Pulling Back the Curtain on the Wizards of Oz

MARTIN PORCHERON, Swansea University, UK
JOEL E. FISCHER, University of Nottingham, UK
STUART REEVES, University of Nottingham, UK

The Wizard of Oz method is an increasingly common practice in HCI and CSCW studies as part of iterative design processes for interactive systems. Instead of designing a fully-fledged system, the 'technical work' of key system components is completed by human operators yet presented to study participants as if computed by a machine. However, little is known about how Wizard of Oz studies are interactionally and collaboratively achieved in situ by researchers and participants. By adopting an ethnomethodological perspective, we analyse our use of the method in studies with a voice-controlled vacuum robot and two researchers present. We present data that reveals how such studies are organised and presented to participants and unpack the coordinated orchestration work that unfolds 'behind the scenes' to complete the study. We examine how the researchers attend to participant requests and technical breakdowns, and discuss the performative, collaborative, and methodological nature of their work. We conclude by offering insights from our application of the approach to others in the HCI and CSCW communities for using the method.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **HCI design and evaluation methods**; *User studies*; Natural language interfaces.

Additional Key Words and Phrases: woz, natural language interfaces, voice interfaces, vuis, robots, collaboration, coordination, research practice, methodology, ethnography, ethnomethodology, cscw

ACM Reference Format:
Martin Porcheron, Joel E. Fischer, and Stuart Reeves. 2020. Pulling Back the Curtain on the Wizards of Oz. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 243 (December 2020), 22 pages. https://doi.org/10.1145/3432942

1 INTRODUCTION

The Wizard of Oz (WOz) method, as a widespread technology research practice, uses humans to drive components of complex digital or physical systems, particularly those with significant automation or elements of artificial intelligence. In other words, core computational logic or mechanisms are absent and are, in fact, performed by a human 'pulling' metaphorical levers of some sort or another. While such 'hidden work' is very familiar to CSCW in the form of crowdwork systems, the hidden work in WOz is more proactively 'deceptive' in that it is used to facilitate experimental studies where participants must believe they are just interacting with a machine. The payoff is that WOz enables researchers to get at a 'way of interacting' that does not involve them committing to building a system before they know how to build it. As such, WOz studies are clearly powerful and have seen significant uptake across a variety of research communities. Curiously, though, we find little inquiry and analysis of *how* WOz as a method is actually performed, and without examining WOz directly, we cannot critically reflect on where it might need improvement.

Authors' addresses: Martin Porcheron, m.a.w.porcheron@swansea.ac.uk, Computational Foundry, Swansea University, Bay Campus, Fabian Way, Crymlyn Burrows, Swansea, Wales, SA1 8EN, UK; Joel E. Fischer, joel.fischer@nottingham.ac.uk, School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK; Stuart Reeves, stuart.reeves@nottingham.ac.uk, School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK.

Existing literature has described the various forms in which WOz methods, often called 'experiments', can or have taken, in terms of which components are simulated and how (e.g. [10, 37]), and has detailed how such studies should be reported to demonstrate validity (e.g. [34]). But, we think there is a lack of focus by researchers on trying to understand how WOz is brought off as a method. We want to ask: what *is* the organisational and interactional work done to implement and perform such a study? CSCW has a tradition of critically examining the social organisation work in a range of collaborative settings such as in London Underground control rooms [19], in usability testing [32], and even disaster response [7], and it is in this vein that this paper proceeds. The goal of this work is to provide an analytic description of how the method is performed in situ, to contribute to an understanding and reflection upon existing research practice, to support the development of improved software tools and practices, and to provoke methodological advancement.

Firstly, we will reflect upon the traditions of the method before turning to examine how it is enacted. We draw upon the philosophical orientation of ethnomethodology [11] and present four fragments of data, consisting of transcripts of action, recorded from WOz studies in which multiple researchers employed the method together. The first fragment sets the scene and introduces the purposes for which the method was used, examining how the study was orchestrated and presented to participants. The second fragment further unpacks the study, explicating the division of labour and coordination between the researchers and the participant, how simulated responses were constructed, and the work of handling contingencies. The third fragment delves deeper into examining the intricate and time-sensitive nature of attending to contingencies, and the strict adherence to a study protocol through which the outcomes of each study are shaped by participant interaction. The final fragment introduces a technical breakdown which leads to coordinated resolution work by the researchers. Through this, we demonstrate the collaborative, rehearsed-yet-improvised, and methodical ways in which WOz studies are constructed and operationalised. By analysing our practice as researchers, we then establish a series of practical implications for researchers in the HCI and CSCW communities to design and conduct WOz studies.

2 RELATED WORK

The method's name stems from the novel *The Wonderful Wizard of Oz* [2], in which the main characters undertake a journey to meet a wizard. The plot twist is the revelation that the Wizard is not as wonderful as promised, but rather consists of a human behind a curtain controlling a machine. In other words, the Wizard is an orchestrated illusion and it is this twist that inspires the name. The earliest examples of the approach took the simpler and unequivocal name of "experimenter in the loop" [17, pp. 1–2] or the epithet of "The Perfect System" [24], with the moniker first appearing in 1982 [8]. These labels exhibit the variety of ways in which the approach is presented in literature for much the same purpose. With this paper, we seek to deepen the representation of the method to examine the specific ways in which such studies are framed and conducted.

2.1 Wizard of Oz in simulations of technology

The method has been used to understand people's interactions with technology which may not (quite) yet exist. A WOz approach offers the ability to prototype and potentially validate—or not—design concepts through experimentation without the costly development time that a full system may require [47]. The method is often used in situations where the ideas involved include the use of *new* technologies or technologies that require significant resources to implement, hence its adoption in studies of interaction with robotics and voice-controlled interfaces.

The earliest use of the method in HCI is Gould et al.'s voice-controlled 'Listening Typewriter' [16], although other studies around the time adopted varying input techniques such as pushing buttons on a touch-tone telephone [23, as cited in [10]]. Dahlbäck et al.'s text-based travel booking system [4],

or Wooffitt's voice-based public service information telephone line [48] are two such early examples. In simulating the technical implementation, which—as remarked upon by Wooffitt [48, pp. 23–24]—must be convincing, researchers have adopted approaches such as adjusting the delivery of the verbal response using vocoders [18, p. 498], or as in more recent studies, using text-to-speech libraries to deliver a synthesised voice (e.g. [26]), "preserving the believability of the simulation" [47, p. 130]. Technological advancement now also allows for the simulation of different technologies that are increasingly conceivable while the technologies previously simulated—such as an automated telephone system—are now commonplace. The method has stayed relevant by serving as a 'sliding scale' by following technological development on what could be called the *cusp* of ready-availability of a technology. For example, Sirkin et al. focuses on eliciting conversation in robot-controlled vehicles [40], and in our case, we simulated voice-controlled collaborative robots in a non-domestic setting, bringing together current research on industrial co-bots and voice interaction [21].

2.2 Wizard of Oz in HCI and CSCW

Erdmann and Neal's work in the 1970s on Airline Ticket Vendor machines demonstrates HCI's long-standing adoption of the approach in design work [5]. More recently, it has particularly found a niche in simulating Machine Learning systems, given the high upfront development and data collection costs involved [3]. It is especially used in the design of natural language [10, 26] and multimodal [39] interfaces, with Wizards supplanting the role of digital systems for myriad reasons [29]. Other recent applications include mobile gaming [6], emotional robotics [33], and playful physiotherapy [25]. Schlögl et al. conducted a literature review and identified 16 'variations' of simulation, in which either (or both) machine input (e.g. automatic speech recognition), automated processing (e.g. machine translation, natural language understanding, dialogue management) and generated output (e.g. text-to-speech) is simulated [37], although here we focus on one particular form. We also note that while applications of the approach often strive to implement an 'optimal system' (e.g. [16, 24]), the method has also been applied in other cases, for example, to compare user performance with variations on different interface designs [22].

There is a growing pool of literature contributing both software [36] and hardware [9] tools to run studies, based on the broadly accepted nature that running such studies is taxing [9], and can benefit from multiple workers to operationalise [47]. While Fraser and Gilbert's descriptive paper on running WOz studies [10] has been highly influential to the development of the approach, there is a notable omission in the description of the orchestration work that unfolds to conduct such studies. Dahlbäck et al. aligns WOz with addressing the need for "high quality empirical data" [4, p. 199] to inform the design of intelligent systems. Therefore to meet this goal, there needs to be a "great deal of care and consideration [...] in the design of such experiments." [4, p. 199]. We examine this *research practice* to reveal how the approach is methodically applied in settings where researchers collaboratively orchestrate a WOz study. As new technologies such as natural language processing become a core design material within interaction design, pressures to avoid premature costly data collection and technical development are likely to exacerbate interest in WOz, as can be seen from a string of recent publications in HCI-based research (e.g. [3, 28, 33, 45, 46, 49]). This work often describes the use of WOz somewhat cursorily, glossing over its application. In contrast, we intend to unpack this gloss and explicate the interactional and organisational work to apply the method. We believe that doing so can uncover important and mostly overlooked aspects of WOz that are methodologically significant.

