International Journal of Child-Computer Interaction 34 (2022) 100540

Contents lists available at ScienceDirect

International Journal of Child-Computer Interaction

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

Research paper

Examining voice assistants in the context of children's speech

Min Kyong Kim [a], Stefania Druga [b], Shaghayegh Esmaeili [c], Julia Woodward [c], Alex Shaw [c], Ayushi Jain [c], Jaida Langham [d], Kristy Hollingshead [e], Silvia B Lovato [f], Erin Beneteau [b], Jaime Ruiz [c], Lisa Anthony [c], Alexis Hiniker [b,*]

[a] Human Centered Design and Engineering, University of Washington, Seattle, WA, USA
[b] Information School, University of Washington, Seattle, WA, USA
[c] Computer & Information Science & Engineering, University of Florida, Gainesville, FL, USA
[d] Spelman College, Atlanta, GA, USA
[e] Institute of Human and Machine Cognition, Ocala, FL, USA
[f] PBS Kids Learning Technologies, Public Broadcasting Service (PBS), Arlington, VA, USA

ARTICLE INFO

Article history:
Received 2 May 2022
Received in revised form 30 August 2022
Accepted 13 September 2022
Available online 1 October 2022

MSC:
0000
1111

Keywords:
Child–computer interaction
Voice assistants
Smart devices

ABSTRACT

An estimated 3.25 billion voice assistants (VAs) are in homes around the world, but these devices are not always able to recognize and respond to children's speech. To inform the design of VAs that support kids, we report on a lab study where 28 5- to 10-year-old participants interacted with a commercial VA to: (1) attempt to execute common VA-supported requests (such as setting an alarm), (2) recite a set of such scripts verbatim, and (3) engage in unstructured conversation. We find that devices only respond appropriately to the content of children's speech half of the time. Frequency of appropriate responses increased with children's age and as their discourse became more standardized. Based on themes in participants' speech, we identify design opportunities in child-VA interaction, such as exploring a topic or responding to a conversational bid. In addition to our empirical findings, we contribute a structured corpus of children's speech.

1. Introduction

Voice assistants (VAs) have become a pervasive presence in children's lives in the developed world (Statista, 2019), with total number of VAs expected to be roughly equal to the global population of eight billion by 2023 (Statista, 2022). These devices have expanded the information-seeking capabilities of young children (Lovato & Piper, 2015; Lovato, Piper, & Wartella, 2019), who often struggle to use traditional point-and-click interfaces and keyboards (Gossen, Kotzyba, Stober, & Nürnberger, 2013). Since VAs have been introduced into children's lives, researchers have begun to examine children's experiences with these interfaces. Prior work has investigated, for example, the types of questions children pose to VAs (Lovato & Piper, 2015; Lovato et al., 2019), the repair strategies children use when these interfaces fail to understand them (Beneteau et al., 2019; Cheng, Yen, Chen, Chen, & Hiniker, 2018; Yarosh et al., 2018), children's perceptions of VA intelligence and personality (Druga, Williams, Breazeal, & Resnick, 2017; Druga, Williams, Park, & Breazeal, 2018; Lovato et al., 2019), the influence of socioeconomic and cultural factors on child-VA interactions (Druga, Vu, Likhith, & Qiu, 2019), and the role parents play in shaping children's interactions with VAs Beneteau et al. (2020) and Porcheron, Fischer, Reeves, and Sharples (2018).

These prior studies have shown that children have varying degrees of success when interacting with VAs. The devices routinely provide incorrect answers to children's questions (Lovato et al., 2019) and rarely provide support that can guide child users to a more productive exchange (Beneteau et al., 2019). Thus, we sought to examine how these technologies might better support and respond to children's VA-directed speech. We build on the growing body of literature in this space by investigating, first, how children of different ages formulate their questions and commands (which we collectively refer to as "utterances" throughout the paper) when talking with VAs and, second, the appropriateness of a current commercial VA's responses. Specifically, we pose the following research questions:

- RQ1: How do children formulate questions and commands when attempting common VA-supported tasks?
- RQ2: When given the freedom to engage with a VA in any way they choose, what do children choose to say and how do they formulate this speech?
- RQ3: How effective are VAs in responding appropriately to children's utterances?

* Correspondence to: University of Washington Information School, Campus Box 352840, Seattle, WA, 98195, USA
E-mail address: alexisr@uw.edu (A. Hiniker).

https://doi.org/10.1016/j.ijcci.2022.100540
2212-8689/

- RQ4: How do these patterns change with age?

To answer these questions, we recruited 28 children between the ages of five and ten years old to participate in a three-part observational lab study. We asked children to: (1) formulate questions and commands to achieve specific tasks, such as setting an alarm, spelling a word, and asking a question about their favorite cartoon character, (2) repeat or read a scripted set of questions and commands to test the recognition accuracy of children's speech with a commercially available VA (i.e., an Amazon Alexa dot), and (3) engage in live, unstructured conversation with a commercial VA (the Amazon Echo). Consistent with prior work, we found that average transcription accuracy (i.e., recognition accuracy, or the percentage of utterances correctly transcribed) for unstructured, child-led conversation was 84%, but children's utterances elicited meaningful, on-topic responses from devices only half of the time. When children attempted tasks pre-specified by the research team, such as setting an alarm, they were more likely to receive a correct response if they formulated their utterance using a common structure, and children's likelihood of producing these common structures increased with age. When children had the freedom to engage in any interaction they chose, they were most likely to engage in relationship-building speech, which was poorly or not at all supported by the VA. Based on these findings, we identify opportunities for designers to improve child-VA interactions, such as adding support for exploring a topic together or responding to a child's conversational bid (e.g., "*I like to slide*"). To support the implementation of future conversational interfaces for young children, we contribute a corpus of over 10 h of raw audio recordings and 1338 utterances characterizing children's speech and formulation patterns when interacting with VAs.

2. Related work

2.1. Child speech and language development

As children grow, they typically follow well-understood patterns of speech and language development. Specifically, by age five, they produce most sounds clearly, although they may continue to have occasional difficulty with particularly challenging phonemes (Dodd, Holm, Hua, & Crosbie, 2003; Stoel-Gammon & Cooper, 1984). Thus, there is reason to expect speech recognition systems to interpret children's speech correctly by this age. The rate at which this development occurs, however, varies between individuals and is influenced by the child's production practice and the amount of exposure a child has to language (Clark, 2009; Stoel-Gammon & Cooper, 1984; Vihman, DePaolis, & Keren-Portnoy, 2009). In addition to learning to produce sounds, children also spend their first years developing two key aspects of linguistic knowledge that are critical to conversational interaction: by age five, typically developing children have both, (1) a robust understanding of syntactic structure (i.e., how words and phrases are combined to form meaning), and (2) an expansive vocabulary. As a result, by this age, children have learned to assign abstract syntactic roles to words and phrases, and they understand the rules that govern combining these speech components (Brooks & Tomasello, 1999; Tomasello, 2000). This allows children to form simple sentences to describe events and actions.

Despite this emerging proficiency, young children's speech continues to lack much of the sophistication of adult speech. Children's lexical knowledge, for example, only expands gradually (Clark, 2009; Dąbrowska & Lieven, 2005; Rowland, 2007; Tomasello, 2000), and they use syntactically simpler and shorter constructions than do adults (Berman, 2007; Nippold, 1998, 2004). As children grow into adolescence, they have increased opportunities to socialize with peers and are exposed to a wider variety of communication contexts, allowing them to further refine their conversational skills (Nippold, 1998). Thus, children acquire a wide variety of linguistic abilities in their preschool years but continue to expand this base into adulthood. Given these emergent abilities, we examine the extent to which children's utterances at this stage of development allow for effective interactions with current VAs.

2.2. Speech recognition and children

Over the last several decades, a number of studies (e.g., Gerosa, Giuliani, and Brugnara (2007), Kennedy et al. (2017), Li and Russell (2002), Liao et al. (2015) and Potamianos, Narayanan, and Lee (1997)) have explored the technical feasibility of automatic speech recognition (ASR) and advanced the state-of-the-art in this space. A small subset of work on automatic speech recognition (ASR) has examined recognition accuracy with children, focusing on the speech of children age five and above (Gerosa et al., 2007; Li & Russell, 2002; Potamianos et al., 1997). The accuracy of ASR for children has lagged behind that of adults, mainly due to the acoustic variability in children's speech and a lack of training data from children. Historically, recognition accuracy rates have ranged from 60%–65% for younger children (ages six to eight) (Potamianos et al., 1997) and for children observed to have "poor pronunciation" as judged by adults (Li & Russell, 2002). A more recent study of off-the-shelf ASR engines such as Google's Speech API (Kennedy et al., 2017) suggests that these technologies successfully recognize children's spontaneous speech only 18% of the time. By age 13, however, ASR accuracy is found to be comparable to that of adults (Gerosa, Giuliani, Narayanan, & Potamianos, 2009; Potamianos et al., 1997). Applying speech normalization techniques has also been shown to improve ASR accuracy for children to above 70% (Gerosa et al., 2007; Li & Russell, 2002).

