Wizard-of-Oz Testing as an Instrument for Chatbot Development

An experimental Pre-study for Setting up a Recruiting Chatbot Prototype

Stephan Böhm, Judith Eißer, and Sebastian Meurer

CAEBUS Center of Advanced E-Business Studies
RheinMain University of Applied Sciences
Wiesbaden, Germany
e-mail: {stephan.boehm, judith.eisser, sebastian.meurer}@hs-rm.de

***Abstract*—Chatbots can be utilized to automate various business processes to add value for companies and users – for example, in the form of efficiency enhancement. Throughout the process of chatbot development, the integration of user feedback within a user-centered conversational design process is essential. In our study, we investigated chatbots in recruiting, a field within human resource management that is characterized by a high proportion of repetitive and standardized tasks. This pre-study applies a Wizard-of-Oz approach in which a basic dialog concept is tested in a very early phase of the project, simulating the chatbot functionality by a human operator. In this way, valuable user feedback on the general suitability of the dialog design can be gathered without coding chatbot functionalities. In total, eight users participated in our 60-minute experiment to conceptionally validate our idea and test the simulated Frequently Asked Questions (FAQ) chatbot. The research brought important insights into the basic concept and allowed us to collect new user intents not considered in the design. As a result, the tested concept proved to be suitable and of value for the users. Despite relatively long response times, only one participant suspected that they were not interacting with a chatbot but a human operator. The feedback on the user satisfaction with the completeness of the predefined answers and competence setup of the simulated chatbot was indifferent and rather moderate. However, most of the participants considered the tested scenario as relevant and stated a high user value for implementing the proposed chatbot in a recruiting process. Moreover, the Wizard-of-Oz approach generated appropriate input for improving the chatbot concept (e.g., intents, entities, criteria for satisfaction and acceptance enhancement) and valuable practical insights for developing a recruiting FAQ chatbot aligned to user needs.**

Keywords–chatbot; Wizard-of-Oz testing; prototyping; chatbot development; recruiting.

I. Introduction

Chatbots as a way to automate repetitive stakeholder (i.e., customers, prospects) inquiries in the form of conversational dialogues are more and more implemented into internal and external business communication processes [1][2]. In order to unfold their potential of enhancing the efficiency of such processes, it is imperative to create a suitable conversational design concept considering the envisioned users' requirements and expectations concerning this automation technology [3][4]. The integration of early user feedback is crucial in the development process [5]–[7], which makes it a common practice in technology development processes [3]. One way to yield stakeholder feedback and thus necessary input for the creation and advancement of a chatbot in an early stage without possessing a functional chatbot system is to conduct a Wizard-of-Oz (WOz) experiment [3][6]. In a WOz test, the executors lead the test subjects to believe that they are interacting with a fully developed technological system, whereas it is the test operators themselves acting as such, in this case serving as chatbot disguising their human form [8]. In our pre-study, on the practical example of a FAQ chatbot for recruiting [9][10], a WOz experiment was conducted within a broader chatbot user testing scenario in order to:

- evaluate the intent database of the developed recruiting FAQ chatbot prototype in terms of relevancy and answer suitability,
- collect feedback on the conversational design and specifically the (1) preliminary content, (2) the perceived user satisfaction, (3) the user's level of acceptance, and (4) utilization limitations, and
- yield not yet considered but relevant content in the form of novel chatbot intents, as well as potential training data for the chatbot.

This work in progress will first shed light on the theoretical background of chatbot prototyping followed by a discussion of Wizard-of-Oz testing in general, as well as the current state of WOz testing applied for chatbots in Section 2. The third section deals with the study approach in terms of the overall goal and the strategic, as well as technical set up of the WOz testing environment and framework. In the fourth section, we present preliminary findings of our pre-test and implications for practice before presenting the study's limitations and conclusions in Section 5.

II. Theoretical Background

Iterative, user-centric design of chatbots is essential for good performance and to ensure the relevancy of the technology to the intended process of deployment. This section deals with the current state of chatbot development and the corresponding role of prototyping. Furthermore, it gives insights into the procedure of WOz experimenting and its application within chatbot development and research.