3 DATA COLLECTION AND ANALYSIS

We combine contributions from existing literature with our insights developed through our data collection to provide the basis for our work.

3.1 Our use of the method

Although our paper does *not* discuss the findings of our forthcoming WOz study, we do need to provide some contextualisation for the reader. The WOz study's purpose was to understand how people who work in a laboratory would issue spoken commands to a voice-controlled robot vacuum cleaner in a 'natural' way. Participants were asked to instruct the vacuum robot to perform certain cleaning and scheduling activities according to five scenarios. We will further unpack this practice through our fragments. The research that is used as the basis for developing this paper's findings received ethical approval from the University of Nottingham's School of Computer Science Research Ethics Committee.

This study required at least two researchers to be present to run the study believably [47]. A participant would stand in the middle of the room (pictured in Figure 1) with a researcher, who explained the study and each scenario that the participant was tasked with completing. Participants would say a command *to* the vacuum robot—which they were told was voice-controlled. Another researcher, performing the role of *the Wizard*, was sitting in the room at the same time, using a laptop and acting as an 'observer' or 'supervisor' of the study. Participants were not told that this person was controlling the NEATO robot during the study. In total, 21 participants took part.

3.2 Studying Wizard of Oz as a practice

We employed an ethnographic approach to examining our use of the WOz method, making use of video data collected during the running of the studies in addition to notes, system logs, and our experiences from running the studies. Two cameras were used throughout the studies to capture the researchers and participants. We also captured the screen of the Wizard's computer used in the studies. Throughout the remainder of this paper we unpack our work in conducting the WOz studies, including the experimental design, systems design, study preparation, and on-the-day coordination. We adopted an ethnomethodological perspective [11] informed by conversation analysis [35] (often abbreviated as EMCA) in our analysis.

Fig. 1. A video still showing the laboratory where the elicitation study took place. The Wizard is shown in the top left of the room (A) and in the cut-out image from the opposite angle. P11 (B) is standing with the researcher (C) looking towards the NEATO robots (D) and the three piles of debris (red ovals for clarity).

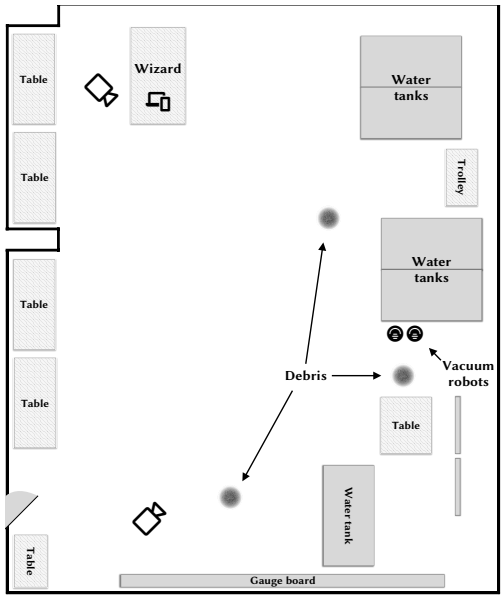


Fig. 2. Schematic of the front half of the lab showing the key features labelled. The orientation is roughly the same as the photo in Figure 1 (i.e. the Wizard is seated in the top left).

The next four sections of this paper each focus on a fragment of data, presented as transcripts of verbal and non-verbal action[1]. These fragments progressively unpack the complexity of how the method was applied as research practice. We will then discuss and use these findings to assemble our implications for researchers in HCI. We start with a fragment which examines how the study is prepared, framed, and presented to participants in much the same way that 'any' lab-based HCI study might be; the second fragment is used to conduct a closer examination of the division of labour between the two researchers and the coordination of their 'backstage' and 'front stage' actions according to a shared protocol (and its practical accomplishment); the third fragment delves into the nitty-gritty of the work in attending to the contingencies of running the study, particularly when, inevitably, elements of the study go awry and remedies must be found to return to the protocol; and the final fragment examines a technical breakdown which entails the two researchers collaborating with the existing tools to repair the problem without revealing the fiction.

4 FRAMING A WIZARD OF OZ STUDY

To begin, we want to introduce the setting by showing how we as researchers brought participants in to do their first task with the robot; as a part of this we will look at a researcher setting the scene for them and then getting the participant to try out their first verbal instruction to the robot. We will also see how the robot responds to instruction. Initially, we want to examine this from the perspective of participants, i.e. the people who approach and treat the study as they might any other lab-based technology trial. Later we will start to tease out where and how the Wizard comes into play.

[1]We minimally draw upon the Jeffersonian notation [1], denoting where there is a short pause (`.`) or pause of a specific length in seconds (`1.2`), utterances that are `cut off-`, two or more latched utterances that immediately `follow on= =from each other`, talk that is `°quiet°` or `LOUD`, where sounds are `elong:::ated` for a specific length of time in tenths`::` of a second, and where there are non-verbal `((sounds))`. Overlapping action is denoted with square brackets (⌜, ⌞, ⌝, and ⌟).

Now, perhaps like 'any' HCI lab-based study, we considered how our experiment was to be framed and presented to participants to engender the 'right kind' of participation. Constructing and presenting the simulated technology to participants necessitates a believable *fiction* [47]. This stems from the idea that participants should believe they are using a computer system with various kinds of functionalities, which are themselves 'artifice' in some way [10]. With this in mind, we introduce Fragment 1 which demonstrates this framing in action. The primary researcher (RES) stands with the participant (P11) a few metres from the two vacuum robot cleaners in the middle of the food chemistry lab. Only one VACuum robot is used in each study, with the other acting as a 'hot spare' should there be a technical issue with one of the robots.

```
1  RES the- the idea is you can instruct the
2       vacuum cleaner
3      there is no er: correct or wrong way to do it
4      the idea is to understand how people are
5       going to instruct the vacuum cleaner
6      so we are going through er five different
7       scenarios to ┌     in┐struct that
8  P11               └oh kay:┘
9  RES so the first one its s- the simple clean
10      scenario an::: in this lab there are three
11      areas that need to be cleaned . and you
12      have to choose an area to clean and instruct
13      the robovac  . to cl┌ean near the du┐st
14 P11                     └yea::s:       ┘
15 RES you have to call them or refer to them as
16      robovac
   ⋮
22 RES so if you can do it
23     (1.4)
24 P11 er::m:: (3.8) <robovac> . move forward and
25      cle↑an?
   ⋮
34 VAC ok . i will clean forward now
35     (1.8)
36 VAC ((beeps and moves forward out of dock))
37     (3.0)
38     ┌((vacuum motor spins up))                 ┐
39 RES └((lifts script up to read next scenario))┘
```

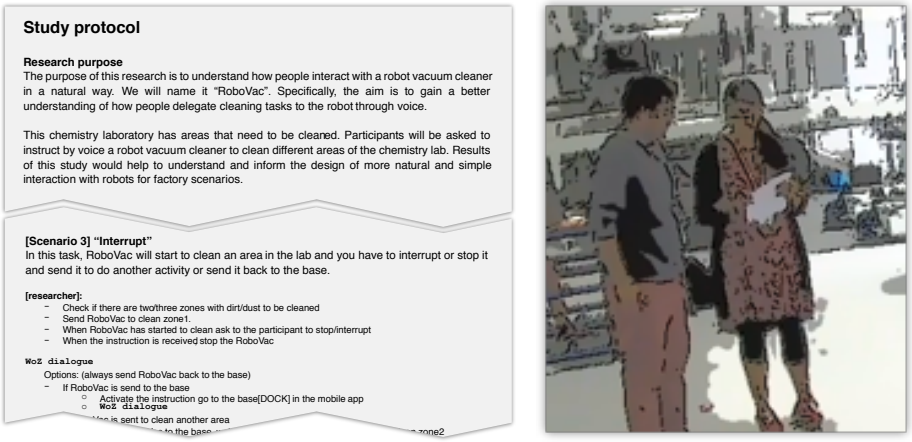
Frag. 1. Participant 11's (P11) study commences

We are interested here in how the primary researcher goes about setting up the framework of participation [15] for the participant to go on with. In many ways this is conducted in a fashion similar to any lab-based study, with the primary researcher (on lns. 1–16) variously: describing the purpose of the study (lns. 1–2, 4–5), reassuring the participant and setting out that instructions can be given in a 'natural' way (ln. 3), framing this first scenario as one of many (lns. 6–7) and then providing instruction about what actions the participant must take for this particular one ("`you have to choose an area to clean and instruct the robovac`", lns. 9–13).