Most of these performance tests have been conducted in controlled settings where children are asked to repeat a fixed set of words (which enables systematic recognition testing) rather than in a natural setting. In addition, very little work has examined the recognition accuracy of children's speech by commercial VAs. One study found that adults and children exhibit different somewhat usage patterns when engaging with commercially available systems (with, for example, adults being more likely to use VAs for assistance and information-seeking, and children more likely to use them for music, jokes, and other entertainment Beneteau et al., 2020), and this may affect the types of statements each of these user groups make. A 2019 in-home study examined five-to six-year-old children's naturalistic interactions with one commercial VA (a Google Home mini) (Lovato et al., 2019). The device correctly matched the researchers' transcription for 89% of utterances. This result suggests the potential for these devices to perform well for children in ecologically valid contexts. We contribute to this body of work by examining transcription accuracy across a range of controlled and free-form tasks and by analyzing the extent to which errors in VA responses reflect inaccuracies in transcription versus failures to correctly interpret the conceptual and contextual dimensions of children's speech.

2.3. Children's interactions with VAs

Most relevant to our current study, a growing body of work in human–computer interaction has examined children's communication practices with VAs in both lab and home settings (Beirl, Yuill, & Rogers, 2019; Geeng & Roesner, 2019; Lovato & Piper, 2015; Lovato et al., 2019; Porcheron et al., 2018). This line of work has found that children frequently use VAs collaboratively with their parents and other family members for information retrieval.

In doing so, both children and adults use classical conversation techniques when addressing the VA, such as prosody changes and strategic silences (Porcheron et al., 2018). Children also engage the device in a wide range of topics, from science and technology to informal jokes (Lovato et al., 2019), and they ask questions mostly to understand the agent embodied by the VA (Lovato & Piper, 2015). Consistent with this finding, others have found that children prefer personified versions of these assistants (Yarosh et al., 2018) and often anthropomorphize them (Druga et al., 2017, 2018; Lovato et al., 2019). Although children attempt to interact with VAs with the expectations and desires they bring to human-to-human conversation, today's VAs are ineffective in participating in conversational exchange that supports these goals (Lovato et al., 2019; Porcheron et al., 2018).

Many studies thus focus on describing the frequent communication breakdowns that occur between children and these interfaces (Beirl et al., 2019; Beneteau et al., 2019; Cheng et al., 2018; Yarosh et al., 2018). When communication breakdowns happen, children struggle to repair the conversation successfully (Beneteau et al., 2019; Cheng et al., 2018; Yarosh et al., 2018) and use repetition as their primary recovery strategy (Yarosh et al., 2018). Prior work recommends that VAs shoulder more of this repair burden; through clarifying questions and other forms of discourse scaffolding, VAs could work collaboratively with children to correct the VA's misunderstandings (Beirl et al., 2019; Beneteau et al., 2019; Cheng et al., 2018).

We build on and extend prior work in several ways. In the present study, we place a greater focus on analyzing children's utterances to VAs: we describe the form of children's utterances, identify the common goals behind children's statements and questions, and characterize VAs' ability to respond to these formulations.

2.4. Child speech and interaction datasets

In the field of HCI, a well-structured dataset can be a valuable contribution to the research community (Wobbrock & Kientz, 2016). However, datasets collected from child users are relatively rare. This is perhaps unsurprising as attentional demands, motivational differences, and challenges in understanding and following instructions can make participating in data collection burdensome for children (Morrow & Richards, 1996; Punch, 2002). As a result, there are very few publicly available datasets of children's speech that can be used to evaluate VAs and other voice interfaces. The **CSLU Kids' Speech** corpus (Shobaki, Hosom, & Cole, 2007) is composed of utterances from 1100 children ages five through 15 and includes both prompted and spontaneous speech. However, these utterances are not specifically tailored to interaction with voice assistants and are now more than ten years old. The most recent and VA-relevant dataset is the **My Science Tutor (MyST) Children's Speech Corpus**. It consists of 393 h of children's speech collected from 1371 third-, fourth-, and fifth-grade students. The participants engaged in spoken dialogue with a virtual science tutor in eight areas of science. A total of 10,496 student sessions of 15 to 20 min produced a total of 228,874 utterances (Ward et al., 2011). Although this valuable dataset provides a rich set of utterances for research in this context, the content is narrowly scoped and is not reflective of the more open-ended child-VA interactions prevalent in families' homes.

In dataset collection studies designed to enable recognition and classification experiments, unlike other types of usability, participatory design, or even elicitation studies conducted with children (Woodward et al., 2018), a key principle is the need to collect consistent and balanced examples of interaction behaviors from multiple children. To this end, Anthony and colleagues have developed a method for collecting such systematic datasets from children (Anthony, Brown, Nias, Tate, & Mohan, 2012). Some of the key components of this method include using prizes at discrete intervals to gamify tasks and balancing the number of elicited examples per child (Anthony et al., 2012; Brewer et al., 2013; Woodward et al., 2016). In this study, we used a modified version of these procedures to collect our data, enabling us to publicly release a **Kids' Voice Assistant Corpus**[1] for future researchers, in addition to the empirical findings presented in this paper.

3. Method

We conducted a three-part lab study to investigate, first, the structure and content of five-to ten-year-old children's VA-directed speech, and second, the appropriateness of VAs' responses to this speech. We recorded children attempting common VA tasks, reciting scripted speech verbatim (i.e., to systematically test VA recognition accuracy), and engaging in free-form conversation with a specific VA (the Amazon Echo).

3.1. Participants

Twenty-eight children (15 girls) participated in our study. All children were recruited from a local elementary school. Children's ages ranged from five to ten years old (M: 7.43 yrs, SD: 1.37 yrs). At the beginning of the study session, we asked participants to tell us their favorite voice input device if they had one. Of the 28 participants, 13 chose Siri (46.4%), five chose Alexa (17.9%), four chose Google (14.3%), one chose Cortana (3.6%), and five others did not specify anything or did not know what these devices were (17.9%). At the time of data collection, children were most familiar with the Siri VA, which had been released in 2011 and was used more widely at that time than the other devices. This was consistent with Lovato and Piper (2015), which reported in 2015 that Siri was used by children more often than other VAs.

Most participants had at least some previous experience with VAs, and a sizeable majority had used a phone with a VA owned by either a family member or the child herself (85.7%). All participants had heard of phones having VAs. Participants were asked to rank themselves as an "expert", "average", or "beginner" at using their favorite VA. Consistent with previous work using similar demographic questions (Anthony et al., 2012; Soni et al., 2019; Woodward et al., 2016), children tended to rank themselves highly: ten said they were experts (35.7%), 12 said they were average users (42.9%), and only six said they were beginners (21.4%).

3.2. Selecting common VA requests

To choose speech tasks children would likely use with VAs, we drew on task examples identified by Lovato et al.'s survey of YouTube videos of children using Siri (Lovato & Piper, 2015). Because other commercial VAs had been developed since Lovato's study, we also conducted our own search for YouTube videos of children using these newer VAs, adopting the search strategy used by Lovato and Piper (2015). We conducted 16 total YouTube searches, consisting of combinations of the words "children", "child", "kid", and "kids" with the four most common VA names (Lovato & Piper, 2015): "Siri", "Cortana", "Alexa", and "Google Home". We saved a link to each relevant video, and when we began to see irrelevant results on a page, we moved on to the next search phrase.

[1] The supplementary corpus can be requested at https://init.cise.ufl.edu/downloads/.

For each video, we documented the search phrase, the video URL, the title of the video, the number of children in the video, estimated age(s), source of the age estimate (e.g., whether it appeared in the video or video metadata or was an experimenter guess), and other notes if applicable. We also recorded the number of results pages returned for each keyword search. Because our goal was to focus on children ages five to ten, we excluded all videos in which the children either appeared or were known to be outside of this age range.