A. Chatbot Prototyping and Development

Chatbots are a kind of conversational interface [4] and belong to the field of Human-Computer Interaction (HCI) research [11]. The need to involve users in the development process becomes apparent as it is the human users who need to see the overall relevancy of the technological system and be able to utilize it appropriately in order for it to add value. As per common practice (e.g., [5][12]), user testings are integrated into the system design process as an essential development step. Overall, there are many requirements to consider when developing a chatbot (see [13] for a multi-perspective overview), such as an adequate and useful re-action to input, behavioral appropriateness, and friendliness. Unlike graphical user interfaces, chatbot development is more difficult to separate interaction with the system from system functionality. Also, for chatbots, clickable dialog flows can be created and visualized for testing (e.g., [14]). However, such prototypes do not react directly to text input and, therefore, strongly abstract from the later usage scenario. Thus, technically, development requires already some sort of a development platform, high levels of programming skills and development experience [15] in order to build a functional chatbot prototype to be tested in a real-world scenario. Contentwise, the intent and response database is essential and determines the quality of the chatbot in the form of response appropriateness [15]. Hence, the creation of a suitable, encompassing intent list with an accompanying set of matching, relevant responses as an adequate reaction is crucial within chatbot development (e.g., [16]). Conversational interfaces can be seen as a progression from visual layout and interaction design [11]. They serve as an interface allowing for a dialogue with human users based on natural language entered by text input. As such, they leave little room for front-end user interface design as text input is rather static and not very variable [7]. Hence, it is the content itself [11] and the way of communication (e.g., chatbot personality [4]) that is in focus in chatbot conversational designing.

B. Wizard-of-Oz Experiments for Technological Innovations

The term Wizard-of-Oz originates from a story in a children's book by [17], in which one of the protagonists hides behind a curtain to control a scene from a remote, through which he can pretend to be a powerful wizard. A Wizard-of-Oz experiment, as coined by [18], is thus a simulation where the researchers interact with the users themselves in a concealed way while posing as a fully functioning technology whereas, in reality, the technological system is in a prototypical, incomplete state [19][20]. WOz studies are conducted to let the participants believe that they interact with a computer system processing natural language dialogues whereas in reality, they are not: a human, called wizard in this kind of experiment, mediates the conversation in order to circumvent the constraints of current technology and thus pretending to showcase an operating, sophisticated kind of technology [8]. The method is not new [6] but still represents a practical, resource-saving way of early user testing within the development process since no full-fledged prototype needs to be built for yielding first feedback. However, due to the integration of a competent, skillful wizard, a WOz scenario does not depict a fully realistic representation of the examined technology and is somewhat idealized so that it cannot be treated as a holistic testing approach -– it rather gives first ideas to build upon [3]. Especially in early stages of prototyping with incomplete functionalities, WOz experiments are advantageous as they resemble realistic, human-like conversational behavior and capable dialogue management as opposed to existing, potentially erroneous systems [21].

WOz studies are integrated into various fields of research and add value to technology development projects of all kinds. Complex technology, such as systems integrating Artificial Intelligence (AI) functionalities are especially well suited for this approach. The following section examines WOz testing in the specific domain of chatbot development.

C. Wizard-of-Oz Experiments for Chatbot Development

Wizard-of-Oz setups are applicable to various systems and architectures for testing before actual implementation [3]. The advantages of WOz experiments, such as the early user feedback on the system to have it comply closely to all relevant user requirements and the savings in (especially technical) resources, can also be exploited within chatbot development. As conversational systems conversing with human users in natural language, chatbots oftentimes encompass AI functionalities and are thus especially suited for WOz tests during the development process: The AI components can be mimicked without the necessity of sophisticated AI framework implementation. Within chatbot development, there are various aspects to consider in terms of technical, content, and design requirements, as presented in Section II-A. Alongside these prerequisites, there are certain restrictions concerning the creation of conversational systems: Chatbots are bound to predefined databases and thus predetermined input, which makes WOz-based prototype tests relevant to cover unexpected and thus non-considered content [22]. This is an ideal setup to assess first user perceptions of the preliminary conversational design while also allowing for new content compilation, which is in line with [8], who states that WOz studies can be utilized to gather data. In this study, the WOz experiment yields relevant intents and accompanying training, as well as test data for the chatbot prototype at hand to be implemented in the chatbot prototype as a next step.