4.1 Resources to frame the study

Despite the fact that this opening may appear casual, there are a host of *methodological obligations* that inform the researchers' conduct that come to the fore. The primary researcher's account to the participant, and the ways in which this must be produced, is framed by the following of

the study protocol. The protocol provides the canonical description of the study with which the primary researcher can consistently explain the narrative to facilitate and scaffold the participants' completion of the scenarios (cf. the protocol in Figure 3a and lns. 1–8 in Fragment 1). Furthermore, this protocol serves as a coordinating tool between the primary researcher's scaffolding of the study and the participant's completion of the scenarios, and the Wizard's 'invisible' work[2]. The Wizard must be able to follow participant-researcher interactions so as to respond to participant instruction to the vacuum robot at the relevant moment, and the primary researcher must maintain awareness of and be sensitive to the contours of this 'hidden' joint project between them.



(a) A snippet of the protocol document showing the framing and scenario 3

(b) The researcher reads from a paper copy of the protocol

Fig. 3. The protocol used in the studies

The protocol, which provides a high-level structure for the study, was used to generate an outline script that guides how input to the robot 'should' be done and responded to by the vacuum robot. Of course, the nature of this being a study to understand how a person instructs the vacuum robot, this script is incomplete and is naturally not shared with participants. It provides the 'sorts' of requests a participant is expected to make in each scenario and the ones which should generally 'work'. The purpose of developing an outline script is not to enumerate the possible input and output combinations—such an elaboration would be impossible and an insurmountable task to navigate during a study. However, there is a 'good faith' notion that the participant will work to complete the scenario explained to them (although there is no such guarantee). The loosely specified nature of the scenarios in combination with this faith enables limited pre-study decision-making on the validity of input by the researchers and the preparation of responses to various sorts of requests. For example, it was decided that requests which instruct the robot to "`clean`" but without a specification of a location to clean should be responded to with "`where should I clean?`"[3].

While these two resources were prepared before the studies commenced, the contingent nature of participants' requests inhibits any comprehensive a priori preparation. The participant can—and may—make any request and it is expected that the robot will respond as per the construction of the study. For example, consider lines 24–34 from Fragment 1:

[2] 'Invisible' in the sense that the participant does not 'see' it—we will expand upon the Wizard's work in the next section.
[3] This was guided by the methodological interests of eliciting as much vocabulary as possible.

```
24 P11 er::m:: (3.8) <robovac> . move forward and
25      cle↑an?
   ⋮
34 VAC ok . i will clean forward now
35     (1.8)
36 VAC ((beeps and moves forward out of dock))
```

A request like this that instructed the robot to "`move`" or "`clean`" with a positional parameter referencing the docked robot (e.g. "`forward`", ln. 24) is treated as a 'valid' instruction in which the robot proceeds to clean in front of it. Of note here is the primary researcher's frame of there being no "`correct or wrong way`" (ln. 3) to instruct the robot. This precludes a static response lookup and instead engenders a continual 'invisible' assessment of participant requests. In this case, the term "*move*" was not considered to be a valid action prior to the studies. As such, lns. 24–25 *could* have been treated as a problematic request were there a rigid application of this script. However, "*move*" was added in response to an earlier participant using the term in their study. At the time, an immediate decision was taken to allow this sort of request and thus the subsequent requests by that and other participants that use the term became valid. To ensure consistency across studies, a 'scratchpad' was kept of these sorts of ad hoc decisions—in each study this was done on sticky notes, which were transferred to a digital document afterwards. The protocol and progressively developed script are emblematic of others' discussions of developing competency through pilots [37]. We note here how this practice can carry on beyond pilot studies and throughout participants' trials.

4.2 The robot's response

Our second point is of the response from the vacuum robot ("`ok . i will clean forward now`", lns. 34–36). Generally speaking, the decision of whether a request was to be responded to in a 'positive' or 'negative' manner[4] was determined by fitting the request against the outline script of pre-determined outcomes. In the case where no prior practice exists, there was an ad hoc examination of both the words and context of the participant's instruction among the sequence of interaction to determine the appropriate action of the robot. `P11`'s request (lns. 24–25) fits the criterion discussed above in relation to the script for the sort of input that is valid based on previous instances of this sort of request, and thus should be responded to 'positively'. `P11` is not provided with a specific formulation for requests, with the researcher providing just a name "`call them . . . robovac`" (lns. 15–16) and the participant left to select the rest of the instruction, as seen in Fragment 1:

```
24 P11 er::m:: (3.8) <robovac> . move forward and
25      cle↑an?
```

The use of a wake word [30]—in this case `robovac`—proffers a veneer of this being a 'voice-controlled' vacuum robot not too dissimilar to commercially-available voice interfaces. This is enhanced by the actions of the researcher in walking participants towards the robot, by directing their gaze at it, and by guiding participants to "`instruct the vacuum cleaner`" (lns. 1–2) to clean one of the areas. The request to and the subsequent response from the vacuum robot are used to buttress the illusion that the participant is the one controlling the vacuum robot.

This illusion, or *fiction*, is the interactional outcome of the researchers to orchestrate the voice-controlled robot. The 'invisible' work—the work that reveals the 'reality' of this orchestration—is abstracted away and participants are only presented with 'enough' of the truth to partake in the study. In this sense, we nominally refer to this as the 'front stage': the primary researcher provides the participant with the outline of a scenario, scaffolding their completion by handing them the 'control' of the robot (lns. 11-22) and the participant then instructs the robot (lns. 24–25), which

[4]Here we consider 'positive' responses to be ones where the participant's instruction matches that of the subsequent action of the robot, although this classification is subjective and the nomenclature arbitrary.

seemingly responds to the participant through a simulated voice (ln. 34) and movement (lns. 37–39), as per the explanation of the study premise to the participant. This sort of interaction is cogently represented as a '*triad of fiction*' in Figure 4:

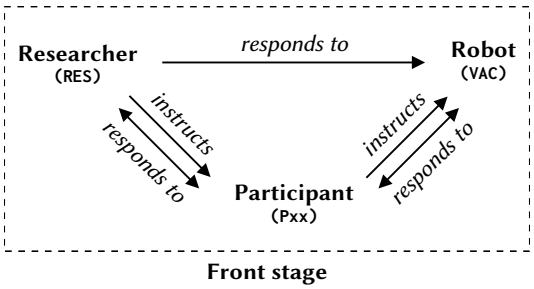


Fig. 4. How interaction ostensibly unfolds from the participant's perspective

In the study, the researcher guides the participant through the scenarios which they complete by instructing the robot—referred to by Martelaro et al. as the "front channel" communication [27]. Moreover, we would contend that the person we have thus far referred to as the primary researcher is fulfilling a role which specifically enables the participant's completion of the study through their 'performance' of sorts[5]. This depiction, however, obscures the invisible work of the other researcher and only consists of the participant's perspective. We now shift our focus on to revealing how the fiction of the voice-controlled vacuum robot is orchestrated.

5 ORCHESTRATING THE VOICE-CONTROLLED ROBOT

We embellish the prior fragment in Fragment 2 by including the 'invisible' work undertaken by the Wizard to orchestrate the voice-controlled robot. As discussed previously, the experiment was consistently framed and presented to the participants to enable the robot to be indiscernible from a fully-implemented system [24]. We will progressively use this second fragment to unpick the coordinated actions of the researchers to explicate how this is accomplished in situ, from both the perspective of the participant (as per the prior fragment) and that of the Wizard.

5.1 The Wizard's involvement

The Wizard responds to the participant by controlling the NEATO vacuum robot as if it were a voice-controlled robot responding to the participant's instruction. Notably, the Wizard's work is completed *in parallel* and *synchronously* to the interaction between the primary researcher, participant, and robot, and hence we present these actions adjacently[6] to the 'front stage' (they are 'backstage', if you will). The Wizard's control of the robot must be aligned with the fiction of the robot presented to participants along with sets of normative conventions connected to voice-controlled interfaces at the time of the study[7] [30].

In the fragment, the Wizard inserts the log marker for the first scenario (lns. 24–25) as the scenario commences. They then craft a verbal response, selecting a partially-completed 'canned' response and inserting the positional parameter from the request ("`forward`", lns. 24–29). In parallel, they send the robot to clean the debris in front of it (set at lns. 1–2 and instructed at lns. 30–33).

