Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots

Iolanda Leite[*], Marissa McCoy[†], Monika Lohani[†], Daniel Ullman[*], Nicole Salomons[*],
Charlene Stokes[†‡], Susan Rivers[†], Brian Scassellati[*]
[*]Department of Computer Science, [†]Department of Psychology
Yale University, New Haven, CT, USA
[‡]Air Force Research Laboratory
{iolanda.leite, marissa.mccoy, monika.lohani, daniel.ullman, nicole.salomons,
charlene.stokes, susan.rivers, brian.scassellati}@yale.edu

ABSTRACT

Robot assistive technology is becoming increasingly prevalent. Despite the growing body of research in this area, the role of type of interaction (i.e., small groups versus individual interactions) on effectiveness of interventions is still unclear. In this paper, we explore a new direction for socially assistive robotics, where multiple robotic characters interact with children in an interactive storytelling scenario. We conducted a between-subjects repeated interaction study where a single child or a group of three children interacted with the robots in an interactive narrative scenario. Results show that although the individual condition increased participant's story recall abilities compared to the group condition, the emotional interpretation of the story content seemed more dependent on the difficulty level rather than the study condition. Our findings suggest that, despite the type of interaction, interactive narratives with multiple robots are a promising approach to foster children's development of social-related skills.

Keywords

socially assistive robotics; child-robot interaction; emotional intelligence; interactive storytelling; type of interaction.

1. INTRODUCTION

Socially assistive robotics applications typically involve one robot and one user [35]. Several authors have also investigated settings with one socially assistive robot interacting with multiple users [37, 13]. However, as robot assistive technology becomes more sophisticated, and as robots are being used more broadly in interventions, there arises a need to explore other types of interactions.

In this paper, we investigate whether socially assistive robots are as effective in small groups of children as they are with an individual. Contrasting the typical "one robot to one user" and "one robot to many users" situations, there are cases where it is desirable to have multiple robots interacting with one user or multiple users. As an example, consider the case of role-playing activities in emotionally charged domains (e.g., bullying prevention, domestic violence or hostage scenarios). In these cases, taking an active role in the interaction may bring about undesirable consequences, while observing the interaction might serve as a learning experience. Here, robots offer an inexpensive alternative to human actors, displaying controlled behavior across interventions with different trainees.

Our goal is to use multiple socially assistive robots to help children build their emotional intelligence skills through interactive role-playing activities. As this is a novel research direction, several questions remain open. What is the effect of having multiple robots in the scene or, more importantly, what is the optimal type of interaction for these interventions? Should the type of interaction of the intervention focus on groups of children (as in traditional role-playing activities) or should we aim for individual interactions, following the current trend in socially assistive robotics?

Considering the nature of most assistive robotic interventions, one might expect individual interactions to be more effective. On the other hand, it has long been acknowledged that groups outperform individuals in a variety of activities, from quantitative judgments [32] to improved individual learning gains [11]. In educational research, for example, many authors highlight the benefits of learning in small groups rather than alone [10, 11, 27], including in learning activities supported by computers. A recent HRI study suggests that children behave differently when interacting alone or in dyads with a social robot [3]. However, it remains unknown whether type of interaction impacts the effectiveness of the robot intervention in terms of how much users can recall or learn from the interaction. To address this theme, we developed an interactive narrative scenario where a pair of robot characters played out stories centered around words that contribute to expanding children's emotional vocabulary [28]. To evaluate the effects of type of interaction, we conducted a three-week repeated interaction study where children interacted with robots either alone or in small groups, and then were individually asked questions on the interaction they had just witnessed. We analyzed interview responses in order to measure participants' story recall abilities, emotional understanding. Our results show that although children interacting alone with the robot were

HRI '15, March 02 - 05 2015, Portland, OR, USA
Copyright 2015 ACM 978-1-4503-2883-8/15/03 ...$15.00
http://dx.doi.org/10.1145/2696454.2696481

able to recall the narrative more accurately, no significant differences were found in the understanding of the emotional context of the stories. We discuss these implications for the future design of robot technology in learning environments.

2. RELATED WORK

A great deal of research has been conducted into the use of virtual agents in the context of interactive storytelling with children. Embodied conversational agents are structured using a foundation of human-human conversation, creating agents that appear on a screen and interact with a human user [7]. Interactive narratives, where users can influence the storyline through actions and interact with the characters, result in engaging experiences [31] and increase a user's desire to keep interacting with the system [12, 14]. *FearNot* is a virtual simulation with different bullying episodes where a child can take an active role in the story by advising the victim on possible coping strategies to handle the bullying situation. An extensive evaluation of this software in schools showed promising results on the use of such tools in bullying prevention [36]. Although some authors have explored the idea of robots as actors [6, 5, 12, 22], most of the interactive storytelling applications so far are designed for virtual environments.

