| | (0.013) | (0.013) | (0.013) | (0.013) |
| Intercept | -1.367* | -1.351* | -0.415 | -0.400 |
| | (0.662) | (0.656) | (0.686) | (0.684) |
| Observations | 148 | 148 | 148 | 148 |
| R-squared | 0.168 | 0.183 | 0.084 | 0.095 |

*Notes:* Standard error in parentheses. The interactive condition was set as the reference group. The dependent variables, children's response rate and length, were standardized, so the mean was zero and the standard deviation was one. Verbal engagement data were missing for three children. Nine children had missing values in their baseline measures. Standard errors in parentheses. **p*<0.05, ***p*<0.01, ***p*<0.001

**Table 7.**

Multilevel logistic regression on Children's Response Rate to Initial Questions

| Fixed effects | Estimates ($\beta$) | Std. Error | z-value |
|---|---|---|---|
| Intercept | 1.954*** | 0.284 | 6.868 |
| Question number | 0.049*** | 0.011 | 4.528 |
| Pseudo-interactive | -0.618 | 0.397 | -1.557 |
| Question number x pseudo-interactive | -0.054*** | 0.014 | -3.777 |
| Random effects | Variance | Std. Dev. | |
| Intercept | 4.645 | 2.155 | |
| Number of observations | 5001 | | |
| Number of groups | 159[1] | | |

Note: The interactive condition was set as the reference group. Each child constituted a random effects group with random intercepts included in the model. The 'glmer' function from the 'lme4' package in R was used to perform the analysis. *p<0.05, **p<0.01, ***p<0.001

[1] Verbal engagement data were missing for one child.

**Table 8.**
Descriptive Statistics of Children's Response Quality

| | | Pseudo-interactive | Interactive | ANOVA |
|---|---|---|---|---|
| Response accuracy | | | | |
| Initial response | M (SD) | 0.41 (0.49) | 0.45 (0.50) | $F(1, 157^1) = 0.86, p = .36$ |
| | Median [min, max] | 0 [0, 1] | 0 [0, 1] | |
| Follow-up response | M (SD) | | 0.54 (0.50) | |
| | Median [min, max] | | 1 [0, 1] | |
| Response relevance | | | | |
| Initial response | M (SD) | 1.21 (0.96) | 1.34(0.91) | $F(1, 157^1) = 2.35, p = .13$ |
| | Median [min, max] | 2 [0, 2] | 2 [0, 2] | |
| Follow-up response | M (SD) | | 1.44 (0.88) | |
| | Median [min, max] | | 2 [0, 2] | |

*Note:* Pseudo-interactive videos did not provide hints to children when they failed to answer the initial questions, therefore, non-responses were included for analysis and coded as 0 for both accuracy and relevance.
[1]Verbal engagement data were missing for one child.

**Table 9.**

Regression on Verbal Engagement Quantity

| | Response Accuracy | | Response Relevance | |
|---|---|---|---|---|
| | Initial response | Follow-up response | Initial response | Follow-up response |
| Pseudo-interactive | -0.140 | -0.464*** | -0.243 | -0.402** |
| | (0.133) | (0.130) | (0.147) | (0.145) |
| Age | 0.148 | 0.157 | 0.206 | 0.202 |
| | (0.106) | (0.103) | (0.117) | (0.116) |
| Science knowledge (LENS) | 0.214* | 0.169 | 0.028 | 0.011 |
| | (0.091) | (0.089) | (0.100) | (0.099) |
| English proficiency | 0.052*** | 0.054*** | 0.057*** | 0.056*** |
| | (0.014) | (0.014) | (0.015) | (0.015) |
| Spanish proficiency | -0.002 | -0.001 | -0.014 | -0.013 |
| | (0.011) | (0.011) | (0.012) | (0.012) |
| Intercept | -2.168*** | -2.027*** | -2.096** | -1.976** |
| | (0.567) | (0.554) | (0.626) | (0.620) |
| Observations | 148 | 148 | 148 | 148 |
| R-squared | 0.376 | 0.394 | 0.250 | 0.259 |