[5]This person's actual function in the research project is somewhat inconsequential to the study proceedings.
[6]We have also used arrows (↑——) to signify the point at which the Wizard commenced an action.
[7]The study took place in 2019.

```
                                                    Wizard's actions
 1 RES the- the idea is you can instruct the        selects the zone in front of
 2      vacuum cleaner                                the robot's base
 :
 :
15 RES you have to call them or refer to them as
16      robovac
17 P11         robovac?
18 RES                 yes
 :
 :
22 RES so if you can do it
23     (1.4)
24 P11 er::m:: (3.8) <robovac> . move forward and  clicks menu option to log start
25      cle↑an?                                    of scenario 1
26     (0.8)
27     ┌                                           double clicks 'Ok, I will
28     │(2.2)                                      clean [zone] now', types
29     └                                           'forward'
30     ┌                                           presses enter with right index
31     │(1.3)                                      finger while turning body
32     │                                           to iPad and moving left hand
33     └                                           to tap start cleaning
34 VAC ok . i will clean forward now
35     (1.8)             ↑-------------------------turns back to laptop
36 VAC ((beeps and moves forward out of dock))
37     (3.0)
38     ┌((vacuum motor spins up))                ┐
39 RES └((lifts script up to read next scenario))┘
```

Frag. 2. The actions taken to control the robot in response to Participant 11's (P11) initial request

The robot is stationary and docked in its base station, thus it logically follows that the participant's use of "forward" refers to the debris in front of the dock[8] (see the schematic in Figure 2).

As well as further furnishing the veneer of the voice-controlled robot, this verbal response accounts for the robot's forthcoming action to both the researcher and the participant. The researcher can use this as a signal of the completion of a scenario, which observably happens in the fragment with the lifting of the script in preparation to read the next scenario's description (ln. 39). For the participant, it allows them to continue without waiting to assess the completion of the task. We adjust (in Figure 5) our prior triad of fiction to demonstrate how this 'invisible' work slots in to the operationalisation of the study.

In this revised depiction, the actions of the researcher remain unaltered as part of their role to ostensibly forgo their understanding of the study organisation. However, the Wizard is instructed (indirectly) by the participant and responds through the software for the robot and voice simulation. The Wizard's *improvised decision making* in response to the participant's requests is their commitment to the study underway and demonstrable of their work to uphold their methodological obligations of supporting the researcher and participant in completing the experiment.

5.2 The Wizard's toolbelt

A key part of the Wizard's work involves support tools that are critical to orchestrating the fiction of WOz, so understanding their role interactionally becomes important. They make use of software

[8]There was no correct or incorrect set of positional parameters for each pile of debris, but rather the participant's instructions were to be interpreted on-the-fly by the Wizard.

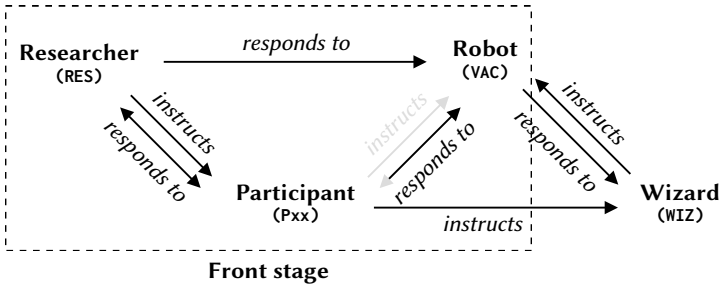
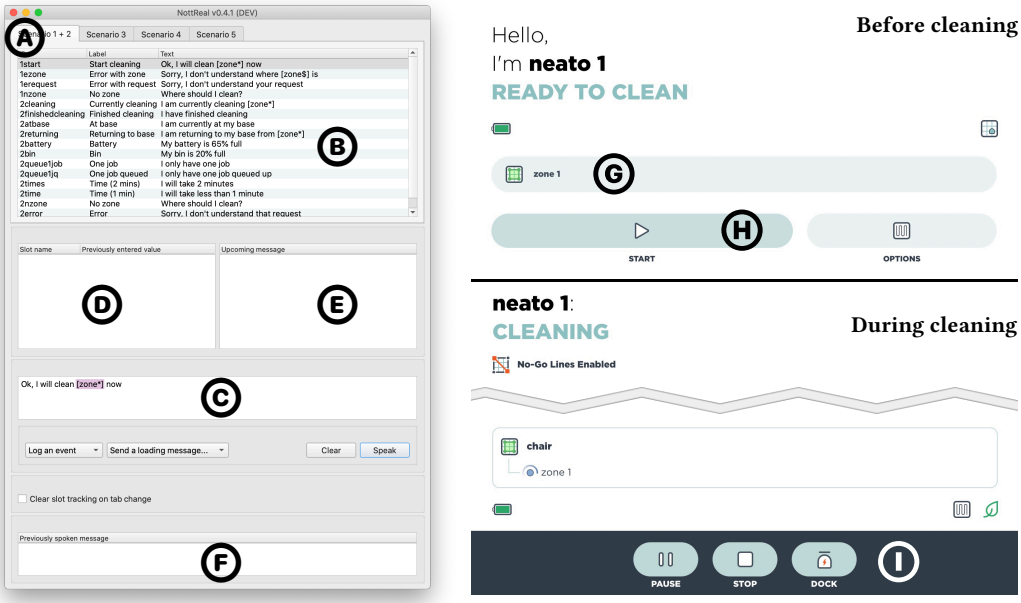


Fig. 5. How the study is orchestrated



(a) NottReal (b) NEATO ROBOTICS® app (© Neato Robotics, Inc. 2020)

Fig. 6. Screenshots of the software used to (a) control the simulated voice and (b) control the robot's movement). NottReal consists of a single control (or 'Wizard') window whereas the Neato app consists of different views with differing sets of controls available at different times, depending on the state of the robot.

to support voice synthesis, which is running on a laptop alongside the Wizard's copy of the study protocol, and an app on an iPad for controlling the robot[9]. In the fragment above, the synthesised voice controlled by the Wizard is noted as being made by the vacuum robot, as are the robot's semi-autonomous[10] movements[11].

[9]To distinguish between the Wizard's use with either of these two devices in the transcripts, the use of the laptop is done through `clicks`, `moves`, `presses` (a key), or `types`, in comparison to the use of the iPad which is done through `taps`.
[10]We use the term *semi-autonomous* because while the Wizard can set the vacuum robot to undertake cleaning of specific zones, they were not controlling the robot's specific motor movements.
[11]Although the vacuum robot and the voice are two decoupled systems, they are present as one on the front stage and thus we amalgamate them into a single 'interactant' in the transcripts.

We developed and used an open-source application for voice-based WOz studies, NottReal (see Figure 6a), to generate and log the synthesised verbal responses [31]. Summarily, a Wizard can send 'pre-scripted' responses by double-clicking them or modify them by single-clicking. Custom responses can be typed. Some pre-scripted responses contain 'slots', shown in square brackets, that must be edited before the response is spoken. We see the Wizard use a number of the app's features in Fragment 2: they log responses for the first scenario (lns. 24–25), they select a pre-scripted response that contains a slot ("`Ok, I will clean [zone] now`", lns. 27–28) and they fill the slot ("`forward`", lns. 28–29) before sending the response (ln. 30).

Figure 6b is two screenshots from the robot manufacturer's app. The first is when the robot is stationary at its 'base' or 'dock'[12], and the second while the robot is cleaning a specific 'zone'[13]. In terms of operation, the Wizard selects a 'zone' to clean (G)—which takes up to around 30 seconds from instructing the app to the request being applied. The Wizard can then instruct the robot to start cleaning (H) this zone. Once cleaning, the Wizard can either pause the cleaning, stop the robot in its current position, or return it to its base (I). In Fragment 2, the Wizard selects the zone in front of the robot's dock to clean (lns. 1–2) and then taps *Start* (lns. 32–33) once they have prepared the verbal response (lns. 27–29). The Wizard's early decision to select this zone can be explained by the fact that following a number of trials of the study it was found that participants typically pick the nearest zone to clean first (i.e. in front of its base). This also reveals the work of the Wizard to *pre-empt* the participant's next move by preparing a candidate next action.

In this section, we have introduced the work to orchestrate the fiction of the voice-controlled vacuum robot, and how this is practically accomplished using two software tools and through coordinated collaborative action between the primary researcher and the Wizard. Next, we will take a closer examination at the tensions resulting from the study's contingencies.

6 DEALING WITH CONTINGENCIES

By contingencies we mean that, as with *any* following of instructions, these instructions will always be insufficient in some way or other in circumstances of enacting them (instructed actions) [12], and thus dealing with emerging contingencies in their application / following becomes a continual problem. We now turn our attention to a situation in which multiple requests are made by a participant during a later scenario. The participant's actions expose the Wizard's limited faculties in dealing with the tensions of upholding their methodological and preformative obligations. Fragment 3 focuses on a different participant—`P18`—who is completing a scenario in which they must interrupt the robot while it is cleaning, and send it to its base or to complete another task.