After finalizing this collection of videos, we transcribed the child-VA interactions in each one, including both the speech produced by the child and the response from the VA. We combined the resulting data set with YouTube data from Lovato and Piper (2015) of children's dialogue with Apple's Siri. We also incorporated a corpus of speech from children interacting with additional VAs (including Microsoft Cortana, Amazon Echo, Apple Siri, and Google Android Assistant) collected by Woodward et al. (2018). From these combined data sources, we created an affinity diagram (Lucero, 2015) to cluster the activities children were engaged in when speaking to the VA. Categories included, for example, math questions, asking for entertainment, and learning about the VA, among others. Finally, we used these clusters to create the speech tasks for our study (described in Section 3.3).

3.3. Procedures

All children participated in a warm-up task followed by three data collection tasks. The three core tasks ranged in degree of openness, including unscripted, scripted, and spontaneous speech. We ordered the tasks purposefully to elicit unconstrained speech from children in the warm-up task first and avoid potentially priming the children with the structured tasks that followed. The order of our tasks controlled for testing effects of increased interaction with the device: real-time feedback from the VA was only provided in the final unstructured Conversation task to ensure that previous tasks reflected children's natural inclinations for interaction. All four tasks were audio-recorded and later transcribed by the research team for analysis.

Warm-up: Open-Ended Speech. Based on prior work, we first prompted children to "ask questions" either (a) of their favorite VA (specified in the demographic survey administered before beginning the speech tasks), or, (b) if they did not have prior experience with a VA, for a "*magic phone*" that can answer any question. No live VA was present during this task. The term "*magic phone*" was selected to be child-friendly and aligned with the hypothetical context of the question, but other terms, like "*a talking computer*" or "*a phone that listens*" could also have been used and might have evoked different responses. If the participant did not use direct speech but instead spoke indirectly about the questions they would ask, the experimenter would prompt more specifically: "*How exactly would you say that to the* [name of voice agent | magic phone]*?*" To avoid biasing the types of questions children asked, we did not provide examples. We intended this task to last about five minutes. If children indicated they could not think of any more questions before time was up, we encouraged them to try to "*think of a few more*" based on their experience. If they still could not think of more questions, we allowed them to move on.

Procedure 1: Requests. During this task, we asked children to imagine making a specific request of their favorite VA or a "magic phone" and asked them to phrase this request in their own words. The researcher asked each child to form ten different requests, chosen based on the common VA requests we had previously identified (described in Section 3.2). Five of the ten requests we asked children to formulate were narrow, with well-defined end-state goals. These included asking a VA for help with each of the specific subtasks.

Narrow Requests:

1 Waking up at 10 in the morning
2 Spelling the longest word the child knows
3 Playing the child's favorite song
4 Sending a text message to someone
5 Finding the location of the child's favorite store

The additional five requests were broader and required children to form questions on topics of their choosing. These five subtasks included formulating any question of their choice for a VA about each of the topics below.

Broad Requests:

1 Their favorite animal
2 A math problem they were working on at school
3 Their favorite cartoon character
4 The device itself
5 Anything fun

The order of the ten requests was block randomized across children (with scenarios grouped into four different groups) to avoid possible effects of children's vocal fatigue on recognition of the utterances during later analysis. No live VA was present during this task.

Procedure 2: Scripted Speech. In this task, we asked children to recite 20 specific, predefined scripts, again chosen based on the common VA interactions we had previously identified (described in Section 3.2). Examples of the 20 scripts used in this task include, "*Call nana*", "*Can you give me a hug?*" and "*Turn on living room lights*". The order of the 20 utterances was also block randomized across children, again to account for fatigue effects. The purpose of this task was to provide a corpus of utterances that varied only by participant, not by structure or content, to systematically test recognition accuracy of current VAs. No live VA was present during this task.

Procedure 3: Conversation. Finally, children interacted with a live Amazon Alexa agent running on an Echo device. We selected the Echo as a representative example of a current (at the time of this study) VA, and one in which there is no accompanying screen-based interaction, which would likely require (and allow) different interaction design decisions. Children were instructed to wake Alexa up by saying the wake word ("Alexa") and to wait for the blue circle to light up before asking each question. In this task, we were particularly interested to see if and how children might construct (or reconstruct) their speech when given real feedback and error messages from the VA.

3.4. Technical setup

Children's utterances were recorded using a high-quality Blue Yeti USB microphone, including a stand and a pop filter, which was connected to Audacity, a free and open-source digital audio recording and editing computer software application. For the conversational task, we used a wireless Amazon Echo two-way smart speaker device.

3.5. Study sessions

Study sessions were held during an after-school program at a local elementary school. Parents had previously consented to their child's participation. At the beginning of the session, each child was given the opportunity to decide of their own volition whether they wanted to participate. If they assented, they were asked to rank four different incentive prizes in order of preference, based on prior work establishing that periodic breaks and prizes help to bolster children's completion rates for lengthy empirical studies (Brewer et al., 2013). The prizes we used were

Table 1
Characteristics of the Kids' Voice Assistant Corpus we collected in this study. Data is available for $N = 28$ children for all procedures.

Procedure	No. audio files (total)	No. audio files per child	Audio Length (total) (hh:mm:ss.0)	Audio length (average) per child	Audio length (per utterance) per child
Warm-up Task	28	1	02:25:34.0	00:05:11.9	n/a
Requests	278	10	02:14:16.0	00:04:47.7	4.3 s
Scripted Speech	584	20	00:50:07.0	00:01.47.4	2.3 s
Conversation	28	1	04:37:12.0	00:09:54.0	n/a

small inexpensive toys to motivate and encourage participants to finish all four tasks. After finishing each task, the participant earned a prize of increasing preference, such that the lowest-ranked prize was assigned to the warm-up task and the highest-ranked prize to the Conversation task. After prize selection, we verbally administered a demographic questionnaire, including questions about experience with VAs. We then began with the warm-up task. Each task was followed by an optional break. The entire session lasted about an hour for each child (including the assent process and any breaks).

3.6. Data set characteristics

We collected 10 h and 7 min of audio recordings across all procedures, including the warm-up task. We release the warm-up and Conversation task audio files intact to preserve context and conversational flow. We split the recordings for the Requests and Scripted Speech by utterance, resulting in a total of 278 audio files (10 per participant, 2 participants missing 1 each) for the Requests and 584 audio files (20 per participant, 0 missing) for the Scripted Speech. These utterances ranged in length from 1 s to 27.3 s (average: 4.3 s) for the Requests, and from 0.09 s to 8.1 s (average: 2.3 s) for the Scripted Speech. We provide a summary of metadata for each task in Table 1.

3.7. Data analysis

3.7.1. Transcription accuracy

We define *transcription accuracy* as the percentage of utterances correctly transcribed by the VA (i.e., Alexa). To determine if an utterance was correctly transcribed, audio captured from the study was played back to a 2019 version of an Amazon Echo Dot. We then used the Alexa history information provided on the account section of the Amazon website to obtain Alexa's transcription of the utterance and its response. An utterance is considered to be correctly transcribed if the transcription obtained from Alexa perfectly matches our transcription of the utterance. For the utterances determined to be incorrectly transcribed, Word Error Rate was calculated to quantify the severity of translation errors. Word Error Rate (WER), based on Levenshtein string distance (Levenshtein, 1966), represents the minimum number of insertions, deletions, and substitutions that have to be performed to convert a hypothesis utterance (i.e., the transcription from Alexa) into the reference utterance (i.e., our transcription). To calculate the WER, the total number of substituted (S), deleted (D), and inserted (I) words is divided by the total number of words (N) in the reference utterance as shown below.

$$WER = \frac{S + D + I}{N} \quad (1)$$

3.7.2. Qualitative analysis

Audio recordings of children's Requests and Conversation were transcribed and segmented by utterance. For Requests–which were not spoken directly to a VA during the study session–we recorded an Echo Dot's responses to 100% accurate transcriptions of children's speech, achieved by the lead researcher's re-reading the child's speech aloud until each word was correctly transcribed by the Echo Dot. This enabled us to control for recognition errors and to focus on the VA's response to the structure and content of the child's speech. Transcriptions of the VA responses were paired with the transcription of the corresponding speech that prompted the response. Two researchers then followed an iterative open-coding process to inductively code the transcribed child speech and VA response, meeting regularly for several weeks to compare codes and refine code categories. Final code categories included: the purpose of the child's speech, the correctness of the response, and patterns in the speech, among others. Once codes were finalized, each researcher coded the entire data set separately; the researchers then reconvened to compare results and resolve any inconsistencies.

Table 2
Analyses per data source. We analyzed children's utterances in four ways, examining: the structure of the utterance, the apparent function or purpose of the utterance, a current VA's ability to transcribe the utterance, and a current VA's ability to respond to the utterance appropriately. Not all analyses were appropriate for all data sources (e.g., the research team determined the function of predefined requests and scripted speech, and thus it would not have been appropriate to assess these utterances for their function).