Note: Standard error in parentheses. The interactive condition was set as the reference group. The dependent variable, children's response accuracy and relevance at initial attempt or with hints, were standardized, so the mean was zero and the standard deviation was one. Verbal engagement data were missing for three children. Nine children had missing values in their baseline measures. Standard errors in parentheses. *$p$<0.05, **$p$<0.01, **$p$<0.001

**Table 10**

The Correlation between Verbal Engagement and Science Learning by Condition

| | Science Learning<br>in Pseudo-interactive Condition | Science Learning<br>in Interactive Condition |
|---|---|---|
| Response rate | | |
| Initial response | .19 | .53*** |
| Follow-up response | | .50*** |
| Response length | | |
| Initial response | .26* | .27* |
| Follow-up response | | .28* |
| Response relevance | | |
| Initial response | 0.29* | 0.62*** |
| Follow-up response | | 0.60*** |
| Response accuracy | | |
| Initial response | 0.45*** | 0.65*** |
| Follow-up response | | 0.64*** |

*Note*: Non-responses were included for analysis. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.01$.

**Table 11**

Regression of Verbal Engagement on Overall Science Learning by Condition with Interaction Terms

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 |
|---|---|---|---|---|---|---|---|---|
| Pseudo-interactive | -0.098 | -0.085 | -0.128 | -0.112 | -0.118 | -0.085 | -0.108 | -0.092 |
| | (0.089) | (0.091) | (0.089) | (0.090) | (0.089) | (0.092) | (0.088) | (0.090) |
| Response rate (initial) | 0.239** | | | | | | | |
| | (0.076) | | | | | | | |
| Response rate (follow-up) | | 0.243** | | | | | | |
| | | (0.078) | | | | | | |
| Response length (initial) | | | 0.142 | | | | | |
| | | | (0.072) | | | | | |
| Response length (follow-up) | | | | 0.147* | | | | |
| | | | | (0.068) | | | | |
| Response accuracy (initial) | | | | | 0.179* | | | |
| | | | | | (0.074) | | | |
| Response accuracy (follow-up) | | | | | | 0.190* | | |
| | | | | | | (0.073) | | |
| Response relevance (initial) | | | | | | | 0.233** | |
| | | | | | | | (0.075) | |
| Response relevance (follow-up) | | | | | | | | 0.229** |
| | | | | | | | | (0.075) |
| ***Interaction terms*** | | | | | | | | |
| Initial rate x pseudo-int. | -0.280** | | | | | | | |
| | (0.093) | | | | | | | |
| Rate with follow-up x pseudo-int. | | -0.279** | | | | | | |
| | | (0.095) | | | | | | |
| Initial length x pseudo-int. | | | -0.146 | | | | | |
| | | | (0.092) | | | | | |
| Length with follow-up x pseudo-int. | | | | -0.151 | | | | |
| | | | | (0.091) | | | | |
| Initial accuracy x pseudo-int. | | | | | -0.160 | | | |
| | | | | | (0.090) | | | |
| Accuracy with follow-up x pseudo-int. | | | | | | -0.168 | | |
| | | | | | | (0.093) | | |
| Initial relevance x pseudo-int. | | | | | | | -0.247** | |
| | | | | | | | (0.092) | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Relevance with follow-up x pseudo-int. | | | | | | | | -0.245** |
| | | | | | | | | (0.093) |
| ***Covariates*** | | | | | | | | |
| Age | 0.095 | 0.091 | 0.136 | 0.133 | 0.119 | 0.115 | 0.091 | 0.093 |
| | (0.071) | (0.072) | (0.072) | (0.072) | (0.072) | (0.071) | (0.072) | (0.072) |
| Science knowledge (LENS) | 0.299*** | 0.303*** | 0.314*** | 0.314*** | 0.283*** | 0.289*** | 0.291*** | 0.295*** |
| | (0.060) | (0.060) | (0.061) | (0.061) | (0.062) | (0.061) | (0.060) | (0.060) |
| Eng. Proficiency | 0.050*** | 0.050*** | 0.049*** | 0.049*** | 0.046*** | 0.045*** | 0.048*** | 0.048*** |
| | (0.009) | (0.009) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) |
| Spa. Proficiency | -0.006 | -0.005 | -0.006 | -0.006 | -0.006 | -0.006 | -0.006 | -0.006 |
| | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) |
| Intercept | -1.842*** | -1.847*** | -2.068*** | -2.069*** | -1.856*** | -1.857*** | -1.746*** | -1.779*** |
| | (0.386) | (0.387) | (0.385) | (0.383) | (0.402) | (0.398) | (0.398) | (0.396) |
| Observations | 148 | 148 | 148 | 148 | 148 | 148 | 148 | 148 |
| R-squared | 0.647 | 0.647 | 0.631 | 0.633 | 0.636 | 0.638 | 0.646 | 0.645 |

