The Impact of Multi-Character Story Distribution and Gesture on Children's Engagement

Harrison Jesse Smith[1], Brian K. Riley[1], Lena Reed[2], Vrindavan Harrison[2],
Marilyn Walker[2], Michael Neff[1]

[1] University of California, Davis, Davis CA 95616
{hjsmith,bkriley, mpneff}@ucdavis.edu
[2] University of California, Santa Cruz, Santa Cruz CA 95064
{lireed,vharriso, mawalker}@ucsc.edu

Abstract. Effective storytelling relies on engagement and interaction. This work develops an automated software platform for telling stories to children and investigates the impact of two design choices on children's engagement and willingness to interact with the system: story distribution and the use of complex gesture. A storyteller condition compares stories told in a third person, narrator voice with those distributed between a narrator and first-person story characters. Basic gestures are used in all our storytellings, but, in a second factor, some are augmented with gestures that indicate conversational turn changes, references to other characters and prompt children to ask questions. An analysis of eye gaze indicates that children attend more to the story when a distributed storytelling model is used. Gesture prompts appear to encourage children to ask questions, something that children did, but at a relatively low rate. Interestingly, the children most frequently asked "why" questions. Gaze switching happened more quickly when the story characters began to speak than for narrator turns. These results have implications for future agent-based storytelling system research.

Keywords: Embodied Storyteller · Listening Comprehension · Primary School · Case Study.

1 Introduction

For many, being read a bedtime story is a fond childhood memory. This comforting experience also creates excitement as the the new world of the story unfolds. While enjoyable, such storytelling also provides the foundation for developing listening comprehension and later reading comprehension skills [28, 19, 20, 38, 48, 16, 59, 63] which are critical to educational attainment.

While the levels of quality home language input low socioeconomic status (SES) children receive is debated [25, 57], it is clear that the absence of such language input can negatively affect a child's early language development skills [26, 27, 8, 14, 4, 35, 44, 64, 67, 31, 56]. If children do not have adequate language skills in the primary grades, they are likely to have persistent academic difficulties [30, 56, 62], leading to long lasting consequences [52].

Computer storytelling apps may provide a way to address this early exposure gap and remediate, at least in part, the early educational deficit by providing high quality language exposure at home or in the classroom. They can be displayed on phones or tablets and deployed at low cost. It remains unclear, however, how to effectively design these apps in order to maximize child engagement.

To help answer this question, we present the results of experiment that looks at two factors in story presentation. The first compares narrator-only storytellings (third-person) with tellings that distribute the text between a narrator and story characters (first and third-person). The second factor varies the amount of nonverbal behavior present in the characters, comparing a condition that only uses beat gestures and subtle head nods with a condition that includes character deixis gestures, turn taking cues, and interaction prompts (see 2.3 for gesture type definitions).

To investigate these factors, we used a custom-built Unity application and cloud-based text-to-speech software to present four Aesop's fables in a repeatable, controlled fashion. The storytelling application is shown in Fig. 1. During the story presentation, we recorded participants' gaze locations. After each story concluded, the system solicited and recorded questions asked by the participant.

The experiment was run as a 2x2, within-subjects study focused on children aged 5-8. Results indicate that a multi-character, distributed telling of the story is more engaging than a narrator-only telling, based on gaze behavior. The impact of nonverbal communication appears complicated, as the additional animation of a conversational turn handover can hold student attention, rather than directing it at the intended target. However, there is some evidence that question prompting gestures can help elicit feedback from children. Question elicitation at the end of the story resulted in questions 20% of the time, most of which (70%) where different types of *why* questions.

The results reported in this paper have implications for future automated and/or interactive storytelling applications. They suggest that presenting stories from multiple characters' first-person points of view is an effective way to increase student engagement. While question-prompting gestures may be a useful way to ellicit questions from students, it is unclear whether nonverbal turn-over gestures are an effective method for signalling the next speaker to children aged 5-8. The frequency and types of questions asked is useful for developing a conversational storytelling framework, which is a long-term aim of this project.

The contributions of this work are as follow:

- We show that a distributed storytelling model results in significantly higher engagement than a narrator-only model.
- We present data suggesting that question-prompting gestures may be effective for eliciting questions from children.
- We present the frequency and types of questions asked by children to an automated storytelling application.

- We present data showing that nonverbal turnover gestures may not be an effective method of signalling the next speaker to children aged 5-8.
- We demonstrate a webcam-based method for collecting gaze data, useful in certain experimental settings.

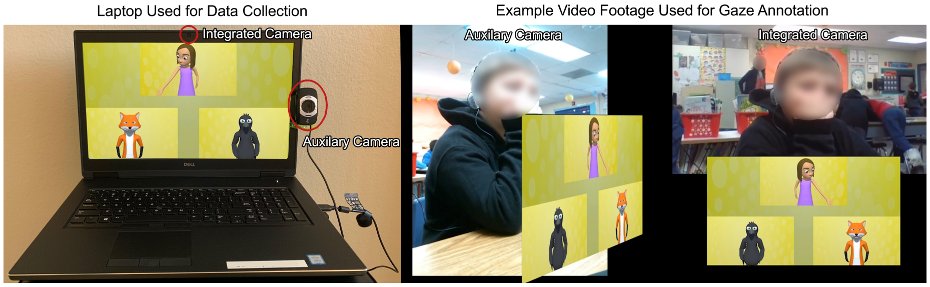

Laptop Used for Data Collection
Integrated Camera
Auxilary Camera
Example Video Footage Used for Gaze Annotation
Auxilary Camera
Integrated Camera

Fig. 1. Left: The laptop and camera placement used to collect video footage of the participants. Right: The video footage used to extract gaze targets. Inverted screen capture overlays are added in a post-processing step to provide additional context for the annotators.