The interface itself is predetermined as well in the form of a certain social media channel or messaging application as most common access point for chatbots [10]. Hence, the WOz framework needs to be integrable into this environment and must fit in a way that it cannot be distinguished from the expected fully developed chatbot. The WOz approach has commonly been applied to chatbot research (e.g., [20][21][23]–[25]). In the focused field area of chatbots for human resources, a few first studies exist as well (e.g., [26][27]). However, no study is known to the authors providing more detailed insights on the WOz framework and its implementation, as well as the findings generated for the user-centered improvement of a chatbot concept. This study seeks to close this gap.

III. Methodological Approach

The study at hand focuses on WOz testing for the simulation of an advanced chatbot. The chatbot is applied to the use case of answering FAQs on different topics and process steps within an electronic (i.e., web-based) recruiting process. In this section, the methodology of the study, including its goals, the chatbot concept, and the WOz framework, are presented.

A. Goals of the Wizard-of-Oz Study

There are three overarching goals of this ongoing study:

1) Intent matching and answer suitability assessment: As introduced, chatbot concepts can be simulated in a WOz testing environment. To get a real user feedback, the reactions of the wizard must reflect not only the functions but also the limitations of the intended chatbot. The wizard, therefore, does not answer freely but must follow predefined rules and settings. In our case, we did use the underlying content in terms of an initial intent set and corresponding predefined answer phrases developed during a previous project work on which the prototype is based on. Via the experiment, we tried to evaluate for which user inquiries the wizard could match an existing user intent to answer the user request, in which cases the wizard had to modify the answers, or no predefined intent was found at all, and thus a response had to be formulated based on the wizard's expertise. All in all, the completeness and suitability of our initial intent set should be assessed.

2) Conversational design evaluation: In addition, after setting up the first version of our recruiting FAQ chatbot, its (1) content, and (2) the experience with the chatbot in the specific application area of recruiting FAQ – assessed via the user's satisfaction and perceived usability, as well as the performance of the demonstrated solution – are studied by gathering user feedback via a qualitative (thinking aloud) as well as a quantitative (user survey) approach.

3) Intent generation: Besides testing of the topics already implemented in our recruiting FAQ chatbot concept, further information needs, and corresponding user intents need to be identified and integrated into the intent set. For acceptance reasons, this set must be extended to a point so that the chatbot provides relevant answers for the most prevalent questions. Apart from intent generation, potential training data can be derived from the WOz testing by the integrated collection and assignment of user input phrases to intents. However, to gain relevant amounts of data, this would require a larger scale of testing than in this pre-study. Furthermore, such use of WOz experiments might get more important and productive in later phases when the chatbot solution is implemented and needs to be trained. Training is necessary as the natural language understanding and intent matching components of advanced chatbots are based on (pre-trained) machine learning algorithms and thus rely on domain-specific training data to evolve and improve [28].

The WOz approach is utilized to test and validate the recruiting FAQ chatbot prototype from the corresponding perspectives as presented above. Based on the findings, the chatbot will be iteratively adapted, enhanced, and further developed.

B. *Chatbot Composition and Configuration*

This pre-study is part of the research project CATS (Chatbots in Applicant Tracking Systems, for further information see acknowledgment section) that focuses on the identification of value-adding chatbot use cases and implementation of chatbot functionalities in applicant tracking systems. The general relevance of the specific use case of a FAQ chatbot to support applicants and answer questions in the recruiting process was already the subject of previous research [29].

In order to satisfy the needs of the target group, intents were collected from different sources: (1) potential candidates on the verge of applying to a job were asked to walk through an application process in an applicant tracking system and to formulate questions on problems and challenges, (2) questions and answers in existing FAQs on websites on career websites and job portals were screened and consolidated, and (3) information inquiries from other channels (e.g., e-mail requests to employers with job offers) were collected. Moreover, recruiting experts were involved in reviewing and improving the resulting set of intents, suitable answers, and an initial set of example user questions (to make the intents easier to understand and as initial training data). In total, 113 intents have been identified to be included in the FAQ recruiting chatbot concept. This intent set with the accompanying answers has been utilized as the wizard's database throughout the experiment.