Notes: Standard error in parentheses. The interactive condition was set as the reference group. The dependent variables were standardized, so the mean was zero and the standard deviation was one. Verbal engagement data were missing for three children. Nine children had missing values in their baseline measures. Standard errors in parentheses. *p<0.05, **p<0.01, **p<0.001

**Figure 1**

*The Dialogue Flow Design of the Interactive Media Character, Elinor*

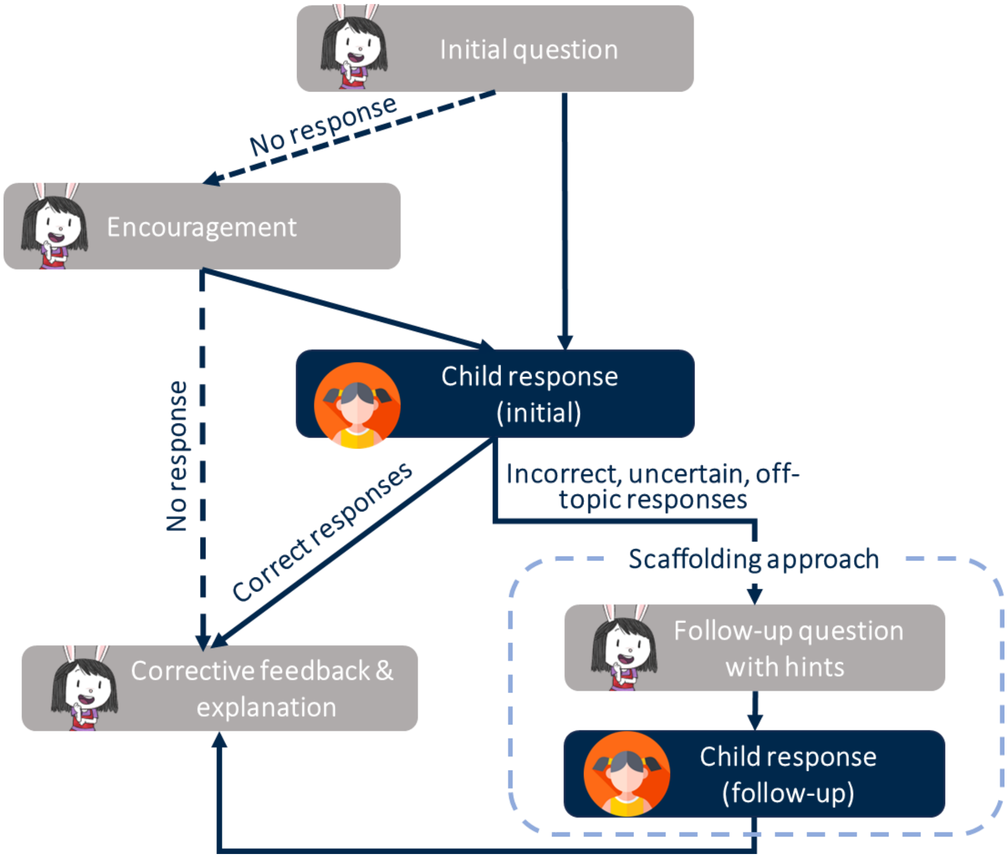
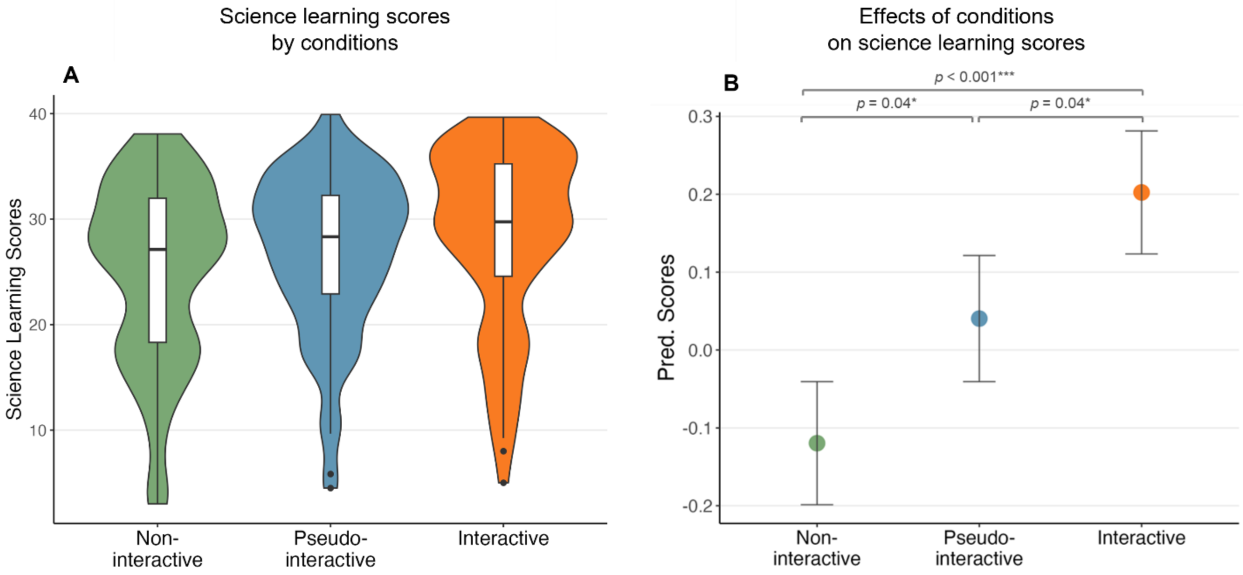