	Narrow requests	Broad requests	Scripted Speech	Conversation
Structure of utterance	X			
Function of utterance				X
Transcription accuracy	X	X	X	X
Appropriateness of VA response	X	X		X

4. Results

We analyzed children's utterances for their structure (i.e., their component parts and organization) and their function (i.e., the apparent purpose or motivation behind the utterance). We also analyzed a current VA's ability to transcribe the words of the utterance and to respond appropriately to it. Not all of our data sources lent themselves to all of these analyses (for example, the scripted speech was defined by the research team, and thus did not reflect the child's motivation). The data source(s) used for each analysis are listed in Table 2 and results of each analysis are described below.

4.1. Children's speech when addressing VAs

We first examined patterns in the ways children spoke to VAs across all three tasks. Here, we report on common formulations and themes in the structure of children's speech. We also analyzed children's speech during the Conversation task for underlying goals, as this was the only task where children had the freedom to speak on any topic.

4.1.1. Common formulations

We observed several common types of formulations of speech that appeared in Conversation and in the Broad Requests, described below.

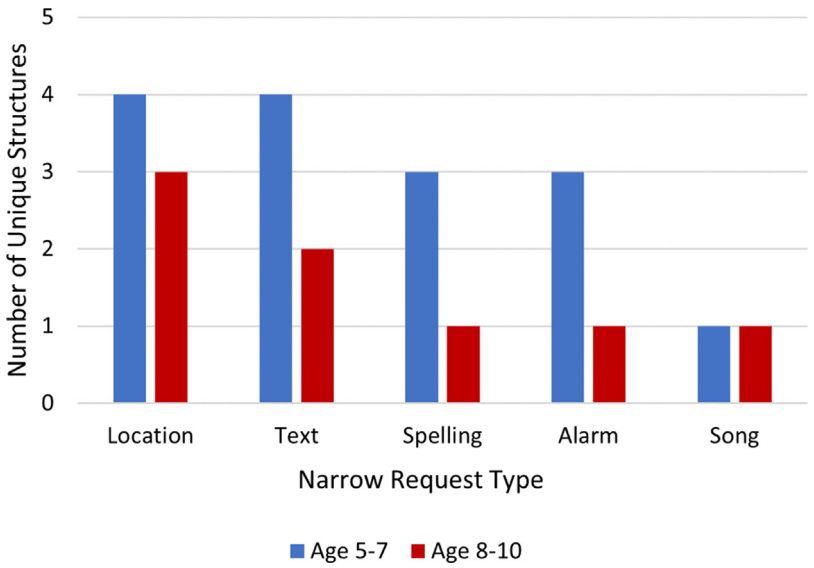

Fig. 1. Younger children used significantly more unique structures than older children.

- *Refined Statements*. Children's speech was often clear and to-the-point; 45.7% of utterances were easily interpretable with no additional context, as judged by the researchers, and scoped to a single specific topic. These included statements and questions like, "*How fast can a hummingbird fly?*" and "*What's 10 minus 10?*" Utterances were coded as "*refined*" if they were entirely clear and complete to a human listener, irrespective of how a VA might respond.
- *Human–Machine Conflation (H–M)*. In 32.1% of utterances, children ascribed human properties to the VA. These included questions and statements like, "*Do you play an instrument?*" and "*What color is your hair?*" In these instances, children's utterances were built on the embedded assumption that human characteristics extend to VAs.
- *Underspecified Details*. Sometimes children formed utterances that left out contextual details that invited follow up (8.7%). For example, children asked, "*How old is, like, the character from Teen Titans?*" and "*Can you please help me with my math problem, Alexa?*" We coded an utterance as underspecified if it invited a specific follow-up question, such as, "**Which** *character from Teen Titans?*" that would prompt the child to replace a generic expression (e.g., "*the character*") with a specific one.
- *Topic Exploration*. In some instances, children asked a question that introduced a topic broadly rather than requesting a narrow and specific piece of information (7.6%). In these cases, they said things like, "*Why are you a magic phone?*", or "*How would you make balloons?*" These questions invited an exploration rather than a targeted response. Some of these questions appeared very difficult or impossible to answer. Such questions included, for example, "*What does a pig do when it's bored?*" and "*How many patterns are there in the world?*" However, they still provided an opportunity for exploration and discussion.
- *Conversational Bids*. Conversational bids were statements that offered the opportunity for conversation without demanding a response (2.3%). Unlike questions and commands, conversational bids provided bits of information about the child or openings for discussion, such as, "*I like to play on the playground*", or "*I want to buy a camera.*"
- *Unintelligible*. Only 1.7% of the utterances were so ambiguous and difficult to parse that they were nearly meaningless to the research team, such as, "*Please get a animals,*" "*What pictures of penguins?,*" and "*Like Matthew found three grapes and he ate one and makes two grapes gone.*"
- *Invalid Premise*. Finally, in a small number of cases, children's speech reflected an underlying misconception about the world (1.6%). For example, one child asked "*How do snakes jump?*", built on the assumption that a snake can jump. Others asked, "*Where can I get a free phone?*" and "*Does [Spongebob] really laugh like that in real life?*"

4.1.2. Structural patterns in children's speech

We also examined children's speech for patterns in structure. Conversation and Broad Requests both allowed children to bring up topics of their own choosing, leading to more varied responses and entirely divergent constructions. However, Narrow Requests, where children's goals were scoped and predefined by the research team–such as setting an alarm or playing a song–led to structural patterns in children's speech. For example, when asking a VA for help waking up, one third of participants used the pattern [action][object][time] saying things like, "*set [action] an alarm [object] for 10 AM [time]*".

Older children's utterances were more likely to conform to a common structure and, collectively, condensed into a small set of forms. Younger children produced a long tail of unique structures that diverged and reflected less consistency across participants. For example, where older children typically asked for help waking up by making statements represented by the structures [action][object][time] or [agent][action][object][time], younger children's statements were less likely to reflect these components and component-orderings. Instead, they made statements like, "*Can you send my mom to say I am out of school, and can you come pick me up? Send*", and "*What are some ways to help me get up?*" A structure was only considered "unique" if the speaker was the *only* participant to use it. This divergence was significant, and across the five Narrow Requests, younger children (age 5 to 7) produced significantly more unique structures than older children (age 8 to 10), $\chi^2(1, N = 28) = 8.08, p = .004$, see Fig. 1.

The number of structured patterns participants collectively produced also varied across requests. Children used the greatest number of variations (N = 16) for the alarm scenario and the fewest for the location scenario (N = 6). Thus, our results suggest that: (1) children's utterances to enact common VA-supported requests have systematic structure, (2) some tasks are

more conducive to using a common pattern than others, and (3) as children get older, they become more likely to use common patterns. We further found that common patterns were more likely than unique patterns to elicit a correct response from the VA, as described in the results section examining VAs' responses (see Section 4.2.3).

4.1.3. Function of conversational speech

Finally, we coded each Conversation utterance for the underlying goal the child appeared to hope to achieve with their speech, as judged by the researchers. We identified five distinct functions of children's speech:

- *Building Relationships*. A plurality of utterances (40%) reflected an interpersonal motive, either seeking to understand personal aspects of the agent or to express the child's own feelings and thoughts. Rather than asking the device for specific information, the core purpose of these utterances was to elicit or share information about the self, for example, "*What is your favorite animal?*"
- *Learning*. The second-most frequent type of goal was obtaining knowledge. We defined "knowledge" as persistent information (such as a historical fact), rather than fleeting information (such as the current temperature). In 34% of utterances, children attempted to cultivate knowledge about a variety of topics, asking the device, for example: "*What is a black hole?*" and "*What do penguins eat?*".
- *Seeking Information.* In contrast to seeking knowledge, 10% of utterances sought information at the specific moment of interaction, such as the weather, time, or population (e.g., "*What time will it get dark tonight?*")
- *Functional*. Another 9% of the utterances aimed to achieve some effect through the device. These statements and questions specified an action for the agent to perform, such as playing a song ("*play calming music*"), or setting an alarm ("*wake me at eight*").
- *Testing the Device's Mind or Knowledge.* In 7% of utterances, it appeared (as judged by the researcher, drawing on all contextual cues from the audio recording) that the child's primary goal was to examine how the device would respond. These questions were often about something the child already knew, such as, "*Do you know what [a] stamp is? There's one right here.*"

4.2. VAs' responses to children

We also evaluated VAs' responses to participants' utterances. We first examined transcription accuracy, measuring the device's ability to recognize children's speech. Then, holding transcription accuracy constant by feeding a perfect transcript to the device, we evaluated the quality of its response to the structure and content of the child's utterance.