Kim and Baylor [15] posit that the use of non-human pedagogical agents as learning companions creates the best possible environment for learning for a child. Virtual agents are designed to provide the user with the most interactive experience possible; however, research by Bainbridge et al. [2] indicates that physical presence matters in addition to embodiment, with participants in a task rating an overall more positive interaction when the robot was physically embodied rather than virtually embodied.

Furthermore, research by Leyzberg et al. [20] found that the students who showed the greatest measurable learning gains in a cognitive skill learning task were those who interacted with a physically embodied robot tutor, as compared to a video-represented robot and a disembodied voice. Research by Mercer [25] supports talk as a social mode of thinking, with talk in interaction between learners beneficial to educational activities. However, Mercer identifies the need for focused direction from a teaching figure for the interaction to be as effective as possible.

Shahid et al. [33] conducted a cross-cultural examination of variation between interactions in children who either played a game alone, with a robot, or with another child. They found that children both enjoyed playing more and were more expressive when they played with the robot, as compared to when they played alone. Still, not surprisingly, children who played with a friend showed the highest levels of enjoyment of all groups. With this previous research serving as the foundation, we posited that a combination of interactions with a robot and peers in a group setting could benefit information retainment and understanding of the interaction.

3. INTERACTIVE NARRATIVES WITH MULTIPLE ROBOTIC CHARACTERS

We developed an interactive narrative system such that any number of robotic characters can act out stories defined in a script. This system prompts children to control the actions of one of the robots at specific moments, allowing the child to see the impact of their decision on the course of the story. By exploring all the different options in these interactive scenarios, children have the opportunity to see how the effects of their decisions play out before them, without the cost of first having to make these decisions in the real world. This section describes the architecture of this system and introduces RULER, the validated framework for promoting emotional literacy that inspired the first interactive stories developed for this scenario.

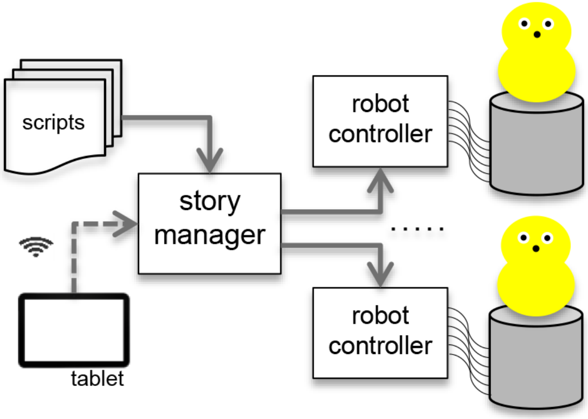

Figure 1: System architechture.

3.1 System Architecture

The central component of the narrative system is the story manager, which interprets the story scripts and communicates with the robot controller modules and the tablet (see diagram in Figure 1). The scripts are JSON files describing every possible scene episode. A scene contains the dialogue lines of each robot and a list of the next scene options that can be selected by the user. Each dialogue line contains an identifier of the robot playing that line (robotID), the path to a sound file[1] and a descriptor of a nonverbal behavior for the robot to display while "saying" that line (e.g., happy, bouncing). When the robots finish playing out a scene, the next story options are presented on the tablet as text with an accompanying illustration. When the user selects a new story option on the tablet, the story manager loads that scene and begins sending commands to the robots based on the scene dialogue lines.

The system was implemented on Robot Operating System (ROS). The robot platforms used in this implementation were two MyKeepon robots (see Figure 2) with programmable servos controlled by an Arduino board [1]. MyKeepon is a yellow, snowman like robot with three dots representing eyes and a nose. Despite their minimal appearance, research has shown that these robots can elicit social responses from children [16]. Each robot has four degrees of freedom: it can pan to the sides, roll to the sides, tilt forward and backward, and bop up and down. To complement the pre-recorded utterances, we developed several non-verbal behaviors such as idling, talking and bouncing. All the story authoring was done in the script files, except the robot animations and tablet artwork. In addition to

[1] Even if generated by TTS in real time, we consider prerecorded utterances.

Table 1: Summary of the story scenes in each session.

	Session 1	**Session 2**	**Session 3**
Feeling Word	Included	Frustration	Cooperation
Difficulty Level	Easy	Hard	Medium
Intro Scene	Leo is new at school and doesn't know anyone. Another student in class, Marlow, called Leo's hat stupid. What should Berry do to help Leo feel included?	Berry tells Leo that he just started a new book as part of an assignment, but some of the words are too hard for him to read. What should Berry do to get through his frustration?	Berry has just mastered a big, hard book on his own. Leo asks Berry to be his reading buddy. Leo wants to read an easier book that's on his reading level, while Berry wants to try reading the hardest books. What should Berry do to be cooperative?
Optional Scenes	- Talk bad about Marlow - Tell Leo how cool Marlow is - Ask Leo to play	- Ask Leo to read the book - Wait for the teacher - Try again	- Find another reading buddy - Choose a book both can read - Choose a hard book anyway

increased modularity, this design choice allows non-expert users (e.g., teachers) to develop new content for the system.