2 Background

2.1 Automated Storytelling and Editing

Recognizing the importance of reading and storytelling for children's development, related work has also focused on improving children's reading skills. Project LISTEN was one of the first systems in this area: it aimed at computer tutors that could listen to a child read aloud and provide help where needed with pronunciation and other types of reading aloud errors [47, 2]. Other work has focused on virtual peers for pedagogical purposes, and tested the effect of having the peer model more advanced storytelling behaviors [15, 10, 55]. Storytelling agents have also been explored with robots as reading companions and tutors [17, 36, 65, 49], including studies placing robots in classrooms over extended period of time [32].

An automated storyteller could potentially tailor the text of a story based on the needs of the current user. Adjusting vocabulary level or narration point-of-view could result in a more effective and rewarding storytelling session. Due to the complexities of natural language, however, it is challenging to create robust methods for automatically editing text in non-trivial ways. Several researchers in the field of natural language processing have focused efforts on this problem. One set of researchers presented methods for automatically generating dialogues from monologues [51]. In another work, the author presented methods for generating full virtual-agent based performances, given only input text [50].

2.2 Eye Tracking in Multimedia Learning

Multimedia learning materials, which distribute conveyed information across multiple visual and/or audio channels, are widely used and are an effective way to foster meaningful learning outcomes in students [42]. In the past, the efficacy of such materials were commonly assessed using post-intervention interviews and behavioral assessments [53]. While such techniques are useful to measure the overall learning outcomes induced by the materials, they do not provide the resolution necessary to link detailed behaviors of a participant to on-screen causal elements [43]. Such linkages, and the insights they can provide, may aid in the creation of valuable design principles for different categories of multimedia applications.

An alternative to post-assessments is tracking participant gaze behavior. It is a useful measure for understanding how a viewer allocates their visual attention and how this engagement temporally fluctuates as a function of on-screen events [29]. Analyzing such engagement is particularly useful when developing design guidelines for interactive storytelling applications, whose primary purpose is fostering listening comprehension in the viewer. Such applications should engage the viewer without resorting to seductive details (motions or other stimuli that are pleasant but distracting, and which do not further comprehension).

While interest in eye tracking has increased rapidly in recent years [3], relatively few researchers have studied the eye movement of early grade school students interacting with multimedia stimuli [46, 60] . Neither study reported eye tracking movement of students observing the stimuli in an in-use classroom. This may be due to the chaotic nature of such classroom, the expense and sensitivity of eye-tracking software, and the difficulty of properly calibrating and controlling the behaviors of a young child during a sedentary experiment. In contrast, the current study focuses on engagement and attention of early grade school students within in-use classrooms and makes use of multiple web cameras to record gaze behaviors.

2.3 Gesture

To further engage the child, we will endow the child-like narrator with nonverbal communication behaviors, as endorsed by the PAL framework [34] and other related work on pedagogical agents and agent personality [11, 37, 41]. Studies of teacher communication have found a cluster of nonverbal behaviors that are particularly effective in the teaching context. Termed "immediacy", these factors generate positive affect and include eye contact, smiling, vocal expressiveness, physical proximity, leaning towards a person, using appropriate gestures and being relaxed [58, 5, 6, 33]. They are consistently shown to impact affective learning [54, 7, 18], which impacts students predisposition towards material and motivation to use knowledge [6, 9]. Their impact on cognitive learning is less clear, with mixed findings [18, 54]. Deictic (or pointing) gestures help ground the conversation by establishing shared reference [45] and can help children distinguish ambiguous speech sounds [61]. Speech that is accompanied by gesture leads to better recall than the same speech without gesture [13]. In teaching settings, gesture can provide a second representation, and multiple representations are known to enhance learning [23].

Beat gestures [45] are small, downward movements of the arms and hands that accompany the cadence of the speech and may add emphasis, but do not convey clear meaning. They are used in this work to make the characters appear more alive. Deictic gestures [45] are used to create reference, such as by pointing. Backchanneling, such as head nods and affirmative utterances, are used by the listener to signal their agreement with the speaker [66]. Conversational turn management in human dialog is largely nonverbal [66], motivating its use here.

3 Method

Participants. Participants from four K-2 classrooms in two schools in the United States participated in this experiment. Consent from school administration, classroom teachers, parents, and an institutional review board was obtained prior to the study. All participants spoke English and had normal or corrected-to-normal vision. In total, 33 participants, 12 girls and 21 boys, were included in the final analysis. Their ages ranged from 5-8 years old (M = 6.4, SD = 1.05).