6.1 Cracks in technical implementations

Our first point is about the Wizard's response to the participant's request to the robot to *stop*. The fragment starts with the vacuum robot moving around the so-called 'forward' zone and vacuuming, with `P18`'s command to "`stop current operation`" (ln. 5) issued as this cleaning unfolds. The Wizard prepares to respond to `P18`'s request, that of a stop command, as necessitated by the objective of this scenario, by moving their mouse pointer over the prepared canned response for such a request ("`Ok, I will stop cleaning now`", lns. 7–8) and their hand over the *Dock* button on the mobile app. As the participant completes their instruction, the Wizard double-clicks the trackpad to send the desired response and taps the *Dock* button on the iPad momentarily after (lns. 10–11). Although the

[12] This is the mains-connected charging station for the robot, and where it must be to start cleaning tasks to a specific zone. The manufacturer interchangeably uses the terms 'base' and 'dock' to refer to this unit.

[13] Zones are pre-configured areas a robot can be sent to clean exclusively, but a robot must be in its base to be sent to clean a zone. Additionally, this pre-configuration requires a robot to map a space by moving around it fully and then a user drawing the boundaries of each zone on the resulting map

```
                                                          Wizard's actions
1  VAC ┌((moving forward and vacuuming)) (1.1)┐
2  P18 └(0.7) robovac (2.0)                   ┘
3              ↑-------------------------------------moves right hand to trackpad and
4  VAC ((turns 90 degrees left))                       left index finger over ‘Dock’
5  P18 ┌please stop current operation                 button on trackpad
6      │↑                      (0.3) (0.2) (0.3)┐
7      │---------------------------↑-----↑-----+-----moves pointer over ‘Ok, I will
8      │                           |     |     |      stop cleaning now’
9  VAC └((moving forward))         |     |     ┘
10                                 ------+-----------taps ‘Dock’ button
11                                       ------------double clicks on trackpad
12 VAC ((stops moving and motor spins down))
13 P18 robov┌ac┐
14 VAC      └ok┘ . i will stop cleaning now
15     (1.0)                                         moves pointer over ‘I will return
16 P18 robovac . i’m going to direct you to a new     to my base’
17      area to cle:an
18          ↑---------------------------------------moves pointer over ‘Sorry, I don’t
19     (1.2)                                          understand your request’
20                                                   double clicks on trackpad
21     (0.8)
22 VAC sorry . i don’t understand your request
23 P18 °hm° (1.1) robovac (1.0) there’s another area
24      to clean (1.2) (1.0) ┌you need-   ┐
25               ↑-----↑------+------------+-----------moves pointer over ‘Where
26                     |      |            |            should I clean?’
27                     -------+------------+-----------double clicks on trackpad
28 VAC                        └where should┘ i clean?
```

Frag. 3. P18 issues multiple instructions to the robot

Wizard has the ability to stop the robot by tapping the *Stop* button, piloting of the study revealed that pausing or stopping the robot was problematic—the robot would be in the middle of the space and upon resumption would be unable to verify its location due to limitations of the robot’s sensors. Furthermore, the robot starts each task from its base, requiring a return to it irrespective of the participant issuing a subsequent command to clean or not. As a result, tapping *Dock* pauses the robot’s movement for half a minute or so before the robot begins returning to its base autonomously, allowing time for a subsequent command by a participant to be issued. The study protocol reflects this requirement, as shown in Figure 3a.

What is of interest here is not necessarily the peculiarities of what the Wizard is forced to do in this study, but rather that these oddities are emblematic of the work in running WOz studies with partially incomplete systems. Namely, that such studies require *papering over the cracks* to present participants an integrated and working ‘intelligent’ system. The definition and utilisation of the method in iterative development approaches will see some WOz studies occur with less-refined or integrated systems that may be improved at a later stage of the development cycle. Papering over these cracks can become a core effort in running a WOz study, and represents an elaboration of the work to establish and maintain the *fiction*. In our study, to meet the obligations of the fiction while masking the underlying implementation details, the Wizard stops the robot within a second of the participant’s request (lns. 5–12) while selecting the pre-scripted verbal confirmation and preparing a second verbal response to account for the next semi-autonomous action of the robot.

6.2 Coherence and nuance

This preparatory action by the Wizard, of moving their mouse pointer over a second prepared response ("`i will return to my base`", lns. 15–16) after the robot stops moving (ln. 12) warrants further consideration. Delivering this request is the logical next step, as stated above, as the robot will begin to move towards the base as a result of the *Dock* button being tapped (ln. 10) as well as the task at hand given to the participant. It would be *incoherent* for the synthesised voice not to correspond to the physical movements of the robot, or rather, you could say that this prepared response is necessary to adequately account for the robot's pending movement and maintain the coherence between the simulated voice and the physical movement of the robot.

While the Wizard's role in the study is to ensure this coherence between voice and motion is maintained as part of simulating the state of the robot, this becomes challenged by a higher-order obligation. Namely, `P18` issues a follow-up request in response to the robot coming to a stop ("`i'm going to direct you to a new area to clean`", lns. 16–17), one for which there is no rehearsed response. This case serves as an exemplar of a situation that was not practised or scripted, for which the Wizard has no stock answer to fall back on—the Wizard must improvise. Furthermore, the Wizard demonstrably changes their planned next response ("`i will return to my base`", lns. 15–16), to respond to the participant's subsequent request:

```
13 P18 robov┌ac┐
14 VAC      └ok┘ . i will stop cleaning now
15     (1.0)                                        moves pointer over 'I will return
16 P18 robovac . i'm going to direct you to a new    to my base'
17      area to cle:an
18          ↑-----------------------------------------moves pointer over 'Sorry, I don't
19     (1.2)                                         understand your request'
```

This reveals a tension in the Wizard's work in that they must maintain a *turn-by-turn coherence* in interaction, in addition to the coherence of the robot's movements with the voice interface. While such situations may not be all too common, they do highlight the tension of the Wizard's work, in which participant requests can be potentially unbounded by what they ask and Wizards must, within a moment, be prepared to abort a previously planned action in favour of another more pressing and seemingly more coherent one. As such, coherence becomes a multifaceted obligation for the Wizard to uphold in their contribution to the study.

As we stated before, requests by participants to clean without the inclusion of a specific area should be responded to with the question "`where should i clean?`" (as shown on lns. 24–28). However, the Wizard selects a 'negative' response (ln. 22) to `P18`'s request of "`. . . a new area to clean`" (lns. 16–17). In retrospect, this discrepancy likely stems from the nuance of language and the idea of enacting a typical 'natural language understanding' system. `P18`'s request ("`i'm going to direct you to a new area to clean`", lns. 16–17) *could* be interpreted as a preparatory account of forthcoming guidance for the robot to move to where the participant specifies either through gestures or through an elaborated series of instructions. Alternatively, this could be interpreted as a request for the robot to clean, but without a specific area. In other words, there are two possible interpretations.

The Wizard responds to the request as if `P18` is to transcend into a series of actions not supported by the study protocol (either manual control of the robot or gesture recognition)[14]. Yet interpreting the request lexically, the Wizard could have responded with "`where should i clean?`". As such, given the outline script being followed, the response by the Wizard could be classified as a *Wizard error* [34]. Poignantly, this situation exemplifies the tension of improvising *quickly* while dealing with the nuance of language in parsing requests and selecting/assembling appropriate responses.

[14]In prior cases in which a participant attempted either of these, the response generated for the participant was "`sorry, i don't understand your request`".

Above, we have listed some of the tensions the Wizard faces in dealing with various contingencies while upholding their methodological obligations. Namely, the Wizard is required to paper over cracks in technical implementations as well as dealing with challenges arising from competing demands for coherence, and the subtle nuances of language that can be difficult to assimilate in an on-the-fly research study. In our final tranche of data, we will look at what happens when things go awry and a technical issue leads to a breakdown in the study proceedings.

7 BREAKDOWN

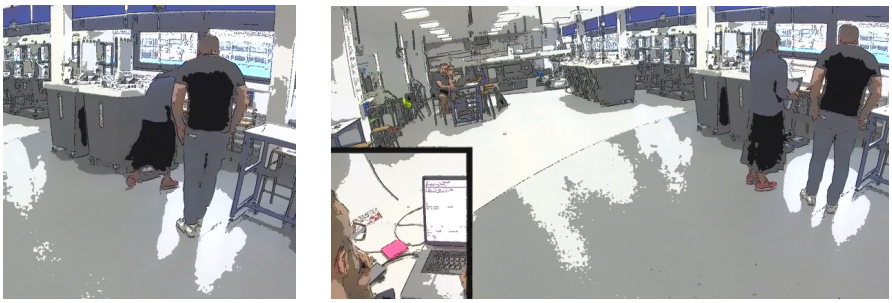
Our final fragment, which is an exemplar of a technical problem with the robot unfolding during the study, follows on from the previous fragment. The robot has returned to its base and has been sent out to clean the 'forward' zone once again. P18 has been asked to queue up a task after the current one completes but the robot has stayed in front of its base for half a minute, it then beeps before returning to its base.