3.2 RULER

RULER is a validated framework rooted in emotional intelligence theory [30] and research on emotional development [9] that is designed to promote and teach emotional intelligence skills. Through a comprehensive approach that is integrated into existing academic curriculum, RULER focuses on skill-building lessons and activities around Recognizing, Understanding, Labeling, Expressing and Regulating emotions in socially appropriate ways [28]. Understanding the significance of emotional states guides attention, decision-making, and behavioral responses, and is necessary in order to navigate the social world [30, 21, 4].

This study employs components of RULER, including the Mood Meter, a tool that students and educators use as a way to identify and label their emotional state, and the Feeling Words Curriculum, a tool that centers on fostering an extensive feelings vocabulary that can be applied in students' everyday lives. The story scripts are grounded in the Feeling Words Curriculum and are intended to encourage participants to choose the most appropriate story choice after considering the impact of each option. Our target age group was 6 to 8 years old. Prior to beginning the study, we gathered feedback from elementary school teachers to ensure that the vocabulary and difficulty levels of story comprehension were age-appropriate. A summary of the scenes forming the scripts of each session are displayed in Table 1. All three stories followed the same structure: introduction scene, followed by three options. Each option impacted the story and the characters' emotional state in different ways.

4. EXPERIMENT

We conducted the user study described in this section to evaluate the impact of type of interaction (i.e., individual versus small groups) on children while interacting with multiple robots. Considering this is the first study in this domain, we did not formulate specific hypotheses, but rather outlined the following exploratory questions to investigate:

- How does type of interaction impact information recall?
- How does type of interaction impact children's emotional understanding and vocabulary?

As previously outlined, socially assistive robotic applications are typically one-to-one, but educational research suggests that children's learning gains may increase in a group [11, 27].

4.1 Study Design

We used a between subjects design with participants randomly sorted into one of two conditions: individual (one participant interacted alone with the robots) or group (three participants interacted with the robots at the same time). We studied groups of three children as three members is the smallest number of members considered to be a group [26]. Our main dependent metrics focused on participants' recall abilities and emotional interpretation of the narrative choices.

Each participant or group of participants interacted with the robots three times, approximately once per week. Participants in the group condition always interacted with the robots in the same groups. The design choice to use repeated interactions was not to measure learning gains over time, but to ensure that the results were not affected by a novelty effect that robots often evoke in children [19].

4.2 Participants

The participants in the study were first and second grade students from an elementary school where RULER, a social and emotional learning (SEL) program, had been implemented. A total of 46 participants were recruited in the school where the study was conducted, but six participants were excluded for various reasons (i.e., technical problems in collecting data or participants missing school). For this analysis, we considered a total of 40 children (22 females, 18 males) between the ages of 6 to 8 ($M = 7.53, SD = .51$). Ethnicity, as reported by guardians, was as follows: 17.5% African American, 17.5% Caucasian, 25% Hispanic, 27.5% reported more than one ethnicity, and 12.5% did not report. The annual income reported by guardians was as follows: 30% in \$0-\$20,000, 42.5% in \$20,000-\$50,000, 10% in the \$50,000-\$100,000 range, and 17.5% not reported.

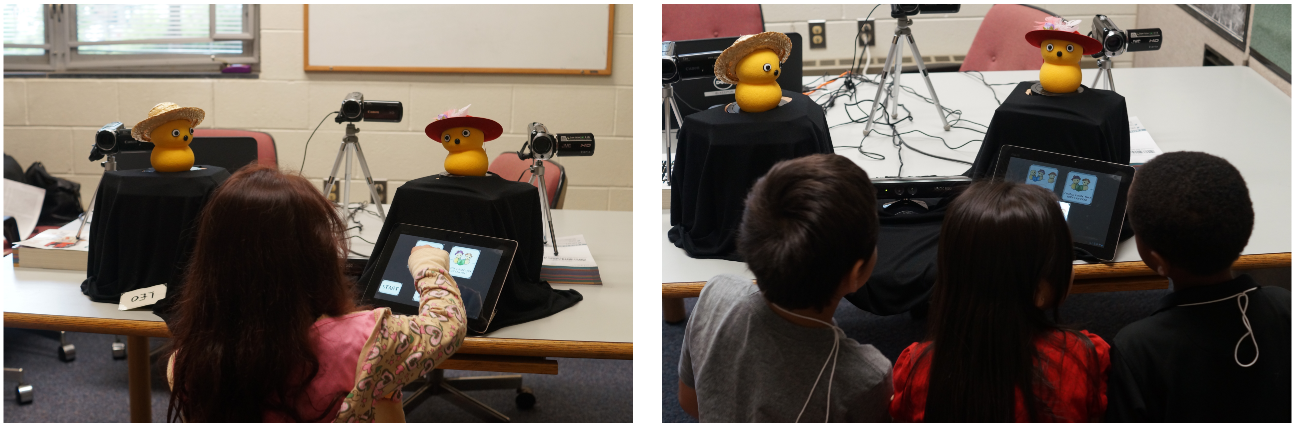

Figure 2: Children interacting with the robots in the individual (left) and group (right) conditions.