Design. The study used a 2x2 experimental design in which every participant observed all four stimuli combinations. A within-subjects design was used to minimize sources of non task-related variance, such as participant's base attention spans or moods on the day of the experiment. The first factor was **Storytelling Perspective**, which employed a **Narrator Only** level and a **Distributed** level. The second factor was **Gesture Types**, which employed a **Complex Gesture** level and a **Simple Gesture** level. A single story was used for each condition combination (see Fig.2).

Materials. Four Aesop's Fables were selected for use in the experiment. Aesop's Fables are commonly used in studies on (oral) narrative comprehension and are often used in teaching materials for the K-2 age group. We selected four fables that could be animated using Narrator, Fox and Crow characters. These were *The Fox and the Grapes*, *The Fox and the Crow*, *The Dog and His Shadow* and *The Crow and the Pitcher*. The fable *The Dog and his Shadow* was converted to *The Fox and his Shadow* in order to use the Fox's character model and gestures.

The original text of the stories came from the versions of Aesop's Fables distributed as part of Elson's Drama Bank [21, 1]. For each story we produced (by hand) a version of the story with simpler sentences and simpler vocabulary: these story versions were double-checked by a learning scientist for their age appropriateness. Because all the original stories are presented in third person by a narrator, we used the Fabula tales natural language generation engine to generate first person direct versions of story sentences for half of the stories [39, 40].

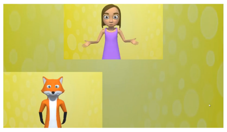
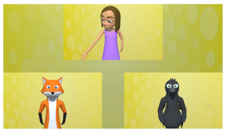
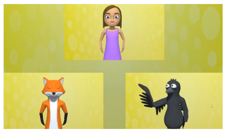
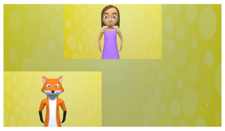
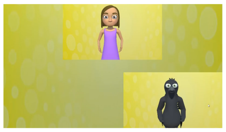
	Condition	Story Name	Example Images	Story Text
a	Narrator Only Complex Gesture	The Fox and the Grapes		Narrator: A hungry Fox saw bunches of tasty looking grapes hanging from a vine. The vine was high up in the air. The Fox tried really hard to reach them by jumping as high as he could into the air. But the fox couldn't reach the grapes. He gave up and walked away. He acted like he didn't care. The Fox said, "I thought those grapes would be tasty, but now I see that they are sour."
b	Distributed Complex Gesture	The Fox And the Crow	 	Narrator: A Crow was sitting on a branch of a tree. She had a piece of cheese in her beak when a Fox saw her. Fox: That cheese looks good! I want to get the cheese from that Crow. Narrator: The Fox stood under the tree and looked up. Fox: What a beautiful bird I see! Nobody is as beautiful as she is, the color of her feathers are amazing. If her voice is as sweet as she is pretty, than she is definitely the Queen of the Birds Narrator: The Crow was very happy with what the Fox said. She wanted to show the Fox that she could sing, so she made a loud caw. Crow: Caw! Caw! Caw! Oh no, the cheese has fallen out of my beak! Narrator: The Fox snatched it up. Fox: You can sing, madam, but you are not smart
c	Narrator Only Simple Gesture	The Fox And His Shadow		Narrator: A Fox crossed a bridge over a stream. He had a piece of meat in his mouth. While walking he saw his own reflection in the water. He thought it was another fox who had a piece of meat twice as big. He let go of his own and jumped at the other fox to get the larger piece. What happened was that he got neither pieces of meat. One was only a reflection and his piece was carried away by the stream.
d	Distributed Simple Gesture	The Crow and the Pitcher		Narrator: A crow was thirsty. Crow: I am so thirsty. Narrator: She found a glass of water, but there was only a little water in it. Crow: My beak can't reach the water, it's too far down. I'm scared that I will die of thirst. Narrator: She came up with a plan to get the water. Crow: I'm going to drop rocks into the glass. Narrator: The rocks made the water rise to the top of the glass. Crow: I can reach the water! I'm so glad I can get a drink.

Fig. 2. Overview of the conditions, along with story names, example images, and story text. Example image in row A shows the *Question Gesture* and example images in row B show *Nonverbal Turnover Gestures*.

Fig. 2 provides examples of how each story was told and how the content was distributed amongst the characters. In the **Narrator Only** condition the Narrator recounted the entire story (see Rows a and c of Fig. 2). The Narrator refers to the story characters in third person, and all utterances and gestures are produced by the Narrator. The story characters appear on the screen but do not speak.

The first person, direct speech, versions of the stories are used in the **Distributed** condition, and thus the story telling is split between the onscreen characters (Rows b and d of Fig. 2). The Narrator only produces the utterances that describe actions. Utterances that provide content for character speech and thought are converted to first person direct speech and spoken by the character to whom the speech or thought is attributed, e.g. *What a beautiful bird I see! Nobody is as beautiful …* in Row b of Fig. 2.