Initial question
No response
Encouragement
Child response (initial)
No response
Correct responses
Incorrect, uncertain, off-topic responses
Scaffolding approach
Follow-up question with hints
Child response (follow-up)
Corrective feedback & explanation

**Figure 2**

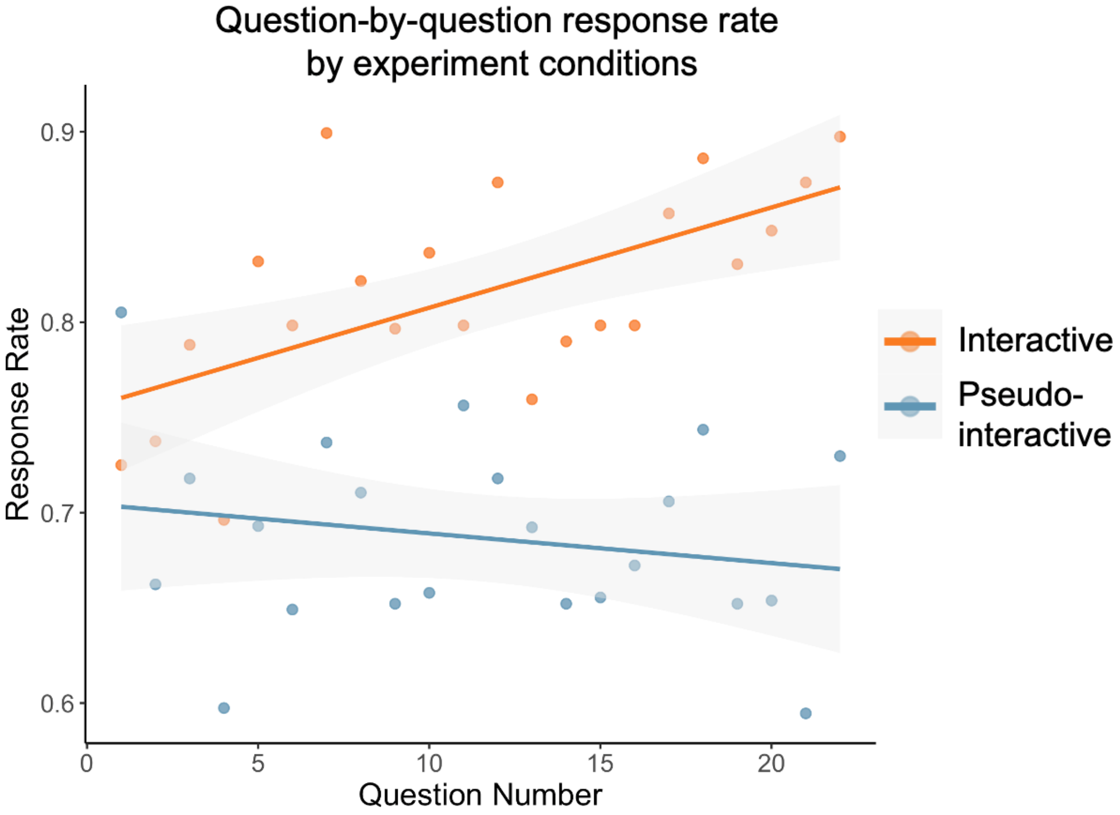
*Descriptive Statistics and Regression Estimates of Science Learning Outcomes by Conditions*