4.3 Procedure

Consent forms were distributed in classrooms that had agreed to participate in the study. Participants were randomly assigned to either the individual (19 participants) or group condition (21 participants). Each session lasted approximately 30 minutes with each participant. The participant first interacted with the robots either individually or in a small group (approximately 15 minutes), and then was interviewed individually by the experimenter (approximately 15 minutes).

Participants were escorted from class by a guide who explained that they were going to interact with robots and then would be asked questions about the interaction. The child was introduced to the experimenter and asked for verbal assent. The experimenter began by introducing the participants to Leo and Berry, the two main characters (MyKeepon robots) in the study. The first half of each session involved the participants interacting with the robots as the robots autonomously role-played a scenario centered around a RULER feeling word. After observing the scenario introduction, participants were presented with three different options from which to choose. Participants were instructed to first select the option they thought was the best choice, and were told they would then have the opportunity to choose the other two options. In the group condition, participants were asked to make a joint decision. The experimenter was present in the room at all times, but was outside participants' line of sight so as not to distract participants from the interaction.

After interacting with the robots, participants were interviewed by additional experimenters. The interviews had the same format for both conditions, which means that even participants in the group condition were interviewed individually. Interviews were conducted in nearby rooms. Experimenters followed a standardized protocol that asked the same series of questions (one open-ended question, followed by two direct questions) for each of the four scenes (i.e., Introduction, Option 1, Option 2, Option 3) that comprised one session. The same three repeated questions were asked in the following order:

1. What happened after you chose <option>?
2. After you chose <option>, what color of the Mood Meter do you think <character>
3. What word would you use to describe how <character> was feeling?

These questions were repeated for a total of 36 times (3 questions * 4 scenes per session * 3 sessions) over the course of the study. If a participant remained silent for more than 10 seconds after being asked a question, the experimenter asked, "Would you like me to repeat the question or would you like to move on". The interviewer used small cards with artwork representing the different scene choices similar to the ones that appeared on the tablet near the robots. All three sessions followed the same format (i.e., robot interaction followed by the series of interview questions). Interviews were audio-recorded and transcribed verbatim for coding.

5. DATA ANALYSIS

In this section, we describe how interview data was coded and how the main evaluation metrics were calculated.

5.1 Word Count

The number of words uttered by each participant during the interview were counted using an automated script. Placeholders such as "umm" or "uhh" did not contribute toward word count. This metric was mainly used as a manipulation check for the other measures.

5.2 Story Recall

Responses to the open-ended question "What happened after you chose <option>?" were coded as the variable Story Recall. Story Recall was further broken down into Narrative Structure Score (NSS), Narrative Accuracy (NAC) and Narrative Inaccuracy (NIN). Similar recall metrics have been previously used in HRI studies with adults [34].

For Narrative Structure Score (NSS), we followed the coding scheme used in previous research by McGuigan and Salmon [24] and McCartney and Nelson [23], in which participants' verbal responses to open-ended questions were coded for the presence or absence of core characters (e.g., Leo, Berry) and core ideas (e.g., Leo doesn't know anyone, everyone is staring at Leo's clothes). This score provides a snapshot of the participants' "ability to logically recount the fundamental plot elements of the story" [24, 23]. For session S and participant i, NSS was computed using the following formula:

$$NSS_{S,I} = \frac{Mentioned(CoreCharacters + CoreIdeas)}{All(CoreCharacters + CoreIdeas)}$$

A perfect NSS of 1.0 would indicate that the participant mentioned all the core characters and main ideas in all four open-ended questions of that interview. The first mention of core characters and core ideas were given a point each, with additional mentions not counted. The sum of core characters and core ideas for each interview session were combined to generate the Narrative Structure Score. The average number of characters in each story was three (Leo, Berry, and Marlow or the teacher), while the number of core ideas varied depending on the difficulty of the story, ranging from an average of four in the easier story to six in the hardest.

Previous coding schemes were followed for Narrative Accuracy and Narrative Inaccuracy [18, 24, 23]. These metrics capture students' ability to move beyond simply recounting overarching story themes, instead describing a more granular or nuanced account of the story. The same responses to the open-ended question were also coded for *correct event details*, *extra-event details*, *intrusions* and *distortions*. Event details included actions, objects or descriptors that were part of a story event but not considered core ideas, and extra-event details were references to the participant's opinions, feelings or thoughts (e.g., "Jake is new to my class and I asked him to play"). Intrusions were mentioned actions, objects or descriptors that were not part of the event, while distortions were considered any actions, objects, or descriptors that were part of the event but inaccurately described (e.g., "Marlow said Leo's shoes are stupid"). Narrative Accuracy (NAC) and Narrative Inaccuracy (NIN) were calculated using the following formulas:

$$NAC_{S,I} = EventDetails + 0.5 * ExtraEventDetails$$

$$NIN_{S,I} = Intrusions + 0.5 * Distortions$$

Higher NAC and NIN scores denote a greater number of correct story descriptors or story errors during story recall, respectively. NAC scores for each scene were summed to create an aggregate NAC score for each interview session, as were NIN aggregate scores. The number of correct and extra-event details, intrusions and distorsions normally ranged from 0 to 4.

