Wizards in the Middle: An Approach to Comparing Humans and Robots*

Zachary Henkel, Kenna Baugus Henkel and Cindy L. Bethel[1]

Abstract— **While Wizard-of-Oz (WOz) techniques are frequently used to supplement a machine's abilities, extending this approach to human entities can increase experimental control in studies comparing evaluations of humans and machines in the same role. This article describes the design, implementation, and use of a WOz system for facilitating controlled verbal interactions between children and robot or human interviewers. A collaborative interface allows multiple remote wizards to combine participant responses with interaction-specific goals in order to direct a robot or human interviewer's behavior in a consistent manner. While robot interviewers are controlled directly, human interviewers receive direction through a tablet device or via a projection system concealed from participants. In addition to the system's technical design, we describe the division of responsibilities between wizards and insights from using the system across three extensive interview studies to facilitate a total of 217 interactions with children.**

I. INTRODUCTION

The Wizard of Oz (WOz) technique, in which a machine perceived by users to function autonomously is covertly aided by humans [1], has enjoyed widespread use by researchers and system designers across the fields of human-computer interaction (HCI) and human-robot interaction (HRI). While the WOz technique is valuable in allowing interactions to be explored without the need for high technical investment, it can also act as mechanism for providing consistency in lab studies.

While conducting research comparing social robots and humans as forensic interviewers we found that in comparison to humans, robot interviewers operating via WOz were more easily able to adhere to best practice protocols intended to eliminate biases because their behaviors were guided by a team of wizards [2]. The need to reduce unwanted behavioral variability between human and robot interviewers led to the development of a "wizards in the middle" (WITM) paradigm that allows both machines and humans to leverage the resources and consistent direction of well-trained wizards.

Figure 1 shows how the WITM paradigm expands the traditional WOz approach by providing a wizard layer to process input and provide output cues for all types of agents, including humans. This technique strengthens experimental designs by allowing isolation between the observable agent engaging in an interaction and the underlying resources generating the behavior.

This article details the design and use of a WOz system that uses the WITM paradigm to support interview-style

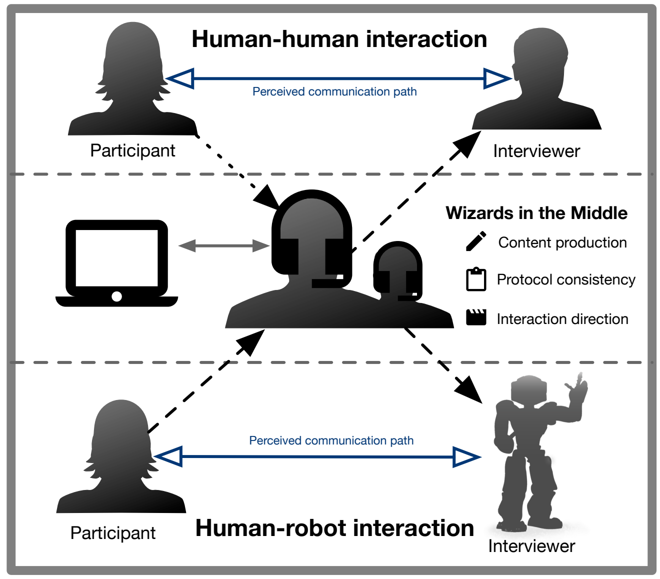

Fig. 1: The "wizards in the middle" (WITM) paradigm allows separation of the delivering agent and processing agent(s) (wizards), regardless of their forms (machine or human).

interactions between a human and either a robot or human interviewer. The work presents two unique contributions: (1) an approach for directing human and robot agents through the same wizarding system and (2) a collaborative multi-wizard interface used to facilitate fluid interactions with 217 children across three studies.

II. BACKGROUND AND MOTIVATION

As the WOz technique has been elemental in the development of systems across disciplines, researchers have continued to extend and study the technique itself [3]–[7]. This section summarizes prior WOz-related research that is most relevant to our work, followed by an explanation of forensic interviewing and how using robots for this task motivated the development of the WITM paradigm.

A. Wizard-of-Oz Techniques

The WOz technique was first reported by Kelley as a method for allowing the study of a natural language interface for a calendar application [1]. The approach became a valuable tool for facilitating investigations of HCI and HRI in which humans could stand-in for computational components that were not yet trivially accomplished by machines but whose functions were essential to the overall interaction.

*This research was sponsored by NSF Award #1408672.