While all stories employed character blinks, idle breathing motions, and minor head/arm beat gestures, the **Complex Gesture** condition included three different types of gestures not present in the **Simple Gesture** condition: question prompt gestures, deictic gestures, and nonverbal turnover gestures. Question prompt gestures (see example image of Fig.2-a) were performed by the Narrator while she verbally prompted the participants for questions about the story (*"Now tell me, do you have any questions about the story?"*); in the **Simple Gesture** condition, the Narrator only verbally prompted the participants. In the **Narrator Only, Complex Gesture** condition, the Narrator used two deictic gestures, pointing towards the Fox, while verbally referring to him. The form of this gesture was identical to the *nonverbal turnover gesture* demonstrated by the Narrator in Fig.2-b.

In the **Distributed, Complex Gesture** condition, characters performed conversational turnover gestures after they finished speaking, visually indicating which character would speak next (see example images in Fig.2-b). In all stories there was a pause of 1.2 seconds between when one character stopped speaking and the next character began. When present, the conversational turnover gestures began as the character finished talking and took 0.75 seconds, leaving 0.5 seconds before the next character began to speak.

Stories were presented using a custom-built Unity application. The characters, story text, and gestures were provided as input to the system. AWS Polly Text-to-Speech was used to obtain speech audio and the viseme information necessary to drive character lip syncing behavior. At the end of each story, the Narrator would prompt the participant for questions about the story. During this period, the researcher used an external keyboard to control the Narrator in a *Wizard of Oz* fashion, triggering verbal and nonverbal backchanneling behaviors. After the child was finished asking questions, the researcher initiated the next story.

Procedure. Stimuli were shown on a Dell Precision laptop with 17 inch screen in a partially secluded classroom corner. Despite this separation, other students would sometimes distract the participant with their presence, actions, and noises. This environment therefore contained the same types of distractions that a child would experience while reading or working in school.

Upon starting the experiment, each participant watched an introductory segment in which the Narrator introduced herself, explained that she would be telling stories, and invited the participant to ask questions at the end of each story. Then all four stories were shown sequentially. Order was randomized to control for ordering effects. At the end of each story, the Narrator prompted the participant to ask any questions they had about about the story. The entire procedure took, on average, 3.5 minutes. For an example screen recording showing the experimental stimuli presented to participants, please visit the following link: *https://youtu.be/HEeQica-xHY*.

Measures. Due to the in-classroom nature of our experiment, expensive, sensitive eye-tracking hardware was avoided. Rather, two webcams were positioned around the perimeter of the laptop screen to record the gaze behaviors of the participant (see Fig.1-Left) for post-hoc annotation. Simultaneously, Open Broadcast Studio was used to record the contents of the screen. Taken together, this information was sufficient to determine when a participant was looking at the stimuli and at which character they were looking. See Fig.1-Right for an example of the resulting video. The webcams also captured the questions each participant asked at the conclusion of each story.

Gaze Annotation. Two undergraduate annotators were hired to annotate gaze behaviors and transcribe the utterances of each participant. Based on the synced screen recording and dual webcam footage, annotators identified the participant's area of focus throughout the duration of the experiment by labeling it with one of four categories: *Narrator*, *Fox*, *Crow*, and *Non-Task*. *Non-Task* was used when the participant was not looking at any of the characters on the screen.

The data from one participant was used to train the annotators; both annotators, along with the lead researcher, collectively discussed and annotated the gaze behavior. Next, data from six participants (21 minutes, 19% of the remaining data) was independently annotated by each annotator. Inter-rater reliability was very high (observed agreement was 97% and Cohen's kappa was 0.93), so data from the remaining 26 participants was split between the annotators.

4 Results

4.1 Visual Attention

Attention To Story. Using the gaze annotations, it was possible to determine the percentage of time participants were actively observing each story (viewing a character versus viewing a *Non-Task* category). These are shown in Fig.3. Summary attention statistics are given in Table 1.

To assess whether attention differed significantly as a function of condition, we conducted a Friedman test of differences using the single factor of 'Condition' with four levels. While a repeated measures ANOVA is commonly used in 2x2 within-factors designs, the percentage values analyzed were not normally distributed, and thus the non-parametric Friedman test was used instead. The test rendered a Chi-square value of 9.13, which was significant (p=0.02). Post-hoc analysis using multiple Wilcox signed-rank tests with Bonferroni correction revealed multiple significant differences (Fig.2, left). **Distributed, Complex Gesture** was significantly higher than both **Narrator Only, Simple Gesture** ($p_{adj} < 0.01$) and **Narrator Only, Complex Gesture** ($p_{adj} = 0.02$). **Distributed, Simple Gesture** was significantly higher than **Narrator Only, Simple Gesture** ($p_{adj} < 0.01$) and almost significantly higher than **Narrator Only, Complex Gesture** ($p_{adj} = 0.09$). Other differences were not significant.

Using the same technique, we evaluated the effect of order on attention (Fig.2, right). As might be expected, attention wanes over time. Attention to the first story was significantly higher than to the third ($p_{adj} = 0.008$) and fourth ($p_{adj} = 0.006$) story, and marginally higher than to the second story ($p_{adj} = 0.10$).

Table 1. Left: Summary statistics on the amount of attention paid to each story as a function of condition. Right: Summary statistics on the amount of attention paid to each story as a function of story order.