5.3 Emotional Understanding

The Emotional Understanding Score (EUS) represents participants' ability to correctly recognize and label character's emotional states, a fundamental skill of RULER [8, 4]. Responses to the two direct questions "After you chose <option>, what color of the Mood Meter do you think <character> was in?" and "What word would you use to describe how <character> was feeling?" were coded based on RULER concepts and combined to comprise EUS.

Appropriate responses for the first question were based on the Mood Meter colors and included Yellow (pleasant, high energy), Green (pleasant, low energy), Blue (unpleasant, low energy), or Red (unpleasant, high energy), depending on the emotional state of the robots at specific points in the roleplay. Responses to the second direct question were based on the RULER Feeling Words Curriculum with potential appropriate responses being words such as excited (pleasant, high energy), calm (pleasant, low energy), upset (unpleasant, low energy), or angry (unpleasant, high energy), depending on which color quadrant the participant "plotted" the character. Since participants were recruited from schools implementing RULER, they use the Mood Meter daily and are used to these type of questions. Most participants answered with one or two words when asked to describe the character's feelings.

For the ColorScore, participants received +1 if the correct Mood Meter color was provided, and -1 if an incorrect color was given. In the FeelingWordScore, participants received +1 or -1 depending on whether the feeling word provided was appropriate or not. If participants provided additional appropriate or inappropriate feeling words, they were given +0.5 or -0.5 points for each, respectively. The total EUS was calculated using the following formula:

$$EUS_{S,I} = ColorScore + FeelingWordScore$$

Higher EUS means that participants were able to more accurately identify the Mood Meter color and corresponding feeling word associated with the character's emotional state. For each interview session, EUS scores for each scene were summed to calculate an aggregate EUS score.

5.4 Coding and Reliability

Two researchers independently coded the interview transcriptions from the three sessions. Both coders first coded the interviews from the excluded participants to become familiar with the coding scheme. Once agreement between coders was reached, coding began on the remaining data. Coding was completed for the 120 collected interviews (40 participants * 3 sessions), overlapping 25% (30 interviews) as a reliability check.

Reliability analysis between the two coders was performed using the Intra-class Correlation Coefficient test for absolute agreement using a two-way random model. All the coded variables for each interview session had high reliabilities. The lowest agreement was found in the number of correct feeling words ($ICC(2,1) = 0.85, p < .001$), while the highest agreement was related to the total number of core characters mentioned by each child during one interview session ($ICC(2,1) = 0.94, p < .001$). Given the high agreement between the two coders in the overlapping 30 interviews, data from one coder were randomly selected to be used for analyses.

6. RESULTS

Mixed model Analyses of Variance (ANOVA) models were conducted with type of interaction (individual versus group) as the between-subjects factor and session (1, 2, and 3) as the within-subjects factor. For all the dependent measures, we planned to test the individual versus group differences in each session.

6.1 Word Count

We examined whether there were any differences between individual versus group level in the number of words spoken by the participants during the interview sessions. An ANOVA model was run with the number of words spoken as the dependent measure. Neither a main effect nor interaction effect was found to be significant. Thus, overall, there

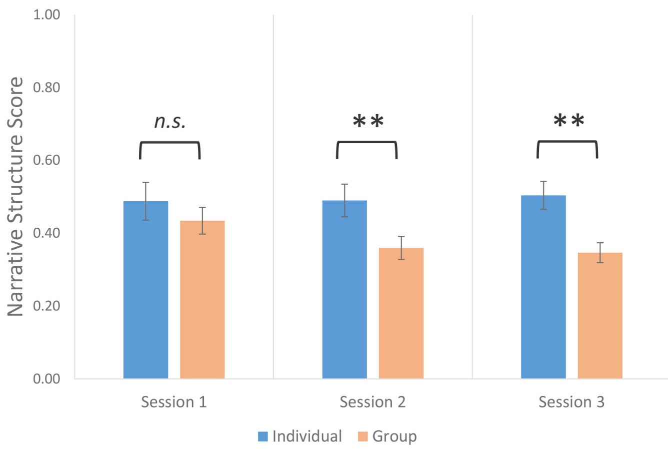

Figure 3: Average Narrative Structure Scores (NSS) for participants in each condition on every interaction session. (**) denotes $p < .01$.