[1]The authors are with the Department of Computer Science and Engineering, Mississippi State University, Mississippi State, Mississippi, USA `zmh68,kbb269,clb821@msstate.edu`

Research across fields dealing with human-machine interactions has produced a rich variety of specific techniques, frameworks, and guidelines related to the use of WOz [8]–[14]. The frequent use of WOz for verbal or language understanding tasks (72% in a 2012 review of WOz HRI studies [15]) has led to work focused on improving the speed of wizard responses. For example, a study comparing interfaces with either free text input or preset response buttons found each had pitfalls [3], necessitating a more nuanced design like pre-scripted statements with variable elements [16], [17]. In addition to improving wizard input interfaces, researchers have also examined using partially unreliable autonomous systems with wizards acting as a filter to mitigate incorrect output [12].

As the cognitive capacity of a single wizard is limited [18], more complex efforts have leveraged multiple wizards coordinating with one another to allow a system to perform beyond what a single wizard could reliably achieve [6]. Though there are complexities associated with this approach [19], it has been implemented successfully in a variety of studies; examples include separation of speech processing and physical tele-operation among wizards in HRI studies [5], [20], [21] and separation of domains among wizards in chat systems [22].

Our research builds upon prior insights and contributes a paradigm and implementation that allows multiple wizards to collaboratively produce fluid interactions for both human and robot agents.

B. Forensic Interviews

We developed the WITM approach to facilitate studies comparing humans and robots acting as interviewers following a well-investigated forensic interview protocol (NICHD Investigative Interview Protocol), which is frequently used in the investigation of sensitive situations [23]. Our studies focused on using the method to obtain information about experiences with bullying at school and in eyewitness memory scenarios. The protocol emphasizes that interviewers provide open-ended prompts and act carefully to avoid undue influence. The protocol progresses in stages, including: establishing guidelines, building rapport, narrative practice, and examining the substantive issue [23].

Though neutrality is critical, building and maintaining rapport is also essential. While humans are able to build rapport in a spontaneous fashion, this approach can be problematic for neutrality and experimental repeatability. Robots seeking to build rapport must consult some knowledge base (wizard facilitated in our case) to produce appropriate output. Since simple differences in rapport building can rapidly impact perceptions of the interviewer [24], we sought to limit these differences by providing both interviewers the same wizard-generated content and direction.

III. SYSTEM DESIGN

Our system was designed using the WITM paradigm to allow humans and robots to conduct forensic interviews in a behaviorally consistent manner. This section first describes the scope of interactions and the strategy used for content development. Next, details about the physical space and equipment used to support the system are described. Finally, the interfaces, roles, and protocols used by wizards and interviewers are presented. While the core mechanisms developed are intended to be usable in different domains, spaces, and equipment, we present the details of our usage to assist others in adapting these mechanisms to new areas of study.

A. Interaction Scope and Content

The system was designed to support verbal interactions in which an interviewer follows a generally planned structure, poses pre-planned questions to an interviewee, acknowledges their responses, and asks appropriate follow-up questions. More formally, the interaction content is structured as an ordered list of prompts (questions and statements), where each prompt may contain some dynamic content determined by the interviewee's prior responses. Each prompt also contains a list of related potential follow-up prompts.

Though the system supports responding to spontaneous requests, the quality of an interaction may degrade if the interaction does not return to the planned structure as soon as possible. The creation of prompts for an interaction is a critical factor in how well wizards will be able to direct an interaction in a natural and consistent manner. For each of our studies we used an iterative approach to prompt development in which prompts were trialed and refined first with researchers and then with members of the target population.

B. Physical Space and Equipment Configuration

The WITM approach is best deployed in a scenario where the wizards occupy a space isolated from the interaction space, as wizards benefit from being able to have verbal exchanges with each other. In our deployments, we used a three-room configuration that included: a room for participants and a researcher to interact and complete questionnaires or evaluations (research room), a room for participants and an interviewer to interact (interview room), and a control room for wizards (control room).

The top portion of Figure 2 shows the interview room used in our deployments of the WITM system. It is configured with two chairs facing each other, slightly angled outward, to accommodate the interviewer and participant. Beside each chair are tables with sensor enclosures, each containing an HD IP camera, a Microsoft Kinect V2, and a cardioid condenser microphone. Network-connected laptop computers for processing and storing Kinect data are located under each chair. A projection system used as part of the interviewer's interface is concealed behind the participant's chair.

The bottom portion of Figure 2 shows the control room space wizards use to guide interactions remotely. The center area contains two displays, the top display provides video feeds and the lower display shows software debug messages, sensor feeds, and notes about upcoming study sessions. Below the displays an audio mixer (BEHRINGER Eurorack

Fig. 2: The interview room (TOP) and wizard control room (BOTTOM).