	Condition				Order			
	Narrator, Complex	**Narrator, Simple**	**Distributed, Complex**	**Distributed, Simple**	**1**	**2**	**3**	**4**
Mean	78.6%	79.5%	91.7%	89.9%	91.4%	85.2%	82.6%	80.5%
Standard Deviation	19.3	21.8	8.0	7.9	10.8%	13.5%	18.6%	19.9%

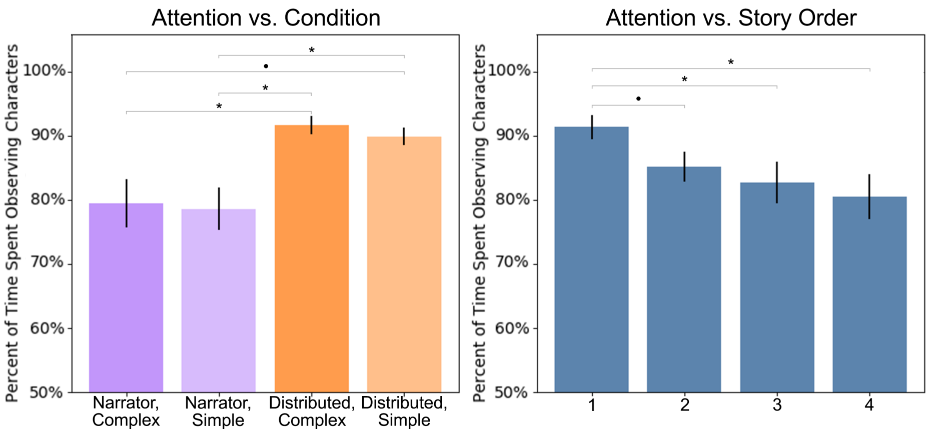


Fig. 3. The percentage of time students gazed at the *Narrator*, *Fox*, or *Crow* (as opposed to *Non-Task*) as a function of story condition. Error bars indicate standard error of the mean. Results significant at $p_{adj} < 0.05$ denoted by asterisk, result approaching significance at this level denoted by dot.

Gaze Behavior during Conversational Turnovers. We next used gaze information to evaluate differences in the amount of time it took participants to focus on the next speaker after a conversational turnover. Because these turnovers only occur when two or more characters take turns speaking, this analysis was conducted only on data obtained from the **Distributed** conditions.

For each conversational turnover, we determined the time at which the new character began to speak. We then calculated, relative to this point, the amount of time it took each participant to first glance at the new speaker. This value was positive if the speaker began talking before the participant looked to them and negative if the participant looked to the speaker before they began to speak.

Using these values, an independent samples t-test was conducted to compare the differences in gaze switching time between the **Distributed, Complex Gesture** condition and the **Distributed, Simple Gesture** condition. The results are shown in Table 2, top. There was a significant difference between these two conditions, with participants taking longer to switch their gaze to the new character in the **Distributed, Complex Gesture** condition.

Table 2. Summary statistics of the amount of time, in seconds, it took participants to switch their gaze to a new speaker after that speaker first began their conversational turn. The top row compares instances in which turnover gestures were present to instances in which the gesture was absent. The bottom row compares turnovers to the Narrator with turnovers to the Fox or the Crow.

	Mean	**Standard Deviation**	**P Value**	**T Statistic**
Distributed, Complex Gesture	0.71	1.47	0.004	2.88
Distributed, Simple Gesture	0.19	1.86		
Turnover to Narrator	1.15	1.64	<0.001	7.72
Turnover to Fox or Crow	-0.08	1.55		

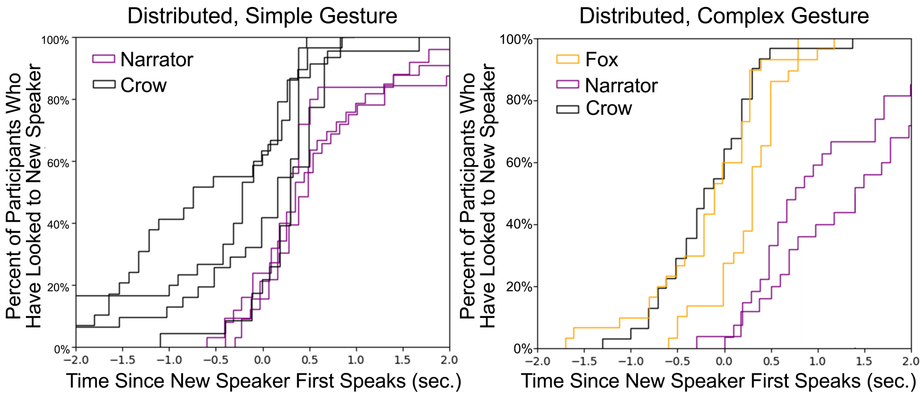


Fig. 4. Cumulative distributions functions showing the percentage of participants who looked to the next speaker relative to when the speaker began talking. Each line indicates a single conversational turnover from the story.

Table 3. Count of questions asked, separated by **Gesture** condition level.