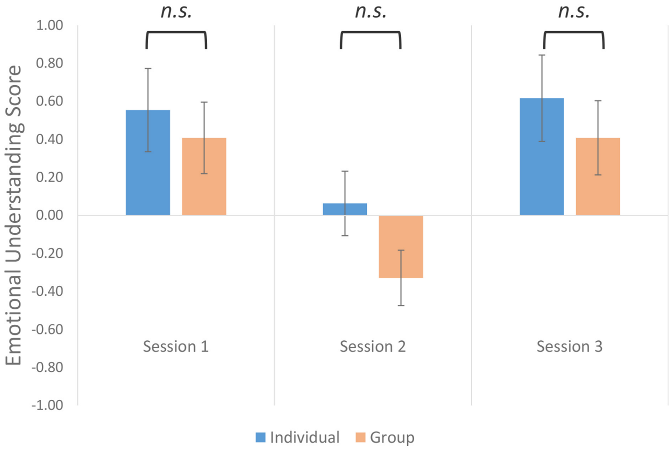

Figure 4: Average Emotional Understanding Scores (EUS) for participants in each condition for sessions 1 (easy), 2 (advanced) and 3 (medium). No significant differences were found between conditions.

was no significant difference in word count between the two groups. The average number of words per interview was 124.82 ($SE = 16.01$). This variability seems to stem from each participant's individual differences and was not related to the participant's condition. This result is important because it serves as a manipulation check for other reported findings.

6.2 Story Recall

We investigated the impact of type of interaction on participants' story recall abilities, measured by the Narrative Structure Score (NSS) and Narrative Accuracy/Inaccuracy (NAC and NIN, respectively). An ANOVA model was run with NSS as the dependent measure. We found a significant main effect of type of interaction (collapsed across sessions), with students in the individual condition achieving higher scores on narrative structure ($M = .49, SE = .03$) than the group condition ($M = .38, SE = .02$), $F(1, 28) = 7.71, p = .01, \eta2 = .22$. Between type of interaction and session, neither a main effect nor an interaction effect was significant (see Figure 3).

Planned comparisons were conducted to test the role of type of interaction in each session. No significant differences were found for session 1. For session 2, students in the individual condition ($M = .49, SE = .05$) had a higher score than the students in the group condition ($M = .36, SE = .03$), $F(1, 36) = 7.35, p = .01, \eta2 = .17$. Similarly, for session 3, students in the individual condition ($M = .50, SE = .04$) had a higher score than in the group condition ($M = .35, SE = .03$), $F(1, 38) = 6.59, p = .01, \eta2 = .15$.

The other measures we used to study participants' story recall abilities were Narrative Accuracy (NAC) and Narrative Inaccuracy (NIN). An ANOVA model with NAC as the dependent measure did not find significant differences between the individual versus the group condition. A significant main effect of session was found, $F(2, 74) = 4.98, p = .01, \eta2 = .12$, but the interaction between type of interaction and session was nonsignificant. None of the type of interaction related (individual versus group) planned contrasts were significant, despite the slightly higher average scores of NAC in the individual condition ($M = .38, SE = .09$), compared to the values on the group condition ($M = .19, SE = .09$). An ANOVA model with NIN as the dependent measure also suggested that there were no significant effects in each session due to type of interaction or session. Yet again, on average, participants in the individual condition performed marginally better (i.e., lower NIN) than participants from the group condition, ($M = .26, SE = .09$) and ($M = .36, SE = .08$), respectively.

These findings suggest that overall the narrative story related recall rate was found to be higher in the individual versus the group level interaction with the robots. In the easier session (session 1), there was no effect on type of interaction, but during the harder sessions (sessions 2 and 3), students were found to perform better in individual than group level interactions. In addition, no type of interaction-related differences were found in NAC nor NIN.

6.3 Emotional Understanding

To investigate our second research question, we tested whether students' Emotional Understanding Score differed in the individual versus group condition. An ANOVA model with EUS as the dependent measure suggested that there was no main effect of type of interaction. The main effect of session was significant $F(2, 62) = 7.39, p = .001, \eta2 = .19$, which aligns with our expectation given that the three sessions had different levels of difficulty. Type of interaction versus session interaction effect was not significant (see Figure 4). Planned comparisons also yielded no significant differences between the individual versus groups in any of three sessions. In sum, the degree of emotional understanding did not seem to be affected by type of interaction in this setting, but varied across sessions with different levels of difficulty.

7. DISCUSSION

Our results yielded interesting findings about the effects of type of interaction on children's interactions with multiple socially assistive robots. Participants interacting with the robots alone were able to recall the Narrative Structure (i.e., core ideas and characters) significantly better than participants in the group condition. On average, participants in the individual condition also enumerated more correct story details in every session, and less inconsistencies in 2 out of the 3 sessions, but these results were not statistically significant across conditions.

Three main interpretations can be taken from these results. First, while the child was solely responsible for all choices when interacting alone, decisions were shared when in the group, thereby affecting how the interaction was experienced. A second interpretation is that in individual interactions, children may be more attentive since social standing in relation to their peers is not a factor. Thirdly, the peers might be simply more distracting.