UB1222FX-PRO) allows selection, adjustment, and distribution of the microphone feeds. A network video recorder (NVR) (Lorex LNR4082C4) records all IP camera feeds while a separate audio recorder (Tascam DR-40) records audio feeds. The video and audio feeds are also routed to an additional recorder (HDML-Cloner Box) which is used for backup and quick review. Each side of the center area contains a laptop computer for accessing the wizarding interface and headphones for monitoring audio feeds.

C. *Wizard Roles and Interfaces*

Interactions in our studies were supported by at least two wizards co-located in the control room. One wizard assumed the role of "director" and was primarily responsible for dispatching prompts to either the human or robot interviewer, while the other fulfilled the role of "producer" and focused on populating prompt variables with participant provided information in a timely manner. In conditions with a robot interviewer, the producer was also responsible for monitoring the robot's automated head tracking and engaging the robot's concurring head nod behavior when listening to a response. In rare circumstances wizards would perform actions outside of their assigned roles to better facilitate an interaction. For example, if the director anticipated the need for a custom response they could verbally ask the producer to prepare the response while continuing to focus on dispatching planned prompts. Wizards relied on verbal communication with each other to coordinate the handling of exceptional events.

While two wizards were sufficient for our interactions, an arbitrary number of additional wizards can be supported for more complex interactions or while training new wizards. In our deployments new wizards benefited from an experienced wizard monitoring and assisting until the new wizard could achieve the performance level required without assistance.

The primary wizarding interface, *Interaction Director*, is a web application that mirrors system state in real-time between all open instances of the interface via a centralized server. This approach allows additional wizards or observers to join the session using any computer in the control room. Figure 3 shows the wizard interface, which consists of the following areas:

(A) **Sticky Top Area** - elements in this area are visible in the interface at all times while other areas are scrollable.

(B) **Text-to-Speech Input** - allows input of arbitrary text for speaking. For robot interviewers "Say It" begins speech, while for human interviewers the text is displayed on their interfaces (tablet and/or the wall behind the participant).

(C) **Common and Useful Utterances Area** - the buttons in this area represent commonly used responses. In one study, this area also held a hobby look-up feature to retrieve domain specific prompts quickly that were related to hobbies.

(D) **Main Script Area** - a scrolling area contains study prompts in the order they are anticipated to occur with conditional or follow-up statements nested to be used as needed. Clicking on a prompt dispatches it to be spoken and updates the current location across interfaces to be this prompt. Hovering over a prompt containing shared variables adds a prominent yellow background to the associated variable input areas in the interface. Each prompt's background color indicates its state (White - top level, not yet dispatched; Green - conditional prompt, not yet dispatched; Blue - most recently dispatched; Yellow - previously dispatched; Cyan - hovering on prompt).

(E) **Shared Variables List** - labeled text input boxes for each variable used in the interaction script. Updates to these values propagate to all interfaces immediately upon a key-down event.

(F) **System Toggle Buttons** - quick access buttons for altering the state of the interaction system. The first group may be used multiple times per session and includes:

- **Prompt Hidden** - toggles the visibility of the projected prompter.
- **Go to Default** - moves robot interviewers' head to a predefined starting position.
- **Tracking** - toggles automatic head tracking for robot interviewers.
- **Nodding On** - toggles concurrence head movements for robot interviewers.

The second group contains configuration options that are typically used once per session:

- **Speech On** - mutes or un-mutes connected robots
- **Using RK** - selects the robot type in use
- **New Voice** - selects the synthetic voice in use
- **New Study** - resets all fields for a new interaction session

(G) **Live Video Overlay** - a system-wide overlay containing a live video feed of the participant from the interviewer's perspective is displayed.