		Wizard's actions
120	RES i think its having a problem now	moves pointer to select text box
121	(0.7)	
122	┌	types 'I have a problem'
123	│(2.3)	hits enter
124	└	
125	VAC ((quickly rotates anti-clockwise so base is	
126	behind it))	
127	VAC i have a problem	
128	((RES and P18 looking at robot))	
129	RES yeah! ┌it's returning to the base? ┐	
130	VAC └((moves slowly backwards into │	
131	base for 5.9s)) ┘	
132	VAC ((chimes to signify docked))	
133	RES ((moves forward and kneels to look	
134	at the robot, adjusts speaker on top))	
⋮	⋮	
146	RES ((stands up and moves back next to P18))	
147	shall it we try again?	
148	ok >i'm going to instruct the<	clicks 'Ok, I will clean [zone*]
149	robovac clean in front of you	when I finish [what]', then highlights
150		and deletes from '[zone*] to end,
151		types 'in front of me', hits enter
152		taps start cleaning button
153	VAC ok . i will clean in front of me	

Frag. 4. Recovering from a technical breakdown that occurs during P18's study

7.1 (Im)perceptibly off-script

The framework of participation [15] enables the researcher to inspect and account for the robot's actions in line with the established fiction—a response from the vacuum robot can be recognisable in its conformance (or not) to this established 'way of working'. The first thing to note here, though, is that what happens in this fragment is 'off-script'. The vacuum robot, by being stationary for an extended period of time, is not performing in an expected manner, which the primary researcher demonstrates through musing of the vacuum robot "having a problem" (ln. 120). This utterance provides an account for the robot's seemingly off-script action in response to what would typically be expected. The presupposed technical breakdown that is occurring introduces new contingencies



(a) RES crouches down to look at VAC (ln. 136)

(b) RES takes a step back and makes a request to the vacuum robot while P18 watches (ln. 148)

Fig. 7. A still from the video corresponding to Fragment 4

that preclude both researchers from undertaking this recognisable and shared way of working, and thus requires a new jointly-managed course of action.

As the vacuum robot begins returning to its base (ln. 125), the Wizard is made aware of an issue that the robot cannot locate itself through an error message displayed in the mobile app. The Wizard responds to the researcher's verbal account of a problem with the vacuum robot by typing and sending the voice response "`i have a problem`" (ln. 130). At this point both researchers continue to uphold their methodological obligations to conceal the Wizard from the participant. The intervention from the Wizard, in response to the researcher, and the moving forward and crouching down to examine the vacuum robot by the researcher (see Figure 7a) marks a momentary pause in the study proceedings, as the participant takes a step back. Moreover, from the perspective of those 'in the know' of the study protocol, i.e. everyone but the participant, the response "`i have a problem`" is perspicuously produced in response to, and directed at, the researcher. It provides an account for the researcher and implicates them into attending to the problem. In this sense, it is an off-script response produced such that it could be interpreted as an automated verbal response from the voice-controlled robot. Potentially, only the researcher would perceptibly orient to this as being from the Wizard. As such, it demonstrably serves as an improvised approach to using the study apparatus to enable communication between the Wizard and the researcher that is ostensibly concordant with fiction to the participant but not to the primary researcher.

Following a readjustment of the portable speaker on the top of the vacuum robot, the researcher states that *they* will "`instruct the [robovac]`" (lns. 148–149). Through this, the researcher is attempting to return the study to the point at which the breakdown occurred, and turn control of the proceedings over to the participant to complete the commenced scenario. This statement is, in effect, the reverse of the Wizard's "`i have a problem`" (ln. 127)—the researcher provides a preparatory account that is inconspicuous among the milieu on the front stage yet implicates the Wizard in treating the forthcoming request as an instruction to which they should normatively respond. Following their account, they take a step back (see Figure 7b), issue their instruction ("`clean in front of you`", ln. 149), and the vacuum robot ostensibly responds, building in the referential parameters from the request ("`in front of me`", lns. 149–153) as per the fictionalised modus operandi of the study.

In this final section we presented an example of a technical breakdown that required resolution work from the primary researcher. Despite being unrehearsed, and involving an unexpected hiatus to study proceedings, both the researcher and the Wizard accountably *coordinated* their actions through means that, from the participant's perspective, are mundane to the study, yet to the researchers they were perspicuously 'off-script'.

8 DISCUSSION

We now bring together our findings on the orchestration of Wizard of Oz studies, the coordinated and cooperative action which drew upon the study resources, and our insights for researchers and designers intending to practise the method.

8.1 Orchestrating Wizard of Oz studies

Firstly, we will synthesise the explicated research practice that unfolded as the researchers worked to uphold the various *methodological obligations*.

8.1.1 Upholding methodological obligations. The primary purpose of this study, and of many WOz studies [37], is to understand how users would interact with a future system[15]. This ties in with the requirements of this future system being specifiable and the simulation possible [10]. To accomplish this, the primary researcher's role in our study was to *scaffold* the participants' interactions with the vacuum robot, providing them with *just enough* information to participate while withholding any specific details of the actual implementation or vocabulary the participant should use. For the most part, this entails explaining the premise of the study, introducing the fiction, and guiding the participant through each scenario. The presentation of the study consists of many resources, beyond the protocol and outline script (see 4.1) that the primary researcher draws upon to do this, including the setting and the visibility/tangibility of the presupposed voice-controlled vacuum robot. The primary researcher consistently designs their actions to be made sense of as *separate* from the operation of the voice-controlled robot, even though they are not.

The Wizard is obligated to support the primary researcher's actions and, ultimately, enable the participant's completion of the study. If the primary researcher's role is to *present the fiction* of the voice-controlled vacuum robot, the Wizard's role is to *enact the fiction* through their covert control of the vacuum robot (see 5.1). As seen when discussing the cracks in the technology used (see 6.1), this introduces tensions that the Wizard must manage by masking the workings of the underlying implementation and presenting an interface that acts ostensibly *coherently* with the described system. As such, the Wizard's role in the study is to control the vacuum robot as if the vacuum robot is controlling itself and, in turn, conceal their presence in orchestrating of the experiment. The Wizard's methodological obligation, which underscores the operationalisation of their role, is that they must treat participants' requests consistently within and between studies. A protocol was established that provided a framework for how each scenario is expected to unfold, and an outline script provided the 'sorts' of requests that were to be treated as valid and the sorts of responses the vacuum robot should 'generate'. However, the contingent nature of participants' requests preclude any comprehensive preparation and thus the Wizard must examine each and every request *on-the-fly* and respond accordingly based upon these resources. As Martelaro et al. remarks, in WOz studies it is important to "improvise in-the-moment based on the user interaction" [27, p. 2074].

In turn, the primary researcher is implicated into treating and accounting for any and all responses from the vacuum robot as of those generated by a system as, after all, it is *their obligation* to present and maintain the fiction. Therefore, the primary researcher and the Wizard *collaborate* to complete each experiment. The primary researcher presents the fiction and accounts for the vacuum robot's supposed actions, while the Wizard works to put into practice the fiction by appropriately controlling the vacuum robot. This is how both parties work together to deliver a system that is "convincing" [10] such that participants do not believe it to be human-operated. Next, we deepen this point by arguing that this collaborative action is *performative* in its accomplishment.

[15] As we briefly covered in our related work section, there are, however, other purposes for Wizard of Oz studies.

8.1.2 Performatively orchestrating the study. Researchers must construct and present a believable story for the participant prior to their exposure to the system, and this fiction must then be *maintained* by the researchers throughout the study as the participant interacts with the system. Rather than this fiction being a 'given' or participant-dependent [10], we have explicated how this is an *interactional outcome* of the actions of those in the study (i.e. the researchers and participants alike) *through* their interaction with each other and the simulated system. The researchers scaffold participants' completion of the study and implement the fiction of the voice-controlled vacuum robot. Participants are told by the primary researcher that the vacuum robot responds to them, and for all intents and purposes, it seems to do so. This, in turn, enables the participant to get on and complete the study and (potentially unwittingly) treat the fictional system in a manner consistent with the established fiction. This is the *'front stage'* work of the study, consisting of only 'front channel' communication [27].

The work of the Wizard, who controls the robot and makes this illusionary 'front stage' experience happen, takes place 'backstage' and thus remains invisible to the participant. We showed how the participants' instructions are acted upon by the Wizard through their control of the robot's movement and synthesised voice. From the perspective of the researcher and participant, the vacuum robot *is* ostensibly voice-controlled and moves in response to the participant's instruction. More so, the primary researcher *must* ostensibly treat the robot's actions as if this intermediary person and 'backstage' were non-existent to sustain this concocted fiction. Yet, however, the Wizard and the researcher coordinate their actions and demonstrably communicate with each other throughout. This 'back channel' communication [27] occurs 'in plain sight' of participants, through the primary researcher and Wizard's shared understanding of the protocol and script. Both parties to these resources can *competently* interpret the actions of each other and the robot without talking to each other directly, and, for example, utilise this shared knowledge to exchange information when there was a system breakdown.