	Question Asked	No Question Asked	P Value	Chi Squared Statistic
Complex Gesture	18	48	0.052	3.77
Simple Gesture	9	57		

After visual inspection of the cumulative distributions of gaze switching vs. time (as shown in Fig.4), it appeared that, when the Narrator took over speaking, participants turned their gaze back to her more slowly and less frequently that with the other two characters. We therefore conducted a t-test to determine if this difference was significant. The results are shown in Table 2, bottom. Participants took significantly longer to turn their gaze back to the speaker when the new speaker was the Narrator ($p < 0.001$). This could be because the participants were less interested in the Narrator (as they see her in every story), participants were more intrigued by the animal characters, and/or participants were more interested in the story characters.

4.2 Question Analysis

Question Frequency. In this study, we only elicited questions from participants at the end of the story. This protocol created 132 possible question opportunities and resulted in 27 questions. 17 participants asked no questions, eight asked one, six asked two, one asked three, and one asked four. To assess whether the **Complex Gesture** condition (and the question prompt gesture it contained) influenced participant's tendency to ask questions, we performed a chi-squared independence test. The results are given in Table 3. While the $p = 0.05$ level of significance was narrowly missed, this could be due to the small total number of questions collected. A larger sampling may reveal that the question prompt gesture is a clear visual indicator encouraging children to interact with the system.

Question Type. We conducted an analysis of the types of questions the children asked in order to determine the needed future capabilities of a conversational storytelling system that can answer questions as the story unfolds. We expected questions about comprehension, and two main types: (1) questions based on understanding the meaning of sentences, based on vocabulary or syntax within a sentence; (2) questions based on inferring causality, since that is a key part of understanding narrative [12, 22, 24]. Our goal is to support these kind of questions from students in a future version of our system, as well as to add question categories based on these to the narrator's repertoire. Examples are shown in Table 4. Q1 illustrates a comprehension question. We expect these would be more frequent if we allowed questions as the story unfolds. There were 19 *why* questions of different types. Questions Q2, Q3 and Q4 illustrate the 12 *why* questions related to causal understanding about how the world works or

Table 4. Example Questions from Participants

ID	N	Type	Example
Q1	1	Comprehension	What is a vine? (vocabulary)
Q2	12	Why, Causal Chain	Why did the Fox try to get the grapes? (hungry)
Q3			Why did the Fox get the cheese?
Q4			Why did she put rocks in the water? That sounds gross.
Q5	5	Why, BackStory	How did the Crow get the cheese?
Q6			Why was the Crow so thirsty?
Q7	5	What Next	Is he going to get the water?
Q8			The Fox will eat the bird?
Q9	2	Why, Storyline	Why wouldn't the Fox know that it was his reflection?
Q10			How is the Fox able to listen to the bird sing when a bird can only chirp?

failure to fill in implicit actions or state changes. Q2 illustrates a very simple causal inference: the Fox is described as *hungry* but two participants asked why the Fox wanted the grapes, while the others involve complex causal reasoning. The other question types target unexpected competencies that would be hard to support in our future conversational storyteller. Q5 and Q6 illustrate the 5 questions about the back-story, about how the situation came to be at the start of the narrative, which is not part of the story content. There were also 5 questions about what might happen in the story world after the end of the story (*What Next*): this is illustrated by questions Q7 and Q8. This could partly be due to the fact that we only asked questions at the end of the story. Finally, in Q9 and Q10 the participants question presuppositions of the story, i.e. that a Fox wouldn't recognize his reflection, and that birds can sing, rather than simply chirp.

5 Conclusion

The greater visual attention children paid to stories presented in first-person by story characters, in addition to the narrator, suggest that such distribution of storytelling may be an effective approach for building engagement. Gaze analysis also showed that children switched attention more quickly to story characters than to the narrator. The use of intentional gestures presents a mixed picture. It appears that gestures to the child are helpful in eliciting questions. Gestures for conversational turn management appeared to hold children's interest, rather than directing them to the next character to speak.

Children did ask questions of the system some of the time and these were frequently *why* questions. In future work we plan to elicit questions and ask questions **during** the storytelling at particular story points, rather than simply at the end of the story. We expect this to increase children's engagement with the story, and hopefully increase their narrative comprehension. We also wish to study deixis in cases where it is non-redundant with the text.

Acknowledgements: This work was supported through NSF grant IIS-1748058.