During an interaction the director dispatches prompts

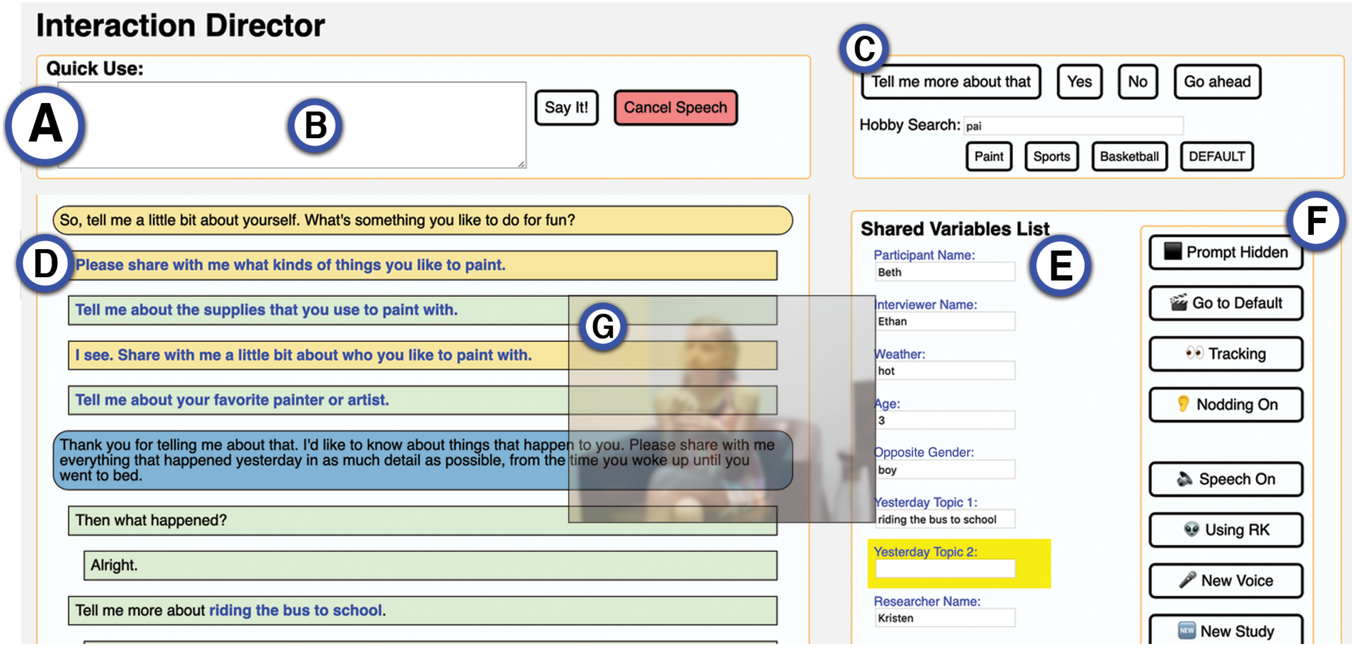

Fig. 3: The primary interface used by the wizards.

to be spoken by a robot or human by clicking on the prompt. Concurrently, the producer listens to the responses of participants and updates the shared variables that are used within future prompts. Updates to shared variables are reflected immediately within prompts and are denoted using a meaningful text color indicating the variable status (Red - missing value, Magenta - default value, Blue - producer input value).

Three shortcuts are implemented to help improve wizard response time:

- **Escape Key** - stops any ongoing speech event, hides prompter projection
- **Right Click Prompt** - puts prompt text into quick use box for quick editing
- **Tab Key** - puts focus on quick use box, if already focused on quick use box clears content

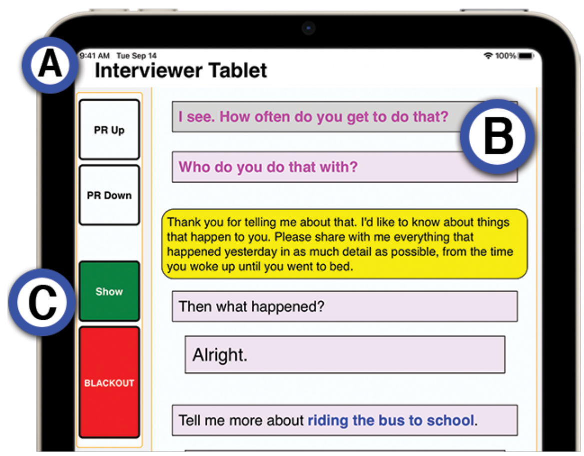

Fig. 4: The tablet interface available to human interviewers.

D. Interviewer / Performer Interfaces

While robot interviewers receive wizard input via network connection, human interviewers receive these inputs through either a tablet interface or projected text that appears behind the participant at eye-level of the participant.

Figure 4 shows the human interviewer's tablet interface, consisting of the following areas:

(A) Top Area - Persistent area displaying a title by default, but also used to display any non-script utterances dispatched.

(B) Main Script Area - a scrolling area containing study prompts in the order they are anticipated to occur with conditional or follow-up statements nested. This area automatically scrolls to keep current prompt in the center of the screen.

(C) Toggle Buttons - a persistent group of buttons to allow adjustments to the projected prompter concealed behind participants, includes:

- **PR Up** - raises the height of projected prompts slightly with each tap
- **PR Down** - lowers the height of the projected prompts slightly with each tap
- **Show** - allows projection of prompts
- **BLACKOUT** - disables projection of prompts

Similar to the wizard interface, background colors distinguish the state of each prompt. Interviewers in the three studies that deployed this system used either a 9.7 inch iPad Air or a 7.9 inch iPad Mini, which they held throughout the interaction.