It is this coordinated and collaborative action—between front and backstage—which elucidates the primary researcher's role as that of some sort of *performer*. Both researchers have, of course, 'rehearsed' this through preparing the study protocol, writing an outline script, and the running of pilots to develop competency [13]. However, with each participant there necessarily entails *improvisation* by virtue of the contingent nature of interaction. Participants can effectively say or do anything[16] and the researchers must respond *in character* to this fortuity to maintain the fiction being spun. This improvisation itself—by both the participants and the researchers—provides a key resource to supporting the overall design process of the novel technologies, allowing for exploration of the design space [14, 41]. By allowing participants to improvise to complete the scenarios, and the Wizard to make ad hoc decisions in their responses, data can be collected throughout the study to inform the design and development of future systems as a requirements-gathering exercise.

Other contingencies may also present themselves which also entail further improvisation. For example, in the last fragment we presented, the Wizard issues an improvised response of the robot having a problem. This response served a duality: it was to convey a technical issue with the robot to the participant and the researcher, and also to validate the researcher's verbal musing of a potential problem. Through this enactment, the Wizard demonstrates how they coordinate with the primary researcher while ensuring the methodological validity of the study by not revealing their control over the robot. In other words, they implicate the primary researcher into cooperating with the resolution of the problem within the study's participation framework, as per the methodological obligations of running the study without revealing the work backstage.

[16]There are obvious assumed limits to what a participant is likely to say or do as per the context and established framework, but participants still retain a broad range of options for how to enact the experimental scenarios presented to them.

8.2 Implications for researchers and designers

Finally, we discuss how researchers and designers can conceptualise and design resources to apply the WOz method based on our findings. While the researcher, Wizard, and participant each had a distinct role in these studies, their actions were coordinated with each other—each person's actions were contingent upon and implicate each others' subsequent actions. Of course, cooperation between participants and researchers is a mundane feature of most, if not all, user studies in HCI, but above we have sought to explicate *how* this cooperation becomes manifested through the coordinated actions in our WOz study. In line with CSCW's pre-eminent concern with the design and use of technologies to support cooperative work [38], we now examine how this coordination drew upon the study protocol, the outline script, the articulated division of labour, the rehearsal of the study, and the improvised nature of running the study.

WOz studies require careful planning to concoct and enact a fiction. We did this by writing a **study protocol** that provided a narrative coherent with the study setting and which was used to frame the study. This protocol, rather than offer a step-wise instruction, served as guidance to inform the practical action of the researchers during the study—it offered a plan but expected the situated detail of how each study unfolds to be defined in its enactment [44]. A complete plan is neither feasible in terms of foresight nor in terms of being used in haste during trials. The primary researcher used this as guidance to inform their interactions with the participant and scaffold their completion of the study on the front stage. The Wizard, conversely, used this document to contextualise their observations of the situation on this front stage, to coordinate their actions with the participant and the researcher [43]. In practical terms, this enabled the Wizard to identify and prepare for the requests that they would be expected to deal with in each succeeding scenario.

We also used this study protocol to write an **outline script**, that we implemented in software designed for WOz studies. This script was crafted by identifying which sorts of requests should and should not 'work'—some requests were to be responded to negatively because they did not specify the detail needed for the methodological concerns of the study. This outline script was pivotal in allowing the Wizard to respond quickly to a vast array of requests and imbuing the robot's responses with 'liveness' [20] by allowing for customised responses— such as participants words being embedded into responses. Our software also enabled the Wizard to create new ad hoc responses, allowing for quick resolution of arising contingencies, from the routine of unexpected requests through to technical breakdowns. By being a *working document*—i.e. it was updated throughout the studies as new requests were encountered—it also supported the Wizard's obligation to ensure consistency within and across studies, bolstering the validity of the approach.

Our study had two researchers undertaking ostensibly independent but coordinated roles. Running a study with two researchers introduces additional challenges to overcome, and requires a closer consideration of how to support this **division of labour** [42] through the limited resources available during trials. In our case, the study protocol and outline script were prepared by both researchers together to support the articulation of the tasks involved in the performance and how these were to be divided among the actors. This articulation of the roles enabled the coordinated and cooperative action between the researchers in both their individual and shared responsibilities during the studies.

There also remains the prohibition of direct communication between the researchers during the studies. As well as relying on the shared resources to coordinate, the two researchers **rehearsed** through pilot studies. Of course, pilot studies help to catch obvious study design issues, but they also enable both researchers to *develop competency* in the upholding their obligations, and due to the informality of these rehearsals, shape the framework of participation that will be introduced with each new participant.

Finally, the researchers will be isolated from each other and both will have to make multiple split-second decisions on-the-fly during the study. There is an obligation to 'roll with it' irrespective of the contingencies faced—this is *the performative* element at work. The primary researcher should attempt to keep the activity of the study in line with the established framework of participation, and the work backstage by the Wizard needs to be focused on orchestrating the presented fiction in line with the methodological obligations. Given this on-the-fly nature, and as *enabled* by the rehearsals, WOz studies are **cooperatively accomplished** by the researchers working to maintain the normative way of working, established in the fiction.

9 CONCLUSION

The Wizard of Oz method is used in studies of users with prototypical technologies that include some form of simulated 'intelligence'. Participants are not made aware of this simulated component, with the approach necessitating users be presented with a fiction on the grounds of methodical validity. We examined our practice of the method in the context of seeking to understand how users instruct a voice-controlled vacuum robot. The analysis in this paper does not just focus on 'the Wizard', i.e. the person who covertly controls the simulated vacuum robot, but also how the method is collaboratively organised between multiple researchers 'running the study' and the participants. We identified how this routine organisation draws upon preparatory resources including a protocol and an outline script, and is rehearsed through pilots. These resources were used to scaffold and coordinate the actions between the researchers during the studies due to there being no opportunity for direct communication. In discussing our findings, we synthesised the methodological, performative, and cooperative work, identifying the on-the-fly nature of implementing the method as required by the experimental contingencies. We closed our discussion with insights for researchers and designers to apply the method and conclude that these experiments should be regarded as cooperatively accomplished trials, enabled by the preparatory resources and rehearsal, and underpinned by the methodological obligation to uphold the established fiction.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/V00784X/1, EP/M02315X/1, and EP/N014243/1] and the Department for International Development. All research data supporting this publication are directly available within this publication.

REFERENCES

[1] J Maxwell Atkinson and John Heritage. 1984. Transcript Notation. In *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK, ix–xvi. https://doi.org/10.1017/CBO9780511665868

[2] L Frank Baum. 1900. *The Wonderful Wizard of Oz*. George M Hill Company, Chicago, IL, USA.

[3] Jacob T Browne. 2019. Wizard of Oz Prototyping for Machine Learning Experiences. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, LBW2621:1–LBW2621:6. https://doi.org/10.1145/3290607.3312877

[4] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies — Why and How. *Knowledge-Based Systems* 6, 4 (1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[5] Robert L Erdmann and Alan S Neal. 1971. Laboratory vs. Field Experimentation in Human Factors—An Evaluation of an Experimental Self-Service Airline Ticket Vendor. *Human Factors* 13, 6 (dec 1971), 521–531. https://doi.org/10.1177/001872087101300603

[6] Andrew Finke. 2019. Lake: A Digital Wizard of Oz Prototyping Tool. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, Article SRC05, 6 pages. https://doi.org/10.1145/3290607.3308455

[7] Joel E Fischer, Stuart Reeves, Tom Rodden, Steve Reece, Sarvapali D Ramchurn, and David Jones. 2015. Building a Birds Eye View: Collaborative Work in Disaster Response. In *Proceedings of the 33rd Annual ACM Conference on Human*

Factors in Computing Systems (CHI '15). ACM, New York, NY, USA, 4103–4112. https://doi.org/10.1145/2702123.2702313

[8] W Randolph Ford and Raoul N Smith. 1982. Collocational Grammar As a Model for Human-computer Interaction. In *Proceedings of the 9th Conference on Computational Linguistics - Volume 2 (COLING '82)*. Academia Praha, Czechoslovakia, 106–110. https://doi.org/10.3115/990100.990122

[9] Juliano Franz and Derek Reilly. 2017. TangiWoZ: A Tangible Interface for Wizard of Oz Studies. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2584–2590. https://doi.org/10.1145/3027063.3053254

[10] Norman M Fraser and Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language* 5, 1 (jan 1991), 81–99. https://doi.org/10.1016/0885-2308(91)90019-M

[11] Harold Garfinkel. 1967. *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ, USA.