References

1. Aesop, Jones, V.S.V., Rackham, A., Chesterton, G.K.: Aesop's Fables: A New Translation. W. Heinemann (1933)
2. Aist, G., Kort, B., Reilly, R., Mostow, J., Picard, R.: Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: adding human-provided emotional scaffolding to an automated reading tutor that listens. In: Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on. pp. 483–490. IEEE (2002)
3. Alemdag, E., Cagiltay, K.: A systematic review of eye tracking research on multimedia learning. Computers & Education **125**, 413–428 (2018)
4. Alexander, K.L., Entwisle, D.R., Dauber, S.L.: First-grade classroom behavior: Its short-and long-term consequences for school performance. Child development **64**(3), 801–814 (1993)
5. Andersen, J.F.: The relationship between teacher immediacy and teaching effectiveness. Ph.D. thesis, ProQuest Information & Learning (1979)
6. Andersen, J.F.: Instructor nonverbal communication: Listening to our silent messages. New Directions for Teaching and Learning **1986**(26), 41–49 (1986)
7. Andersen, J.F., Norton, R.W., Nussbaum, J.F.: Three investigations exploring relationships between perceived teacher communication behaviors and student learning. Communication Education **30**(4), 377–392 (1981)
8. August, D., Hakuta, K.: Improving schooling for language minority students: A research agenda. Improving Schooling for Language Minority Students: A Research Agenda (1997)
9. Baylor, A.L.: Promoting motivation with virtual agents and avatars: role of visual presence and appearance. Philosophical Transactions of the Royal Society of London B: Biological Sciences **364**(1535), 3559–3565 (2009)
10. Baylor, A.L., Kim, Y.: Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In: Intelligent tutoring systems. vol. 3220, pp. 592–603. Springer (2004)
11. Baylor, A.L., Ryu, J.: The effects of image and animation in enhancing pedagogical agent persona. Journal of Educational Computing Research **28**(4), 373–394 (2003)
12. Bloom, C.P., Fletcher, C.R., Broek, P.V.D., Reitz, L., Shapiro, B.P.: An on-line assessment of causal reasoning during comprehension. Journal of Memory and Cognition (1990)
13. Breckinridge Church, R., Garber, P., Rogalski, K.: The role of gesture in memory and social communication. Gesture **7**(2), 137–158 (2007)
14. Brooks-Gunn, J., Duncan, G.J.: The effects of poverty on children. The future of children **7**, 55–71 (1997)
15. Cassell, J.: Towards a model of technology and literacy development: Story listening systems. Journal of Applied Developmental Psychology **25**(1), 75–105 (2004)
16. Catts, H.W., Adlof, S.M., Hogan, T.P., Weismer, S.E.: Are specific language impairment and dyslexia distinct disorders? Journal of Speech, Language, and Hearing Research **48**(6), 1378–1396 (2005)
17. Chang, A., Breazeal, C.: Tinkrbook: shared reading interfaces for storytelling. In: Proceedings of the 10th International Conference on Interaction Design and Children. pp. 145–148. ACM (2011)
18. Chesebro, J.L.: Effects of teacher clarity and nonverbal immediacy on student learning, receiver apprehension, and affect. Communication Education **52**(2), 135–147 (2003)

19. Cunningham, A.E., Stanovich, K.E.: Early reading acquisition and its relation to reading experience and ability 10 years later. Developmental psychology **33**(6), 934 (1997)
20. Duncan, G.J., Dowsett, C.J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., Pagani, L.S., Feinstein, L., Engel, M., Brooks-Gunn, J., et al.: School readiness and later achievement. Developmental psychology **43**(6), 1428 (2007)
21. Elson, D.: Dramabank: Annotating agency in narrative discourse. In: LREC. pp. 2813–2819 (2012)
22. Fletcher, C.R., Hummel, J.E., Marsolek, C.J.: Causality and the allocation of attention during comprehension. Journal of Experimental Psychology **16**(2), 233–140 (1990)
23. Goldin-Meadow, S., Singer, M.A.: From children's hands to adults' ears: gesture's role in the learning process. Developmental psychology **39**(3), 509 (2003)
24. Graesser, A.C., Singer, M., Trabasso, T.: Constructing inferences during narrative text comprehension. Psychological review **101**(3), 371 (1994)
25. Hart, B., Risley, T.R.: Meaningful differences in the everyday experience of young American children. Paul H Brookes Publishing (1995)
26. Hoff, E.: Environmental supports for language acquisition. Handbook of early literacy research **2**, 163–172 (2006)
27. Hoff, E., Naigles, L.: How children use input to acquire a lexicon. Child development **73**(2), 418–433 (2002)
28. Hoover, W.A., Gough, P.B.: The simple view of reading. Reading and writing **2**(2), 127–160 (1990)
29. Hyönä, J.: The use of eye movements in the study of multimedia learning. Learning and Instruction **20**(2), 172–176 (2010)
30. Juel, C.: Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of educational Psychology **80**(4), 437 (1988)
31. Juel, C., Griffith, P.L., Gough, P.B.: Acquisition of literacy: A longitudinal study of children in first and second grade. Journal of educational psychology **78**(4), 243 (1986)
32. Kanda, T., Hirano, T., Eaton, D., Ishiguro, H.: Interactive robots as social partners and peer tutors for children: A field trial. Human-computer interaction **19**(1), 61–84 (2004)
33. Kennedy, J., Baxter, P., Belpaeme, T.: Nonverbal immediacy as a characterisation of social behaviour for human–robot interaction. International Journal of Social Robotics **9**(1), 109–128 (2017)
34. Kim, Y., Baylor, A.L.: A social-cognitive framework for pedagogical agents as learning companions. Educational technology research and development **54**(6), 569–596 (2006)
35. Korenman, S., Miller, J.E., Sjaastad, J.E.: Long-term poverty and child development in the united states: Results from the nlsy. Children and Youth Services Review **17**(1-2), 127–155 (1995)
36. Kory, J., Breazeal, C.: Storytelling with robots: Learning companions for preschool children's language development. In: Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on. pp. 643–648. IEEE (2014)
37. Lee, K.M., Peng, W., Jin, S.A., Yan, C.: Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. Journal of communication **56**(4), 754–772 (2006)
38. Literacy, D.E.: Report of the national early literacy panel washington. DC National Institute for Literacy (2008)