*Note:* Panel A displays the descriptive statistics of the overall science learning outcomes by condition. Panel B displays the predicted mean scores adjusted for control variables in the regression model for each condition.

**Figure 3**

*Comparison of Response Rates by Question*



*Note:* Each data point on the graph signifies the mean response rate for an individual initial question. Trend lines represent the fitted average response rates across the sequence of questions for each experimental condition.

## Appendix A

*Examples of children's conversations with Elinor*

**1. When a child provided a correct response.**

Elinor: We made a really fast car! It's so fun to try different things and figure out what works and what doesn't work to make the car go faster [giggle]. What did we change to make our car go faster?

Child: Shape.

Elinor: You are right! The pointy shape on the front made the car streamlined, so it goes faster.

**2. When a child provided an incorrect response**

Elinor: Ari thinks that changing the color of our car will make it as fast as Camila's car. What can we do to test if changing the color can make our car go faster?

Child: Change the engine

Elinor: Hmm, we want to figure out if changing the color will make our car go as fast as Camilar's car! To test this idea, should we change the color of the car or should we add a cupholder to it?

Child: Change color

Elinor: Yeah, that's exactly what we're going to do! We're going to paint our car yellow to see if it goes faster.

**3. When a child provided an off-topic response.**

Elinor: I learn so much from my Dad! What moves through the air slower? Boxy things or pointy things?

Child: We need to have wings

Elinor: Think about what my dad said! The air pushes against boxy things, but moves around pointy things. So which one does the air slow down more, boxy things or pointy things?

Child: I don't know.

Elinor: My dad said that the air pushes against boxy things and slows them down.

**4. When a child did not verbally respond**

Elinor: Did you ever ride in a car going really fast?

Child: (Silence)

Elinor: Hmm, I didn't quite hear what you said! Did you ever ride in a car going really fast?

Child: Yeah.

Elinor: Yeah, that's so cool! Let's see if we can build a car that goes fast!

**Appendix B.**

*Quality of Verbal Engagement Outcomes with Non-Responses Excluded*

| | Pseudo-interactive | Interactive |
|---|---|---|
| Response relevance | | |
| *M* (*SD*) | 1.78 (0.23) | 1.75 (0.25) |
| Median [min, max] | 1.83 [1.00, 2,00] | 1.84 [1.00, 2.00] |
| Response accuracy | | |
| *M* (*SD*) | 0.58 (0.23) | 0.58 (0.22) |
| Median [min, max] | 0.59 [0, 1.00] | 0.61 [0, 1.00] |