In addition to the human interviewer's tablet interface, the current prompt is projected onto the wall behind the participant. Participants are unaware of this interface, as it is inactive when they enter and exit the room and does not produce any visible effects from their typical vantage point during the interview. The primary purpose of this interface is to allow human interviewers to maintain a forward facing gaze, as robot interviewers do, throughout the interview.

Fig. 5: A low-cost digital projector was combined with an angled mirror to allow the covert projection of text prompts onto the wall behind participants.

Figure 5 shows the hardware configuration for the projection system. A low-cost 3200 lumen projector (1280x800 native resolution) connected to a small single board computer was combined with an angled mirror to allow covert projections to occur on the wall behind participants near eye level of the participant. The projector was hidden from sight using an enclosure made of black foam board. A web interface was designed to provide the appropriate output for the projector given the system state. Prompts were displayed on a black background in white letters which were easily readable by interviewers. The displayed text was also constrained to be behind the participant's chair and not extend too far horizontally. To account for participant height and posture, human interviewers were able to use controls on the tablet to make fine adjustments to the projected text's height on the wall.

In our studies, human interviewers were encouraged to use the interface that facilitated the best interaction with the participant. In general, the projected text was preferred by interviewers as it reduced their need to look down at the tablet interface. The interviewer as well as the wizards were able to disable the projection interface during an interaction to prevent participants from becoming aware of it. This occurred infrequently overall and was mostly used with younger participants, as they were more likely to alter their vantage points mid-interview. Across all sessions one participant communicated awareness of the projected interface.

While human interviewers became practiced in delivering most prompts and often only used their interface as a cue for which prompt to deliver, prompts that incorporated participant-provided information were novel to the interviewer each time they were delivered. As a result the wizard acting as director would adjust their performance so that the next prompt would be displayed as soon as the interviewer delivered the previous prompt. This allowed the interviewer to see their next statement while the participant was responding to the previous prompt.

E. Data capture and processing

All wizard interfaces, interviewer interfaces, sensors, and robots were connected through a server application written in JavaScript executed in the Node.js run-time environment. The application provided interfaces as web pages over the HTTP protocol and facilitated data transfer between clients and the server via Web Sockets. All data was transferred through the server application and timestamped at millisecond resolution for later analysis [7].

IV. USAGE EXAMPLE

The workflow facilitated by the system may be best appreciated from the vantage point of the control room in a typical session. Table I provides a representative sample of events that occurred during a study session that involved conducting a forensic interview about a child's experiences with bullying. Though these example events include some items relevant to only robot interviewers (e.g., checking pronunciation of names by a text-to-speech engine) the general flow of events is similar regardless of the agent acting as the interviewer.

In addition to the events described in Table I, there are other circumstances that may prompt wizards to confer with each other while making decisions during an interaction. For example, the wizard acting as the director adheres to the study's protocol to decide which follow-up prompts are merited given a participant's response, but may verify their choice with the wizard acting as the producer if uncertain. Similarly, if a wizard perceives a participant may be showing signs of distress while recalling a past experience, other wizards and researchers can be alerted and assist in ensuring the well-being of the participant. This may entail the director changing the order of prompts or the producer being cognizant of sensitive areas while choosing content to be used later in the interaction.

V. DEPLOYMENTS

The system was initially deployed for use in a study comparing the use of humans and robots conducting a forensic interview about bullying experiences with 86 children ages 8 to 12 and in a slightly modified version of the same study conducted with 56 children ages 12 to 17. The system was also deployed to facilitate an eye-witness memory study

Pre-session	**Producer** updates the shared variables for names of interviewer, researcher, and participant (and verifies pronunciation).
00 min	Participant arrives, completes the assent process, **Director** begins system recording.
06 min	Researcher begins to lead the participant to the interview room. Researcher inadvertently forgets to use the participant's name in the introduction to the interviewer. **Producer** quickly deletes the name, allowing the first prompt to fall back to a generic greeting. **Director** dispatches the first prompt.
07 min	Researcher leaves the interview room and enters the control room to monitor the session **Director** proceeds through the first few prompts and follow-up prompts **Producer** captures information provided by the participant as the interview continues and supervises the robot's head tracking and nodding.
14 min	The participant slowly describes an experience with relational aggression After extended silence, the **Director** dispatches the next prompt, but the participant continues speaking **Director** immediately uses the escape key to stop the speech, and clicks the quick response utterance, "Go ahead". Participant finishes explaining their relational aggression experience.
25 min	As the session reaches the final steps, the **Director** cues the researcher to prepare to return to the interview room **Director** dispatches the final prompt, asking the participant about their summer plans.
27 min	Once the participant has nearly completed a response, the researcher knocks on the door to the interview room, enters, and asks if the interview is about finished. **Director** dispatches a "yes" response. The researcher asks the participant to follow them back to the research room.