39. Lukin, S.M., Reed, L.I., Walker, M.: Generating sentence planning variations for story telling. In: 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. p. 188 (2015)
40. Lukin, S.M., Walker, M.A.: A narrative sentence planner and structurer for domain independent, parameterizable storytelling. Dialogue & Discourse **10**(1), 34–86 (2019)
41. Mairesse, F., Walker, M.A.: Towards personality-based user adaptation: psychologically informed stylistic language generation. User Modeling and User-Adapted Interaction **20**(3), 227–278 (2010)
42. Mayer, R.E.: Cognitive theory of multimedia learning. The Cambridge handbook of multimedia learning **41**, 31–48 (2005)
43. Mayer, R.E.: Using multimedia for e-learning. Journal of Computer Assisted Learning **33**(5), 403–423 (2017)
44. McLoyd, V.C.: Socioeconomic disadvantage and child development. American psychologist **53**(2), 185 (1998)
45. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)
46. Molina, A.I., Navarro, s., Ortega, M., Lacruz, M.: Evaluating multimedia learning materials in primary education using eye tracking. Comput. Stand. Interfaces **59**(C), 45–60 (Aug 2018). https://doi.org/10.1016/j.csi.2018.02.004, https://doi.org/10.1016/j.csi.2018.02.004
47. Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M.B., Tobin, B.: Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. Journal of Educational Computing Research **29**(1), 61–117 (2003)
48. Network, N.E.C.C.R., et al.: Child care and child development: Results from the NICHD study of early child care and youth development. Guilford Press (2005)
49. Park, H.W., Gelsomini, M., Lee, J.J., Breazeal, C.: Telling stories to robots: The effect of backchanneling on a child's storytelling. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. pp. 100–108. ACM (2017)
50. Piwek, P., Hernault, H., Prendinger, H., Ishizuka, M.: T2d: Generating dialogues between virtual agents automatically from text. In: International Workshop on Intelligent Virtual Agents. pp. 161–174. Springer (2007)
51. Piwek, P., Stoyanchev, S.: Generating expository dialogue from monologue: motivation, corpus and preliminary rules. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 333–336. Association for Computational Linguistics (2010)
52. Ritchie, S.J., Bates, T.C.: Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. Psychological science **24**(7), 1301–1308 (2013)
53. Rodrigues, P., Rosa, P.J.: Eye-tracking as a research methodology in educational context: a spanning framework. In: Eye-tracking technology applications in educational research, pp. 1–26. IGI Global (2017)
54. Rodríguez, J.I., Plax, T.G., Kearney, P.: Clarifying the relationship between teacher nonverbal immediacy and student cognitive learning: Affective learning as the central causal mediator. Communication education **45**(4), 293–305 (1996)
55. Ryokai, K., Vaucelle, C., Cassell, J.: Virtual peers as partners in storytelling and literacy learning. Journal of computer assisted learning **19**(2), 195–208 (2003)

56. Snow, C.E., Burns, M.S., Griffin, P.: Preventing reading difficulties in young children committee on the prevention of reading difficulties in young children. Washington, DC: National Research Council (1998)
57. Sperry, D.E., Sperry, L.L., Miller, P.J.: Reexamining the verbal environments of children from different socioeconomic backgrounds. Child Development **90**(4), 1303–1318 (2019). https://doi.org/10.1111/cdev.13072, https://onlinelibrary.wiley.com/doi/abs/10.1111/cdev.13072
58. Staudte, M., Crocker, M.W., Heloir, A., Kipp, M.: The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. Cognition **133**(1), 317–328 (2014)
59. Storch, S.A., Whitehurst, G.J.: Oral language and code-related precursors to reading: evidence from a longitudinal structural model. Developmental psychology **38**(6), 934 (2002)
60. Takacs, Z.K., Bus, A.G.: Benefits of motion in animated storybooks for children's visual attention and story comprehension. an eye-tracking study. Frontiers in psychology **7**, 1591 (2016)
61. Thompson, L.A., Massaro, D.W.: Children' s integration of speech and pointing gestures in comprehension. Journal of Experimental Child Psychology **57**(3), 327–354 (1994)
62. Torgesen, J.K.: Avoiding the devastating downward spiral: The evidence that early intervention prevents reading failure. American Educator **28**(3), 6–19 (2004)
63. Vellutino, F.R., Scanlon, D.M., Zhang, H.: Identifying reading disability based on response to intervention: Evidence from early intervention research. In: Handbook of response to intervention, pp. 185–211. Springer (2007)
64. Wertheimer, R.F., Moore, K.A., Hair, E.C., Croan, T.: Attending kindergarten and already behind: A statistical portrait of vulnerable young children. Child Trends Washington, DC (2003)
65. Westlund, J.K., Breazeal, C.: The interplay of robot language level with children's language learning during storytelling. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts. pp. 65–66. ACM (2015)
66. Whittaker, S.: Theories and methods in mediated communication: Steve whittaker. In: Handbook of discourse processes, pp. 246–289. Routledge (2003)
67. Zill, N.: Promoting educational equity and excellence in kindergarten. The transition to kindergarten pp. 67–105 (1999)