4.2.1. Transcription accuracy

Scripted speech. We began by examining transcription accuracy for Scripted Speech, in which utterances were the most controlled across children and thus well-suited to a systematic accuracy analysis. As shown in Table 3, transcription accuracy increased with age. Analysis of Variance (ANOVA) found a significant main effect of age on transcription accuracy ($F_{5,22} = 8.54, p < .001$). Tukey post-hoc comparisons using Bonferroni correction showed the 5 and 6-year-old groups had significantly lower transcription accuracy than the 8, 9, and 10-year-old age groups ($p < .01$ in all cases). Post-hoc analysis also showed accuracy for the 7-year-old age group to be significantly lower than the 10-year-old group ($p < .01$). There were no other significant differences between age groups. We also observed a significant strong positive correlation between age and transcription accuracy ($r(28) = 0.80, p < .0001$), suggesting that as age increased, so did transcription accuracy.

Table 3
Transcription accuracy and word error rates (WER) means (standard deviations in parentheses) by age for Scripted Speech.

Age	Accuracy	Word Error Rate (WER)
5	20.0% (0.0%)	0.64 (0.33)
6	28.0% (5.7%)	0.70 (0.33)
7	41.9% (15.6%)	0.55 (0.33)
8	61.4% (16.0%)	0.51 (0.31)
9	61.2% (18.9%)	0.36 (0.28)
10	78.3% (11.5%)	0.33 (0.20)

Next, we analyzed WER of transcription errors by age group. An ANOVA found a significant main effect of age on WER ($F_{5,22} = 6.26, p < .001$). Tukey post hoc comparisons using Bonferroni correction showed 5- and 6-year-old age groups had significantly higher WER than 9- and 10-year-old age groups ($p < .05$ in all cases). Post hoc analysis showed no other significant differences between age groups. A Pearson correlation between age and WER showed a significant negative relationship between age and WER ($r(28) = -.72, p < .001$), i.e., as age increased, WER decreased.

Requests and conversation. While the Scripted Speech provides an understanding of speech recognition accuracy when children recite pre-specified utterances, it may not represent real-world use where children are free to choose their vocabulary. Therefore, we also examined transcription accuracy and WER for Requests and Conversation. Transcription accuracy for Requests ranged from 20%–100% ($mean = 59.6\%, s.d. = 19.0\%$) across children. However, an ANOVA found no significant main effect of age on transcription accuracy ($F_{5,22} = 1.29, n.s.$). For the Conversation task, transcription accuracy ranged from 33%–100% ($mean = 84.1\%, s.d. = 15.9\%$) across children. A Shapiro–Wilk normality test showed our data was not normal ($W = 0.86, p < .01$); therefore, we performed a Kruskal–Wallis Rank Sum test which showed no significant effect of age on transcription accuracy ($\chi^2 = 5.58, df = 5, n.s.$). Similar to transcription accuracy, an ANOVA showed no significant effect of age on WER for either Requests ($F_{5,22} = 1.752, n.s.$) or Conversation ($F_{5,22} = 1.32, n.s.$). Thus, while younger children had lower transcription accuracy and a higher WER for Scripted Speech, there were no significant differences between age groups in tasks where children were free to choose their utterances. To further understand accuracy across tasks, we performed a repeated-measures ANOVA (RM-ANOVA) to explore the effect of age and task on transcription accuracy. A Shapiro–Wilk normality test showed our data was not normal ($W = 0.913, p < .001$); therefore, we applied an Aligned Rank Transform (ART) (Wobbrock, Findlater, Gergle, & Higgins, 2011) to the data. The RM-ANOVA showed significant effect of task ($F_{2,44} = 67.2, p < .001$) on transcription accuracy, but not age ($F_{5,22} = 1.34, n.s.$). Post-Hoc analysis using Bonferroni correction showed transcription accuracy of the request task to be significantly lower than that of the scripted and conversation tasks ($p < .001$ in both cases). No other pairwise differences were observed. This suggests that age differences in Scripted Speech might be influenced by the vocabulary of the utterances, and that participants were able to adapt their utterances when provided feedback from the agent.

4.2.2. Types of VA responses

We next examined the appropriateness of an Amazon Echo Dot's responses to the self-formulated utterances children produced during the Requests and Conversation tasks. We saw that the Echo's responses clustered into five categories:

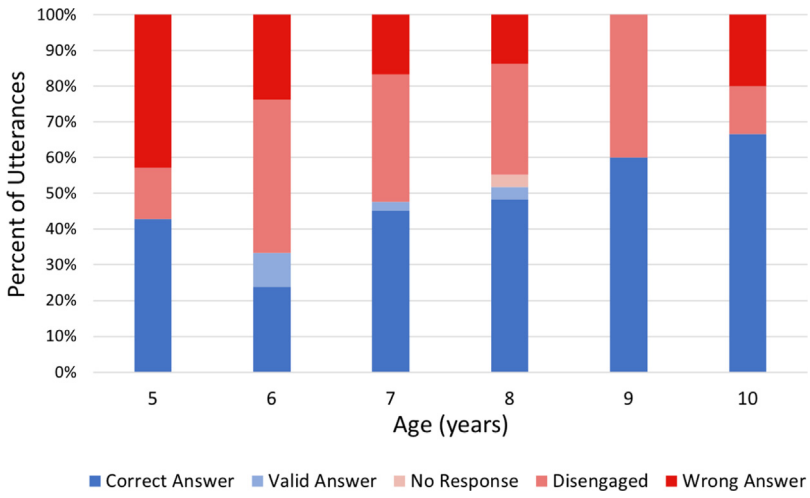

Fig. 2. Appropriateness of VA responses to children's **Broad** Requests by child age.

- *Correct Answers*. In these instances, the Echo replied with a statement that was conversationally consistent with the child's utterance and provided the requested information or desired action. For example, "*'Fight Song,' by Rachel Platten, starting now, on Amazon Music*", was considered a correct answer in response to the utterance, "*Can you put on 'Fight Song'?*"
- *Valid Answers*. In these instances, the Echo did not perform the requested action or return the requested information, but it informed the user that it could not do so, explaining why, offering alternatives, or otherwise indicating that it understood the child's goal. For example, "*I couldn't find that one, but I'm working on adding more local businesses*", was considered a valid (but not correct) response to, "*Where is the nearest [local grocery store chain]?*" In these cases, although children were effective in engaging the device in a valid conversational exchange, they were unable to achieve their end-goal.
- *No Response*. In a few instances, the Echo did not respond verbally and was either silent or beeped, despite detecting and transcribing the utterance.
- *Disengagement*. In these instances, the Echo returned an opaque answer indicating it was unable to engage further, such as, "*I don't have an opinion on that*", "*Hmm, I'm not sure*", or "*I'm sorry I can't help with that*". These responses were ambiguous, and the VA did not specify why it was disengaging.
- *Wrong Answers*. Finally, in other instances, the Echo provided a substantive reply or took meaningful action, but in a way that was not conversationally valid or was misaligned with what the child had said. For example, when one child asked, "*Alexa, can you do anything about my hair?*", the Echo replied, "*Here's a fact about Hair. The hair is a constitutional part of the skin, the epidermis, and the pilosebaceous unit.*" Though detailed and informative, this tangential answer did not directly address the child's original question. Similarly, in another instance, the child commanded, "*Count by fives*", and the Echo replied: "*Here's a station for you based on the song 'Count by Fives' by Pinkfong, on Amazon Music*", taking action, but not the requested action.

We use this taxonomy to analyze the appropriateness of the Echo's responses to children's speech for both the Request and Conversation tasks. To do so, we computed appropriateness by converting the above taxonomy into a 5-point scale, with fully correct answers as most appropriate (5) and wrong answers scored as least appropriate (1).

4.2.3. Appropriateness of VA responses to requests

VAs were only partially successful in responding appropriately to children's speech during the Requests task. Of the 262 Requests children produced, only 53% led to a correct or valid response, while the rest returned a disengaged response, wrong answer, or no response at all. In addition, we evaluated the appropriateness of VA responses as a function of child age. We found a significant positive correlation between age (in years) and average percent appropriateness across all Requests ($r(28) = .44, p = .02$). We also examined this relationship by calculating, for each child, the fraction of utterances that produced a fully correct response. Here again, we found a significant positive correlation between age and the likelihood of eliciting a fully correct response ($r(28) = .47, p = .01$).