[12] Harold Garfinkel. 2002. Instructions and Instructed Action. In *Ethnomethodology's Program: Working Out Durkheim's Aphorism*, Anne Warfield Rawls (Ed.). Rowman & Littlefield Publishers, Inc., Lanham, MD, USA, Chapter 6, 197–218.

[13] Harold Garfinkel and D Lawrence Wieder. 1992. Two Incommensurable, Asymmetrically Alternate Technologies of Social Analysis. In *Text in Context: Contributions to Ethnomethodology*, Graham Watson and Robert M Seiler (Eds.). Sage Publications, Newbury Park, CA, USA, 175–206.

[14] Elizabeth Gerber. 2007. Improvisation Principles and Techniques for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1069–1072. https://doi.org/10.1145/1240624.1240786

[15] Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press, Philadelphia, PA, USA. 344 pages.

[16] John D Gould, John Conti, and Todd Hovanyecz. 1983. Composing Letters with a Simulated Listening Typewriter. *Commun. ACM* 26, 4 (April 1983), 295–308. https://doi.org/10.1145/2163.358100

[17] Paul Green and Lisa Wei-Haas. 1985. *The Wizard of Oz: a tool for rapid development of user interfaces. Final report.* Technical Report. University of Michigan.

[18] Marc Guyomard and Jacques Siroux. 1988. Experimentation in the Specification of an Oral Dialogue. In *Recent Advances in Speech Understanding and Dialog Systems*, H Niemann, M Lang, and G Sagerer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 497–501.

[19] Christian Heath and Paul Luff. 1992. Collaboration and control: Crisis management and multimedia technology in London Underground control rooms. *Computer Supported Cooperative Work* 1, 1990 (1992), 69–94. https://doi.org/10.1007/BF00752451

[20] Jonathan Hook, Guy Schofield, Robyn Taylor, Tom Bartindale, John McCarthy, and Peter Wright. 2012. Exploring HCI's Relationship with Liveness. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 2771–2774. https://doi.org/10.1145/2212776.2212717

[21] IoTUK. 2017. *4IR - The Next Industrial Revolution*. Technical Report October. Digital Catapult. https://iotuk.org.uk/wp-content/uploads/2017/10/Digital-Catapult-4IR.pdf

[22] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. ACM, New York, NY, USA, Article 9, 9 pages. https://doi.org/10.1145/3405755.3406119

[23] Christopher Labrador and Pai K Dinesh. 1984. Experiments in speech interaction with conventional data services. In *1st IFIP International Conference on Human-Computer Interaction (INTERACT '84)*, Brian Shackel (Ed.). Amsterdam, Netherlands, London, UK, 225–229.

[24] Ashok Malhotra. 1975. Knowledge-Based English Language Systems for Management Support: An Analysis of Requirements. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1 (IJCAI'75)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 842–847.

[25] Elena Márquez Segura, Annika Waern, Luis Márquez Segura, and David López Recio. 2016. Playification: The PhySeEar Case. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16)*. ACM, New York, NY, USA, 376–388. https://doi.org/10.1145/2967934.2968099

[26] Nikolas Martelaro and Wendy Ju. 2017. WoZ Way: Enabling Real-Time Remote Interaction Prototyping & Observation in On-Road Vehicles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 169–182. https://doi.org/10.1145/2998181.2998293

[27] Nikolas Martelaro, Sarah Mennicken, Jennifer Thom, Henriette Cramer, and Wendy Ju. 2020. Using Remote Controlled Speech Agents to Explore Music Experience in Context. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. ACM, New York, NY, USA, 2065–2076. https://doi.org/10.1145/3357236.3395440

[28] Anna Lisa Martin-Niedecken, Elena Márquez Segura, Katja Rogers, Stephan Niedecken, and Laia Turmo Vidal. 2019. Towards Socially Immersive Fitness Games: An Exploratory Evaluation Through Embodied Sketching. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (CHI PLAY '19 Extended Abstracts)*. ACM, New York, NY, USA, 525–534. https://doi.org/10.1145/3341215.3356293

[29] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 277–284. https://doi.org/10.1145/169059.169215

[30] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[31] Martin Porcheron, Joel E Fischer, and Michel Valstar. 2020. NottReal: A tool for voice-based Wizard of Oz studies. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. ACM, New York, NY, USA, Article 35, 3 pages. https://doi.org/10.1145/3405755.3406168

[32] Stuart Reeves. 2019. How UX Practitioners Produce Findings in Usability Testing. *ACM Trans. Comput.-Hum. Interact.* 26, 1, Article Article 3 (Jan. 2019), 38 pages. https://doi.org/10.1145/3299096

[33] Pedro Reynolds-Cuéllar and Cynthia Breazeal. 2017. Emotional Robocoaster: An Exploration on Emotions, Research Methods and Introspection. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17 Extended Abstracts)*. ACM, New York, NY, USA, 561–567. https://doi.org/10.1145/3130859.3131337

[34] Laurel Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction* 1, 1 (Aug 2012), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

[35] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (jan 1974), 696. https://doi.org/10.2307/412243

[36] Stephan Schlögl, Anne Schneider, Saturnino Luz, and Gavin Doherty. 2011. Supporting the Wizard: Interface Improvements in Wizard of Oz Studies. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction (BCS-HCI '11)*. BCS Learning & Development Ltd., Swindon, UK, 509–514. https://doi.org/10.14236/ewic/HCI2011.85

[37] Stephan Schlögl, Gavin Doherty, and Saturnino Luz. 2014. Wizard of Oz Experimentation for Language Technology Applications: Challenges and Tools. *Interacting with Computers* 27, 6 (may 2014), 592–615. https://doi.org/10.1093/iwc/iwu016

[38] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW Seriously: Supporting Articulation Work. *Computer Supported Cooperative Work* 1, 1-2 (mar 1992), 7–40. https://doi.org/10.1007/BF00752449

[39] Marcos Serrano and Laurence Nigay. 2009. Temporal Aspects of CARE-Based Multimodal Fusion: From a Fusion Mechanism to Composition Components and WoZ Components. In *Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI-MLMI '09)*. ACM, New York, NY, USA, 177–184. https://doi.org/10.1145/1647314.1647346

[40] David Sirkin, Kerstin Fischer, Lars Jensen, and Wendy Ju. 2016. Eliciting conversation in robot vehicle interactions. In *2016 AAAI Spring Symposium Series*. AAAI, Palo Alto, CA, USA, 8.

[41] David Sirkin and Wendy Ju. 2015. Embodied Design Improvisation: A Method to Make Tacit Design Knowledge Explicit and Usable. In *Design Thinking Research. Understanding Innovation*, Hasso Plattner, Christoph Meinel, and Larry Leifer (Eds.). Springer, Cham, Switzerland, 195–209. https://doi.org/10.1007/978-3-319-06823-7_11

[42] Anselm Strauss. 1985. Work and the Division of Labor. *The Sociological Quarterly* 26, 1 (1985), 1–19.

[43] Lucy Suchman. 1997. Centers of Coordination: A Case and Some Themes. In *Discourse, Tools and Reasoning: Essays on situated cognition*. Springer Berlin Heidelberg, Berlin, Germany, 41–62. https://doi.org/10.1007/978-3-662-03362-3_3

[44] Lucy A Suchman. 1985. *Plans and Situated Actions: The Problem of Human Machine Communication* (1 ed.). Cambridge University Press, Cambridge, UK. 220 pages.

[45] Hamish Tennent, Wen-Ying Lee, Yoyo Tsung-Yu Hou, Ilan Mandel, and Malte Jung. 2018. PAPERINO: Remote Wizard-Of-Oz Puppeteering For Social Robot Behaviour Design. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18)*. ACM, New York, NY, USA, 29–32. https://doi.org/10.1145/3272973.3272994

[46] Greg Walsh and Eric Wronsky. 2019. AI + Co-Design: Developing a Novel Computer-Supported Approach to Inclusive Design. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. ACM, New York, NY, USA, 408–412. https://doi.org/10.1145/3311957.3359456

[47] Kevin F White and Wayne G Lutters. 2003. Behind the Curtain: Lessons Learned from a Wizard of Oz Field Experiment. *SIGGROUP Bull.* 24, 3 (Dec. 2003), 129–135. https://doi.org/10.1145/1052829.1052854

[48] Robin Wooffitt. 1994. Applying Sociology: Conversation Analysis in the Study of Human-(Simulated) Computer Interaction. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 43, 1 (1994), 7–33. https://doi.org/10.1177/075910639404300103

[49] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 185, 12 pages. https://doi.org/10.1145/3290605.3300415

Received June 2020; accepted October 2020