TABLE I: A representative sample of events and wizard actions during a forensic interview study session.

conducted with 75 children ages 8 to 12. In total 125 interactions took place with robot interviewers (across three different robot platforms) and 92 interactions were conducted with human interviewers (using three different human interviewers). A total of six individuals acted as wizards over the course of the three studies. All studies were approved by and conducted with guidance from the Mississippi State University Institutional Review Board.

This section briefly summarizes each study and discusses results and insights relevant to the performance of the wizarding system and the WITM paradigm.

A. Forensic Interview Studies

The system was first used in studies that conducted forensic interviews with children about bullying experiences (see [25] for comprehensive design and results). The interview interactions in these studies were typically between 15 - 30 minutes. The first study with children ages 8 to 12 compared human and robot interviewers of both genders for a total of four different conditions. The second study with children ages 12 to 17 compared a human male interviewer, male RoboKind robot, and humanoid Nao robot.

A semi-structured verbal interview was used to understand each participant's perceptions of the human or robot that performed their forensic interview regarding bullying. While the results indicated overall favorable perceptions of both human and robot interviewers, some significant differences were found between interviewer types. On average participants were more likely to be uncertain if a robot interviewer understood how they felt or could provide advice in comparison to a human interviewer, but attributed more socially protective factors to robots than humans. No differences were found in areas of likability or helpfulness of interviewers (>90% positive responses for all) [25].

A review of session logs and interaction recordings demonstrated that the WITM-based system was able to effectively support both human and robot interviewers in adhering to the forensic interview protocol. All interviewers and wizards were able to use the system reliably with only a single session excluded for a technical failure involving a robot. While we have not yet conducted studies which compare using WITM to non-WITM while holding other aspects constant, our initial attempts at a non-WITM approach to forensic interviewing were riddled with variability between the two types of interviewers. For example, in sessions without WITM, human interviewers had significantly higher word counts during rapport building compared to robots, while in studies using WITM this difference was eliminated.

B. Eyewitness Memory Study

The third study to make use of the system involved participants watching a slide show, being interviewed about what they remembered from the slide show (but were given some misinformation during the interview phase), and then completed a final memory assessment about the slide show. In this application prompts about bullying experiences and definitions were replaced with prompts asking the participant to provide a narrative of what they recalled about the slide show. Follow-up questions focused on encouraging elaboration of provided details. The interviewer concluded by asking specific questions about the slide show, with some questions introducing misinformation about the slide show. On average, interactions in this study lasted approximately 10 minutes. The study compared a human male interviewer and two versions of the humanoid Nao robot with varying levels of social awareness.

While data from this study is currently being analyzed, interactions between participants and human and robot interviewers occurred without issue and a preliminary review of qualitative feedback indicates both were generally rated positively. The system continued to perform reliably and both robot and human interviewers were able to effectively engage with participants while adhering to a standardized protocol.

This study also involved expanding the Interaction Director interface to include a lookup widget to automatically provide relevant prompts for hobbies or interests mentioned by participants in the rapport building portion of the interview. Using data from the previous studies, a small database of common hobbies and relevant follow-up prompts was created and was able to be queried by the producer to quickly select an appropriate set of prompts. This feature was successfully adopted by wizards and human interviewers with the majority of sessions resulting in successful timely retrieval and dispatch of appropriate prompts to the interviewer. Although a small example, this illustrates how the WITM approach extends to allow both human and non-human agents to seamlessly incorporate knowledge and direction from human and non-human sources.

VI. Insights and Future Development

Through a review of session log notes and feedback from wizards and interviewers, this section describes the most salient observations that occurred from the implementation and use of the system and discusses future development paths. Among the lessons learned throughout the course of this work, the standouts include:

- **Thoughtful content design reduces delays:** Delays in responding are noticeable artifacts of the wizarding process, but well-planned content design helps to eliminate this by having relevant response options available for selection immediately. A dedicated wizard acting as a producer can assist in creating content options dynamically for a directing wizard to select from, but the producer too benefits from templates that are relevant and require the least customization effort. The benefit of well-crafted base content has also been noted in design studies like [26] while using WOz techniques.
- **Sensible default variable values reduce problems:** While including participant-dependent variables in prompts assists in building rapport with participants and produces a more engaging interaction, the ability to fallback to a generic, but appropriate prompt is a necessary safety mechanism. For example, though wizards were able to learn and coerce the synthetic speech engine to correctly pronounce a participant's preferred name prior to its use, if the researcher failed to introduce the participant by name to the interviewer, use of the name would be confusing and perhaps unsettling.
- **Video overlays increase wizard awareness:** In our initial prototypes of the wizard interface the live video feed from the perspective of the interviewer was positioned adjacent to the main interface. Replacing this with a resizable floating system-wide overlay with opacity adjustment allowed wizards to remain engaged with the remote interaction without requiring as large of an attention shift away from their interface. This is similar to systems like [21], which overlays interface elements onto a live video feed.
- **Verbal communication facilitates multi-wizard collaboration:** Though our wizards had well defined roles and synchronized visual interfaces, the use of verbal communication was common to coordinate responses during an interaction. The ability to verbally communicate allowed wizards to remain focused on the interface while communicating with each other. This finding has been discussed in other multi-wizard research, such as [5].
- **Humans and robots have distinct needs:** The WITM approach aims to provide the same wizarding services to all types of agent, but wizards may need to adopt different behaviors for different types of agents. For example, since human interviewers needed to consult the projected interface for their responses, the wizard acting as a director worked to dispatch the next prompt as soon as possible (often during the participant's response) to maximize the amount of time the human interviewer had to observe their next response.

Though our WITM-based system was effective for improving consistency while comparing human and robot interviewers, there are several improvements that may strengthen the system, including:

- **Expanded prompt generation methods:** The prompts used in the forensic interview studies were customized using variables that were populated by the producer's input. Variable fields have been used in several studies involving the WOz technique [16], [17]. In the eyewitness memory study a widget for selecting a set of prompts related to a search term (participant's hobby) was developed and was able to provide a set of tailored follow-up prompts very quickly. The incorporation of other knowledge bases and methods of predicting text or interaction flow could greatly increase a wizard's capacity to respond quickly [3], [4].
- **Improved shared content syncing:** In our deployments it was common for producers to iteratively refine their inputs for variables as participants provided more information, however this could lead to a state of the producer beginning a refinement while the director simultaneously dispatches a prompt using the variable. This was remedied by communication between the two wizards, but this added cognitive load could be relieved by including a mechanism for the producer to commit completed versions of the variable rather than propagating each character edit.
- **Projection interface improvements:** Though basic, the projected prompt mechanism was an effective interface for human interviewers to receive direction from the wizards. A more dynamic version could make use of sensors to automatically keep prompt text in an optimal

location for the interviewer.

Finally, the WITM paradigm itself would benefit from further expansion and exploration of its limitations. Examples of intriguing future use of the paradigm include:

Additional modalities: While our use of WITM has focused on verbal interactions, exploration into other modalities is an area in need of exploration. For example, we posit that WITM could be implemented and deployed with haptic output cues or through augmented reality interfaces to facilitate non-verbal spatial tasks.

Fully concealing wizards: Human interviewers in our studies were aware that content was generated by human wizards, but WITM is able to facilitate interactions where the nature of the wizard is unknown.

The collaborative multi-wizard approach combined with interfaces allowing human agents to gain covert direction from behind the curtain presented in this article proved to be a helpful tool for facilitating the comparison of humans and robots as interviewers. While our evaluations were in the context of guided verbal interactions, we hope that researchers and designers will benefit from adopting and expanding upon the "Wizards in the Middle" paradigm across a wide variety of application areas.

ACKNOWLEDGMENT

In addition to the authors of this article, the research team includes co-PIs: Deborah K. Eakin, David C. May, and Melinda Pilkinton; and student researchers: Allison Bruning, Zachary Buchanan, Ethan Flurry, Alexis Jones, John Kelly, Aaron Kimbrell, Hannah Krapac, and Kristen Stives. We also thank the anonymous reviewers for their helpful feedback.

REFERENCES

[1] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Transactions on Information Systems (TOIS)*, vol. 2, no. 1, pp. 26–41, 1984.

[2] A. Z. Buchanan, "Reducing social demands in child eyewitness memory using robot interviewers," May 2017. [Online]. Available: https://ir.library.msstate.edu/handle/11668/14455

[3] S. Janarthanam, R. Hill, A. Dickinson, and M. Fredriksson, "Click or type: An analysis of wizard's interaction for future wizard interface design," in *Proceedings of the EACL 2014 Workshop on Dialogue in Motion*, 2014, pp. 19–27.

[4] S. Gandhe and D. Traum, "Sawdust: a semi-automated wizard dialogue utterance selection tool for domain-independent large-domain dialogue," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2014, pp. 251–253.