Separately, we examined appropriateness as a function of the type of Request. Across the five Narrow Requests, 59% of utterances produced a fully correct response (see Fig. 3), as compared to 47% of Broad Requests (see Fig. 2), a significant difference ($\chi^2(1, N = 28) = 3.86, p = .05$). In evaluating the appropriateness of responses to each type of Narrow Request (see Fig. 4), we returned to the common structures that we saw across participants (see Section 4.1.2) to examine how the structure of an utterance relates to the appropriateness of the VA's response. As mentioned above, using a common structural pattern was significantly more likely to elicit a correct or valid response than using a unique pattern ($\chi^2(1, N = 28) = 5.55, p = .02$), meaning that as children's utterances begin to conform to common structural patterns, VAs become more likely to deliver an appropriate response. This suggests that children are more successful in eliciting a correct response from VAs when they: (1) attempt narrowly scoped, closed-ended tasks, (2) use a common structure ("common" in that it appeared frequently in our sample), and (3) are older.

4.2.4. Appropriateness of VA responses to conversation

We found that children also had mixed success eliciting accurate, meaningful responses from the Echo during Conversation. Of the 516 utterances made across all 28 children, only 261 (51%) produced a correct response. The other 49% were either mistranscribed by the speech recognition system ($N = 72$, 14%) or led to a valid response, disengaged response, wrong answer, or no response from the device ($N = 183$, 35%). As with Requests, the device responded more appropriately to the Conversation utterances of older children, and we found a significant positive correlation between the percentage of correct responses a child received from the VA and the child's age in years ($r(28) =$

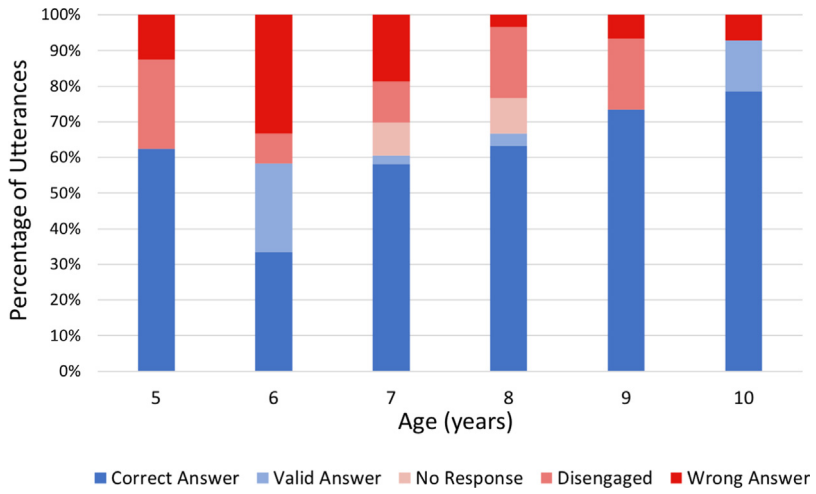

Fig. 3. Appropriateness of VA responses to children's **Narrow** Requests by child age.

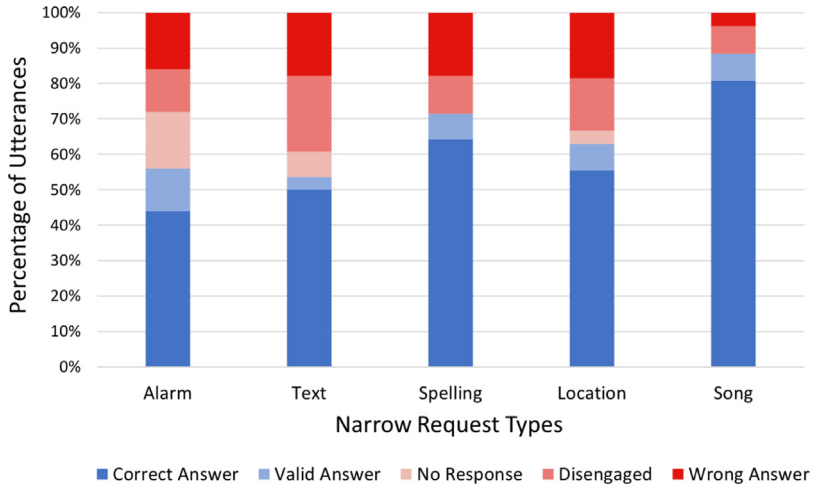

Fig. 4. Appropriateness of VA responses to children's speech across the five Narrow Requests.

.135, $p = .002$). We also tested whether there were significant differences in success rates across the types of utterance functions (see Section 4.1.3). A χ^2 test of the appropriateness of response by function-type did not reveal a significant relationship, suggesting that no specific conversational goals were more likely to generate appropriate responses from VAs than others.

5. Discussion

We saw a number of themes in the ways children constructed VA-directed utterances and in the ways the VA responded to this input. Consistent with prior work (Lovato et al., 2019), we found that children's conversational speech was transcribed accurately more than 84% of the time. When the research team controlled the exact words spoken by children, we saw age-dependent differences, but when children were given the freedom to speak in their own words, we saw no differences by age, and 5-year-olds were just as successful as their 10-year-old peers in producing speech a VA could transcribe accurately. Although historical speech recognition technology has struggled to recognize children's pitch (Gerosa et al., 2007; Kennedy et al., 2017; Li & Russell, 2002; Liao et al., 2015; Potamianos et al., 1997) and pronunciation (Li & Russell, 2002), our results indicate that today's VAs are built on a technical foundation that is sufficiently sophisticated to accommodate young children as first-class users of these systems. And as speech recognition continues to improve, designers can anticipate building on reasonably accurate word recognition when creating VA-enabled experiences for children.

VAs were less successful in responding appropriately to children's speech than they were in transcribing (or recognizing) it. VAs responded meaningfully and accurately to children only about half the time, and they were more likely to do so when responding to older children and when responding to common, narrowly scoped commands. By age ten, children's attempts at these common requests, like setting an alarm or getting directions, led to a correct or valid answer more than 90% of the time. Some tasks were easier for children than others; for example, nearly 90% of participants across all ages were successful when asking a VA to play a particular song. These results suggest that designers targeting young children can expect users to have an emergent ability to command these systems and to be effective at performing common tasks. Prior work has shown that users tend to engage VAs repetitively for a small set of common tasks (Beneteau, Guan et al., 2020; Bentley et al., 2018; Cowan et al., 2017; Sciuto, Saini, Forlizzi, & Hong, 2018); thus, our results suggest that VAs may already respond appropriately to children in many of the most common usage scenarios.

5.1. Designing to support children's use of VAs

Our results also revealed several patterns in children's utterances that suggest design opportunities for enriching children's experiences with these interfaces. For example, children's *Topic Exploration* utterances included broad questions like, "*What if we didn't have animals on earth; what would it be like?*" The device struggled with these exploratory questions and was far more successful responding to narrowly scoped questions with a precise, known answer. However, asking questions of more knowledgeable others is an essential way in which children learn and grow (Callanan & Oakes, 1992; Lever & Sénéchal, 2011; Vygotsky, 1980). Prior work shows that asking questions (Lever & Sénéchal, 2011) and exploring causal relationships (Callanan & Oakes, 1992) with an adult are highly productive mechanisms for expanding children's understanding of the world. Our results suggest that designers have an untapped opportunity to enable VAs to support topic exploration together with child users, as children in our study spontaneously chose to ask broad exploratory questions of VAs. By designing interactions that encourage this exploration, respond to and ask open-ended questions of users, and gradually refine the direction of the conversation, VAs would engage children in a form of dialogue that is known to be both enjoyable and profoundly useful for children as they grow.

In other instances, we saw that children directed statements to the VA that served to initiate conversation without making a specific demand, such as, "*I like to slide*", or "*I want to buy a camera*". As with exploratory questions, we found that the VA did not respond meaningfully to these conversational bids. Here again, prior work has shown that adults make use of these types of openings to engage children in developmentally useful exchanges (Wanska & Bedrosian, 1986). By designing VAs to respond in conversationally appropriate ways to children's bids, designers could support valuable interaction experiences. To do so, designers can leverage evidence-based techniques from parent–child interaction, such as inviting a child to expand on a child-initiated topic (Hoff-Ginsberg, 1987), that support developmentally useful dialogue. Finally, we saw children's utterances at times leave out key contextual details, making these questions and statements ambiguous. This is consistent with findings from prior work suggesting the device provide *discourse scaffolding*, that is, explanation of what exactly it finds ambiguous (Beneteau et al., 2019).