At first glance, our results may seem to contradict previous findings highlighting the benefits of learning in small groups [11, 27]. However, recalling story details is different than increasing learning gains. In fact, no significant differences were found between groups in our main learning metric, Emotional Understanding Score (participants' ability to interpret the stories using the concepts of the RULER framework), despite average individual condition scores being slightly higher for every session. Other than session 2, which had the most difficult story content, all participants performed quite well despite the type of interaction in which they interacted. One possible explanation, in line with the findings from Shahid et al. [33], is that participants in the individual condition might have benefited from some of the effects of a group setting since they were interacting with multiple autonomous agents (the two robots). Moreover, several authors argue that group interaction and subsequent learning gains do not necessarily occur just because learners are in a group [17]. An analysis of the participants' behavior while in the group during the interaction could clarify these hypotheses.

8. DESIGN IMPLICATIONS

There are obvious reasons why having multiple children instead of one child interact with a robot at a time is favorable, including limitations of cost, time and space. Our work focuses on whether the advantages of one-on-one tutoring, which have been well established in the HRI domain, can also apply to one-to-many instruction and what costs might be incurred when this shift happens.

While individual interactions seem to be more effective in the short-term, group interventions might be more suitable in the long-term. Previous research has shown that children have more fun interacting with robots in groups rather than alone [33]. Since levels of engagement are positively correlated with students' motivation for pursuing learning goals [29], influence concentration, and foster group discussions [38], future research in this area should study the effects of type of interaction in long-term interaction with robots.

Another implication of our findings is that instruction primarily concerned with factual recall (such as basic arithmetic facts) might be best served by one-on-one interactions, but that other skill-based and outcome-based instruction (such as interpersonal skill training, as in our study) might be amenable to one-to-many instruction. These results align with previous research comparing individual and group learning gains with computer-based technology. The meta-analysis performed by Pai and colleagues [27], for example, showed that effect sizes of individual versus group learning were smaller for more exploratory tasks.

To keep the gains of individual interactions in group settings, it might be necessary to implement more sophisticated perception mechanisms in the robots. For example, the robots could detect disengagement and employ recovery mechanisms to keep children focused in the interaction. Similarly, robots that capture the complex dynamics of group interactions by perceiving and intervening when a child is dominating an interaction would be useful in group settings.

9. CONCLUSION

The effective acquisition of social and emotional skills requires constant practice in diverse hypothetical situations. In this paper, we proposed a novel approach where multiple socially assistive robots are used in interactive role-playing activities with children. The robots acted as interactive puppets; children could control the actions of one of the robots and see the impact of the selected actions on the course of the story. Using this scenario, we investigated the effects of type of interaction (individual versus small group interactions) on children's story recall and emotional interpretation of interactive stories.

Results from this repeated interaction study showed that although participants who interacted alone with the robot remembered the story better than participants in the group condition, no significant differences were found in children's emotional interpretation of the stories. This latter metric was fairly high for all participants, except in the session with the hardest story content. Despite the promising results of this study, further research is needed to understand how type of interaction affects children's learning gains in longer-term interactions with socially assistive robotics.

10. ACKNOWLEDGMENTS

This work was supported by the NSF Expedition in Computing Grant #1139078 and SRA International (US Air Force) Grant #13-004807. We thank Bethel Assefa for assistance in data collection and coding, Rachel Protacio and Jennifer Allen for recording the voices for the robots, Emily Lennon for artwork creation, and the students and staff from the school where the study was conducted.

11. REFERENCES

[1] H. Admoni, A. Nawroj, I. Leite, Z. Ye, B. Hayes, A. Rozga, J. Rehg, and B. Scassellati. Mykeepon project. http://hennyadmoni.com/keepon. Accessed January 2, 2015.

[2] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati. The benefits of interactions with physically present robots over Video-Displayed agents. *Adv. Robot.*, 3(1):41–52, 1 Jan. 2011.

[3] P. Baxter, J. de Greeff, and T. Belpaeme. Do children behave differently with a social robot if with peers? In *International Conf. on Social Robotics (ICSR 2013)*. Springer, 2013.

[4] M. A. Brackett, S. E. Rivers, and P. Salovey. Emotional intelligence: Implications for personal, social, academic, and workplace success. *Soc. Personal. Psychol. Compass*, 5(1):88–103, 2011.

[5] C. Breazeal, A. Brooks, J. Gray, M. Hancher, J. McBean, D. Stiehl, and J. Strickon. Interactive robot theatre. *Commun. ACM*, 46(7):76–85, July 2003.

[6] A. Bruce, J. Knight, S. Listopad, B. Magerko, and I. Nourbakhsh. Robot improv: Using drama to create believable agents. In *Proc. of the Int. Conf. on*

Robotics and Automation, ICRA'00, pages 4002–4008. IEEE, 2000.

[7] J. Cassell. *Embodied Conversational Agents*. MIT Press, 2000.