[5] M. Marge, C. Bonial, K. A. Pollard, R. Artstein, B. Byrne, S. G. Hill, C. Voss, and D. Traum, "Assessing agreement in human-robot dialogue strategies: A tale of two wizards," in *International Conference on Intelligent Virtual Agents*. Springer, 2016, pp. 484–488.

[6] D. Salber and J. Coutaz, "Applying the wizard of oz technique to the study of multimodal systems," in *International Conference on Human-Computer Interaction*. Springer, 1993, pp. 219–230.

[7] N. Martelaro, "Wizard-of-oz interfaces as a step towards autonomous hri," in *2016 AAAI spring symposium series*, 2016.

[8] G. Hoffman, "Openwoz: A runtime-configurable wizard-of-oz framework for human-robot interaction," in *2016 AAAI Spring Symposium Series*, 2016.

[9] S. R. Klemmer, A. K. Sinha, J. Chen, J. A. Landay, N. Aboobaker, and A. Wang, "Suede: a wizard of oz prototyping tool for speech user interfaces," in *Proceedings of the 13th annual ACM symposium on User interface software and technology*. ACM, 2000, pp. 1–10.

[10] S. Schlögl, G. Doherty, N. Karamanis, and S. Luz, "Webwoz: a wizard of oz prototyping framework," in *Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems*. ACM, 2010, pp. 109–114.

[11] D. DeVault, J. Mell, and J. Gratch, "Toward natural turn-taking in a virtual human negotiation agent," in *2015 AAAI Spring Symposium Series*, 2015.

[12] J.-w. Ahn, P. Watson, M. Chang, S. Sundararajan, T. Ma, N. Mukhi, and S. Prabhu, "Wizard's apprentice: Cognitive suggestion support for wizard-of-oz question answering," in *International Conference on Artificial Intelligence in Education*. Springer, 2017, pp. 630–635.

[13] F. Rietz, A. Sutherland, S. Bensch, S. Wermter, and T. Hellström, "Woz4u: An open-source wizard-of-oz interface for easy, efficient and robust hri experiments," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.668057

[14] M. Porcheron, J. E. Fischer, and S. Reeves, "Pulling back the curtain on the wizards of oz," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW3, jan 2021. [Online]. Available: https://doi.org/10.1145/3432942

[15] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012.

[16] C. Bonial, M. Marge, A. Foots, F. Gervits, C. J. Hayes, C. Henry, S. G. Hill, A. Leuski, S. M. Lukin, P. Moolchandani, *et al.*, "Laying down the yellow brick road: Development of a wizard-of-oz interface for collecting human-robot dialogue," *arXiv preprint arXiv:1710.06406*, 2017.

[17] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, J. H. Ruckert, and H. E. Gary, "Human creativity can be facilitated through interacting with a social robot," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 173–180.

[18] A. Shin, J. Oh, and J. Lee, "Apprentice of oz: Human in the loop system for conversational robot wizard of oz," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 516–517.

[19] M. Nebeling and M. Speicher, "The trouble with augmented reality/virtual reality authoring tools," in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2018, pp. 333–337.

[20] A. Green, H. Huttenrauch, and K. S. Eklundh, "Applying the wizard-of-oz framework to cooperative service discovery and configuration," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, 2004, pp. 575–580.

[21] A. Pereira, E. J. Carter, I. Leite, J. Mars, and J. F. Lehman, "Augmented reality dialog interface for multimodal teleoperation," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 764–771.

[22] A. P. Chaves and M. A. Gerosa, "Single or multiple conversational agents?: An interactional coherence comparison," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 191.

[23] M. E. Lamb, Y. Orbach, I. Hershkowitz, P. W. Esplin, and D. Horowitz, "A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the nichd investigative interview protocol," *Child abuse & neglect*, vol. 31, no. 11-12, pp. 1201–1231, 2007.

[24] I. Hershkowitz, M. E. Lamb, C. Katz, and L. C. Malloy, "Does enhanced rapport-building alter the dynamics of investigative interviews with suspected victims of intra-familial abuse?" *Journal of Police and Criminal Psychology*, vol. 30, no. 1, pp. 6–14, 2015.

[25] Z. Henkel, K. Baugus, C. L. Bethel, and D. C. May, "User expectations of privacy in robot assisted therapy," *Paladyn, Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 140–159, 2019.

[26] P. Worthy, M. Boden, A. Karimi, J. Weigel, B. Matthews, K. Hensby, S. Heath, P. Pounds, J. Taufatofua, M. Smith, *et al.*, "Children's expectations and strategies in interacting with a wizard of oz robot," in *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. ACM, 2015, pp. 608–612.