5.2. Understanding the structure of children's speech

We saw that VAs were more likely to respond appropriately to older children than younger children. In the Narrow Requests task, this was mediated in part by the fact that older children structured their utterances using a small set of common patterns. Prior work has shown that knowledge of social scripts increases with age, wherein children become more familiar with sequences of language and behavior associated with specific everyday activities, such as getting ready for school or buying groceries (Goodman, Duchan, & Sonnenmeier, 1994; Short-Meyerson & Abbeduto, 1997). Our results are consistent with the idea that common VA interactions – such as playing a song or setting an alarm – have associated scripts (learned through experience and interaction with similar devices) that are understood by users and devices alike. As children grow older, they become more proficient with a wider variety of social scripts, enabling them to devote fewer cognitive resources to structuring their speech and allowing them to have more sophisticated conversations (Furman & Walden, 1990; Nelson & Gruendel, 1979). Designers of VA experiences could capitalize on this phenomenon by leveraging existing scripts and guiding users toward new ones, thus giving users and VAs more common ground during their interactions. For example, designers might aim to create experiences for younger children that have simple, widely used scripts, as even the youngest participants in our study were successful in asking a VA to play a song, an action supported by a simple speech pattern. VAs might also provide scaffolding to demonstrate successful structural patterns. Designers should anticipate that younger users will produce a wider, more diverse set of structures that may require working iteratively with the VA to identify the intended interaction.

5.3. Children's interactions with personified interfaces

Finally, there is a great deal of interest, both in the research community and in mainstream discourse, about whether and how interacting with personified interfaces affects children. Some studies have surfaced the potential advantages of forming parasocial relationships with personified interfaces (Brunick, Putnam, McGarry, Richards, & Calvert, 2016; Coninx et al., 2016), while other work has raised concerns about children's willingness to disclose personal and sensitive data to these systems (Kahn, Friedman, Perez-Granados, & Freier, 2006; McReynolds et al., 2017; Williams, Machado, Druga, Breazeal, & Maes, 2018) and to take direction and respond to peer pressure from digital agents (Williams et al., 2018). Our results show, first, that a large percentage of children's VA-directed utterances seek to establish and expand the child's relationship with the interface, even in a short-term lab context. Further, VAs were largely unsuccessful in responding appropriately to these types of interactions, producing a correct response only half the time. The mainstream VA we tested currently offers inconsistent responses to questions that imply it has human-like characteristics: at times, it played into invalid assumptions, telling children, for example, that its favorite car is Lightning McQueen and that it loves to read, but more often disengaging without giving a meaningful reply.

Participants' tendency to ask such questions is consistent with prior work showing that children readily attribute mental and social attributes to intelligent systems (Druga et al., 2017, 2018; Kahn et al., 2012; Lovato et al., 2019; Woodward et al., 2018) and report greater satisfaction when agents in these systems are personified (Purington, Taft, Sannon, Bazarova, & Taylor, 2017; Yarosh et al., 2018). Designers should anticipate that children will seek out parasocial relationships with VAs and design for this eventuality. Future work to: (1) closely examine the statements children make when seeking to cultivate these relationships, and (2) design, develop, and evaluate responses to these relationship-building statements would be valuable. Given the sensitivity of this design question, it would be particularly appropriate for researchers to design responses to relationship-building utterances through participatory practices that prioritize children's perspectives (Fails, Guha, & Druin, 2013).

Prior work also shows that other digital technologies routinely leverage children's parasocial relationships for profit (Radesky et al., 2022), for example, using digital characters to pressure children into making purchases or extending usage time. We advocate designing to enable relationships between VA personas and child users *only* with the end goal of supporting and benefiting the child. It is imperative that designers make such a commitment, given the frequency with which we saw participants spontaneously engage in relationship-building. Designers should also anticipate that interpersonal biases and pressures (such as social reciprocity Fradkin, Grewal, Holtz, & Pearson, 2015; Uehara, 1995 or ingroup bias Koval, Laham, Haslam, Bastian, & Whelan, 2012) will translate to this usage context and systematically shape children's interactions with the system.

5.4. Limitations and future work

Due to our small sample size per age group, we cannot make strong claims regarding how the age differences affect the variation of VAs' accuracy and effectiveness. We identify some initial trends that suggest age differences, but future work remains to characterize these trends across a broader sample of children with experience with the latest VA technologies. We also asked children to address a hypothetical (rather than live) VA, and their utterances likely would have changed in the context of responsive feedback and interactivity. This had the benefit of decoupling our data from any one VA implementation, but it also has the drawback of producing speculative rather than *in situ* data. Separately, future work drawing on the traditions of Ethnomethodology and Conversation Analysis (EMCA) (Porcheron et al., 2018) to analyze longer dialogues between children would provide a valuable complement to the data we present here. We invite other researchers to use and contribute to our public corpus to investigate how VAs could respond to children in different conversational settings. And finally, we note that this data was collected before the onset of the COVID-19 pandemic, which disrupted our research pipeline and delayed this project.

6. Conclusion

In this study, we examined child-VA interaction from several angles. We report on a three-part lab study in which children between the ages of five and ten years old: (1) formed utterances to attempt common VA tasks, (2) recited scripted speech, and (3) engaged in unstructured conversation with one commercially available VA. By examining children's utterances in this mix of contexts, and by evaluating a VA's responses to these utterances, we identify themes in how children address VAs and how appropriately VAs respond to this input. We find that the VA we tested frequently recognizes children's speech accurately but only answers with a meaningful, contextually relevant response half of the time. Several factors predict the likelihood of a child receiving a meaningful response, and responses are more likely to be correct when: the child is older, the child is attempting a common task, and the child organizes their speech into a structure that is commonly used by others for the same task. When children were given the freedom to interact with a VA in any way they chose, their utterances were more likely to be social than task-oriented, highlighting the importance of understanding children's parasocial relationships with these devices. Although children are already effective in using VAs for common tasks, we demonstrate untapped design opportunities for VAs to respond to children's conversational bids and explore topics of interest with them collaboratively.

Selection and participation of children

Children ages five to ten were recruited for this study from an after-school program at a local area K-12 school, with permission from the program and school. Parents were approached during drop-off and pick-up time by researchers to describe the study, distribute informational packets, and answer questions. Parents who consented to have their children participate returned the packets to the researchers, the after-school program, or the school staff. Permitted children were pulled out of the normal after-school program activities and asked to assent to the research of their own volition before beginning. The sessions were held at the after-school program location, in a setting familiar to the children, in a quiet private room with at least two researchers present. After the session, children returned to their regularly scheduled program activities.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ayushi Jain reports a relationship with Amazon.com Inc that includes: employment.

Data availability

The data is shared as public corpus.

Acknowledgments

We would like to thank Isaac Wang and Aishat Aloba for their help with this research. We are also grateful to the P.K. Yonge Developmental Research School. This work was supported in part by the National Science Foundation, USA under awards IIS-1552598 and CNS-1560243 to the University of Florida. This work was also supported in part by a Jacobs Foundation Fellowship, Germany to Alexis Hiniker.

References

Anthony, L., Brown, Q., Nias, J., Tate, B., & Mohan, S. (2012). Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices. In *Proceedings of the 2012 ACM international conference on interactive tabletops and surfaces* (pp. 225–234). ACM.

Beirl, D., Yuill, N., & Rogers, Y. (2019). Using voice assistant skills in family life.

Beneteau, E., Boone, A., Wu, Y., Kientz, J. A., Yip, J., & Hiniker, A. (2020). Parenting with Alexa: exploring the introduction of smart speakers on family dynamics. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13).

Beneteau, E., Guan, Y., Richards, O. K., Zhang, M. R., Kientz, J. A., Yip, J., et al. (2020). Assumptions checked: How families learn about and use the echo dot. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(1), 1–23.

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., & Hiniker, A. (2019). Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13). ACM.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–24.

Berman, R. A. (2007). Developing linguistic knowledge and language use across adolescence.

Brewer, R., Anthony, L., Brown, Q., Irwin, G., Nias, J., & Tate, B. (2013). Using gamification to motivate children to complete empirical studies in lab environments. In *Proceedings of the 12th international conference on interaction design and children* IDC '13, (pp. 388–391). New York, NY, USA: Association for Computing Machinery.

Brooks, P. J., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, 720–738.

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

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233.

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. In *Proceedings of the 17th ACM conference on interaction design and children* (pp. 337–348). ACM.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Coninx, A., Baxter, P., Oleari, E., Bellini, S., Bierman, B., Henkemans, O. B., et al. (2016). Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *Journal of Human-Robot Interaction*, *5*(1), 32–67.