[8] R. Castillo, P. Fernández-Berrocal, and M. A. Brackett. Enhancing teacher effectiveness in Spain: A pilot study of the RULER approach to social and emotional learning. *Journal of Education and Training Studies*, 1(2):263–272, 16 Aug. 2013.

[9] S. A. Denham. *Emotional development in young children*. Guilford Press, 1998.

[10] P. Dillenbourg. What do you mean by collaborative learning? *Collaborative-learning: Cognitive and Computational Approaches.*, pages 1–19, 1999.

[11] G. W. Hill. Group versus individual performance: Are n+1 heads better than one? *Psychological Bulletin*, 91(3):517, 1982.

[12] G. Hoffman, R. Kubat, and C. Breazeal. A hybrid control system for puppeteering a live robotic stage actor. In *Proc. of RO-MAN 2008*, pages 354–359. IEEE, 2008.

[13] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro. A two-month field trial in an elementary school for long-term human-robot interaction. *IEEE Transactions on Robotics*, 23(5):962–971, Oct 2007.

[14] C. Kelleher, R. Pausch, and S. Kiesler. Storytelling alice motivates middle school girls to learn computer programming. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '07, pages 1455–1464, New York, NY, USA, 2007. ACM.

[15] Y. Kim and A. L. Baylor. A Social-Cognitive framework for pedagogical agents as learning companions. *Educ. Technol. Res. Dev.*, 54(6):569–596, 2006.

[16] H. Kozima, M. Michalowski, and C. Nakagawa. Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics*, 1(1):3–18, 2009.

[17] K. Kreijns, P. A. Kirschner, and W. Jochems. Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Comput. Human Behav.*, 19(3):335–353, May 2003.

[18] S. Kulkofsky, Q. Wang, and S. J. Ceci. Do better stories make better memories? narrative quality and memory accuracy in preschool children. *Appl. Cogn. Psychol.*, 2008.

[19] I. Leite, C. Martinho, and A. Paiva. Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, 5(2):291–308, 2013.

[20] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *Proc. of the 34th Annual Conf. of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, 2012.

[21] P. N. Lopes, P. Salovey, S. Coté, and M. Beers. Emotion regulation abilities and the quality of social interaction. *Emotion*, 5(1):113–118, Mar. 2005.

[22] D. Lu and W. Smart. Human-robot interactions as theatre. In *RO-MAN, 2011 IEEE*, pages 473–478, July 2011.

[23] K. A. McCartney and K. Nelson. Children's use of scripts in story recall. *Discourse Process.*, 1981.

[24] F. McGuigan and K. Salmon. The inflience of talking on showing and telling: adult-child talk and children's verbal and nonverbal event recall. *Appllied Cognitive Psychology*, 2006.

[25] N. Mercer. The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction*, 6(4):359–377, Dec. 1996.

[26] R. L. Moreland. Are dyads really groups? *Small Group Research*, 17 Feb. 2010.

[27] H.-H. Pai, D. A. Sears, and Y. Maeda. Effects of small-group learning on transfer: A meta-analysis. *Educational Psychology Review*, pages 1–24, 2013.

[28] S. E. Rivers, M. A. Brackett, M. R. Reyes, N. A. Elbertson, and P. Salovey. Improving the social and emotional climate of classrooms: a clustered randomized controlled trial testing the RULER approach. *Prev. Sci.*, 14(1):77–87, Feb. 2013.

[29] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

[30] P. Salovey and J. D. Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1989.

[31] H. Schoenau-Fog. Hooked!–evaluating engagement as continuation desire in interactive narratives. In *Interactive storytelling*, pages 219–230. Springer, 2011.

[32] T. Schultze, A. Mojzisch, and S. Schulz-Hardt. Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, 118(1):24 – 36, 2012.

[33] S. Shahid, E. Krahmer, and M. Swerts. Child-robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend? *Comput. Human Behav.*, 40(0):86–100, Nov. 2014.

[34] D. Szafir and B. Mutlu. Pay attention!: Designing adaptive agents that monitor and improve user engagement. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '12, pages 11–20, New York, NY, USA, 2012. ACM.

[35] A. Tapus, M. Matarić, and B. Scassellatti. The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine*, 14(1), 2007.

[36] N. Vannini, S. Enz, M. Sapouna, D. Wolke, S. Watson, S. Woods, K. Dautenhahn, L. Hall, A. Paiva, and E. André. Fearnot!: computer-based anti-bullying programme designed to foster peer intervention. *European journal of psychology of education*, 26(1):21–44, 2011.

[37] K. Wada, T. Shibata, and Y. Kawaguchi. Long-term robot therapy in a health service facility for the aged-a case study for 5 years. In *International Conf. on Rehabilitation Robotics*, pages 930–933. IEEE, 2009.

[38] H. J. Walberg. Productive teaching and instruction: Assessing the knowledge base. *Phi Delta Kappan*, pages 470–478, 1990.