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., et al. (2017). What can I help you with?: Infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services* MobileHCI '17, (pp. 1–12). New York, NY, USA: Association for Computing Machinery.

Dąbrowska, E., & Lieven, E. (2005). *Towards a lexically specific grammar of children's question constructions*. Walter de Gruyter.

Dodd, B., Holm, A., Hua, Z., & Crosbie, S. (2003). Phonological development: a normative study of British English-speaking children. *Clinical Linguistics & Phonetics*, *17*(8), 617–643.

Druga, S., Vu, S. T., Likhith, E., & Qiu, T. (2019). Inclusive AI literacy for kids around the world. In *Proceedings of FabLearn 2019* (pp. 104–111). ACM.

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). Hey google is it OK if I eat you?: Initial explorations in child-agent interaction. In *Proceedings of the 2017 conference on interaction design and children* (pp. 595–600). ACM.

Druga, S., Williams, R., Park, H. W., & Breazeal, C. (2018). How smart are the smart toys?: Children and parents' agent interaction and intelligence attribution. In *Proceedings of the 17th ACM conference on interaction design and children* IDC '18, (pp. 231–240). New York, NY, USA: ACM.

Fails, J. A., Guha, M. L., & Druin, A. (2013). Methods and techniques for involving children in the design of new technology for children. *Foundations and Trends in Human-Computer Interaction*, *6*(2), 85–166.

Fradkin, A., Grewal, E., Holtz, D., & Pearson, M. (2015). Bias and reciprocity in online reviews: Evidence from field experiments on airbnb. *EC*, *15*, 15–19.

Furman, L. N., & Walden, T. A. (1990). Effect of script knowledge on preschool children's communicative interactions. *Developmental Psychology*, *26*(2), 227.

Geeng, C., & Roesner, F. (2019). Who's in control?: Interactions in multi-user smart homes. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (p. 268). ACM.

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, *49*(10–11), 847–860.

Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009). A review of ASR technologies for children's speech. In *Proceedings of the 2nd workshop on child, computer and interaction* (pp. 1–8).

Goodman, G. S., Duchan, J. F., & Sonnenmeier, R. M. (1994). In J. F. Duchan, L. E. Hewitt, & R. M. Sonnenmeier (Eds.), *Children's development of scriptal knowledge* (pp. 120–133). Englewood Cliffs, NJ: Prentice Hall, Ch. 9.

Gossen, T., Kotzyba, M., Stober, S., & Nürnberger, A. (2013). Voice-controlled search user interfaces for young users. In *7th annual symposium on human-computer interaction and information retrieval* (pp. 2–5).

Hoff-Ginsberg, E. (1987). Topic relations in mother-child conversation. *First Language*, *7*(20), 145–158.

Kahn, P. H., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies*, *7*(3), 405–436.

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., et al. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, *48*(2), 303.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., et al. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* HRI '17, (pp. 82–90). Association for Computing Machinery.

Koval, P., Laham, S. M., Haslam, N., Bastian, B., & Whelan, J. A. (2012). Our flaws are more human than yours: Ingroup bias in humanizing negative characteristics. *Personality and Social Psychology Bulletin*, *38*(3), 283–295.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lever, R., & Sénéchal, M. (2011). Discussing stories: On how a dialogic reading intervention improves kindergartners' oral narrative construction. *Journal of Experimental Child Psychology*, *108*(1), 1–24.

Li, Q., & Russell, M. J. (2002). An analysis of the causes of increased error rates in children's speech recognition. In *Seventh international conference on spoken language processing* (pp. 1–4).

Liao, H., Pundak, G., Siohan, O., Carroll, M., Coccaro, N., Jiang, Q.-M., et al. (2015). Large vocabulary automatic speech recognition for children. In *Interspeech* (pp. 1611–1615).

Lovato, S., & Piper, A. M. (2015). Siri, is this you?: Understanding young children's interactions with voice input systems. In *Proceedings of the 14th international conference on interaction design and children* (pp. 335–338). ACM.

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. In *Proceedings of the 18th ACM international conference on interaction design and children* (pp. 301–313). ACM.

Lucero, A. (2015). Using affinity diagrams to evaluate interactive prototypes. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2015* (pp. 231–248). Springer International Publishing.

McReynolds, E., Hubbard, S., Lau, T., Saraf, A., Cakmak, M., & Roesner, F. (2017). Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 5197–5207). ACM.

Morrow, V., & Richards, M. (1996). The ethics of social research with children: An overview. *Children & Society*, *10*(2), 90–105.

Nelson, K., & Gruendel, J. M. (1979). At morning it's lunchtime: A scriptal view of children's dialogues. *Discourse Processes*, *2*(2), 73–94.

Nippold, M. A. (1998). *Later language development: The school-age and adolescent years*. ERIC.

Nippold, M. A. (2004). Research on later language development: International perspectives. In *Language development across childhood and adolescence, Vol. 3* (pp. 1–8).

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (p. 640). ACM.

Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. In *Fifth European conference on speech communication and technology* (pp. 1–4).

Punch, S. (2002). Research with children: The same or different from research with adults? *Childhood*, *9*(3), 321–341.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF": Social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems* (pp. 2853–2859). New York, NY, USA: Association for Computing Machinery.

Radesky, J., Hiniker, A., McLaren, C., Akgun, E., Schaller, A., Weeks, H. M., et al. (2022). Prevalence and characteristics of manipulative design in mobile applications used by children. *JAMA Network Open*, *5*(6), 1–11.

Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition*, *104*(1), 106–134.

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). Hey Alexa, What's Up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 designing interactive systems conference* (pp. 857–868). ACM.

Shobaki, K., Hosom, J.-P., & Cole, R. (2007). CSLU: Kids' speech version 1.1. *Linguistic Data Consortium*, URL https://catalog.ldc.upenn.edu/LDC2007S18.

Short-Meyerson, K. J., & Abbeduto, L. J. (1997). Preschoolers' communication during scripted interactions. *Journal of Child Language*, *24*(2), 469–493.

Soni, N., Gleaves, S., Neff, H., Morrison-Smith, S., Esmaeili, S., Mayne, I., et al. (2019). Do user-defined gestures for flatscreens generalize to interactive spherical displays for adults and children? In *Proceedings of the 8th ACM international symposium on pervasive displays* PerDis '19, (pp. 1–7). New York, NY, USA: Association for Computing Machinery.

Statista (2019). Digital assistant and voice assistant adoption 2019. URL https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.

Statista (2022). Number of voice assistants & world population 2023. URL https://www.statista.com/statistics/1034436/worldwide-number-voice-assistant-human-population/.

Stoel-Gammon, C., & Cooper, J. A. (1984). Patterns of early lexical and phonological development. *Journal of Child Language*, *11*(2), 247–271.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209–253.

Uehara, E. S. (1995). Reciprocity reconsidered: Gouldner's moral norm of reciprocity and social support. *Journal of Social and Personal Relationships*, *12*(4), 483–502.

Vihman, M. M., DePaolis, R. A., & Keren-Portnoy, T. (2009). Babbling and words: A dynamic systems perspective on phonological development. In *The Cambridge handbook of child language* (pp. 163–182). Cambridge, England: Cambridge University Press.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wanska, S. K., & Bedrosian, J. L. (1986). Topic and communicative intent in mother–child discourse. *Journal of Child Language*, *13*(3), 523–535.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., et al. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing (TSLP)*, *7*(4), 18.

Williams, R., Machado, C. V., Druga, S., Breazeal, C., & Maes, P. (2018). My doll says it's ok: a study of children's conformity to a talking doll. In *Proceedings of the 17th ACM conference on interaction design and children* (pp. 625–631). ACM.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* CHI '11, (pp. 143–146). New York, NY, USA: Association for Computing Machinery.

Wobbrock, J., & Kientz, J. (2016). Research contribution in human-computer interaction. *Interactions*, *23*, 38–44.

Woodward, J., McFadden, Z., Shiver, N., Ben-hayon, A., Yip, J. C., & Anthony, L. (2018). Using co-design to examine how children conceptualize intelligent interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (p. 575). ACM.

Woodward, J., Shaw, A., Luc, A., Craig, B., Das, J., Hall, P., et al. (2016). Characterizing how interface complexity affects children's touchscreen interactions. In *Proceedings of the 2016 CHI conference on human factors in computing systems* CHI '16, (pp. 1921–1933). New York, NY, USA: Association for Computing Machinery.

Yarosh, S., Thompson, S., Watson, K., Chase, A., Senthilkumar, A., Yuan, Y., et al. (2018). Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM conference on interaction design and children* (pp. 300–312). ACM.