C. *Study Design*

The study at hand was designed to comply with the goals as defined in Section III-A: Intent matching in the form of answer fitness assessment, conversational design evaluation, and intent generation. The experiment resp. study design consists of four sub-sequential parts, as presented in Figure 1, and is described in the following paragraphs.

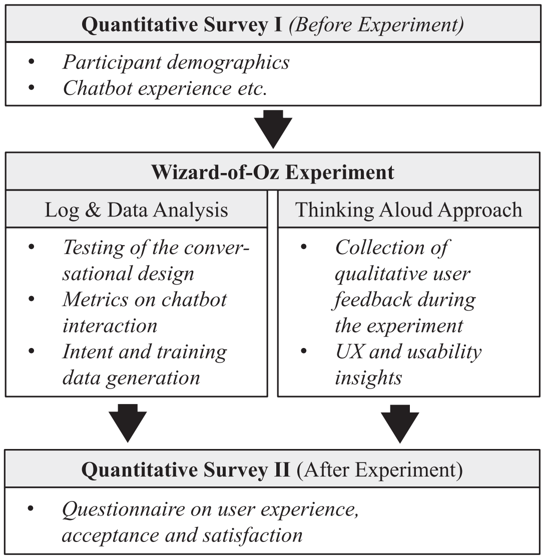
Quantitative Survey I (Before Experiment)
• Participant demographics
• Chatbot experience etc.
Wizard-of-Oz Experiment
Log & Data Analysis
• Testing of the conversational design
• Metrics on chatbot interaction
• Intent and training data generation
Thinking Aloud Approach
• Collection of qualitative user feedback during the experiment
• UX and usability insights
Quantitative Survey II (After Experiment)
• Questionnaire on user experience, acceptance and satisfaction

Figure 1: Flowchart on the Wizard-of-Oz Study Approach

1) Quantitative Survey I: Prior to the chatbot experiment, some socio-demographic characteristics of the participants (e.g., study program and qualification as filter-questions to confirm a required fit with the made-up job ads prepared for the study), as well as their experience with (recruiting) chatbots were surveyed.

2) Log and Data Analysis: At this stage, the WOz experiment started and the participants (students) were asked to apply for one of three pre-chosen open positions presented by job advertisements. The job ads were selected based on the qualification and skill profiles defined for the acquisition of study participants. As per the digital job advertisement landing page, the participants were free to consult the chatbot for any upcoming question or insecurity during their information and application phase up to the final application step of document and information submission. Even though the participants could have applied with their own documents due to their qualifications, application documents were provided for data protection reasons. All interactions were logged for the later data analysis.

3) Thinking Aloud Approach: The participants were asked to conduct a chatbot-supported application process in a thinking aloud approach, thus stating their thoughts, irritations, opinions, and actions while performing the task. The thinking aloud approach helps to gain relevant user experience (UX) insights and is a standard tool within user experience research [30]. Qualitative results are highly valuable for assumption and opinion validation and exploration of usability aspects [31].

4) Quantitative Survey II: After the participants have completed the WOz experiment and successfully submitted their application to the system, they were asked to answer a quantitative survey focussing on their satisfaction with the chatbot support and the corresponding user experience.

A moderator accompanied the participants through this process (on-site or remote) while one of the researchers posed as the wizard in the WOz framework; the details will be discussed in the following description of the WOz experiment setup.

D. Setup of the Wizard-of-Oz Experiment

Depending on the technological system, the WOz experiment concept needs to be integrated in a way that the wizard can operate covertly, which can be problematic for some setups [6]. However, the users must be led to believe that they are interacting with the technology itself for the WOz experiment to become successful and measuring the intended aspects. In the following, the WOz testing strategy and setup will be explained from conceptual, as well as from the technical perspective.

1) Wizard-of-Oz Experiment Concept: Maulsby et al. [31], who conducted a study on WOz testing with an automation agent, stress the importance of a strict behavioral plan for the wizard (they even recommend implementing an algorithm). This is important to maintain consistent behavior and, thus, experimental reliability [31]. In the four parts of the experiment as presented in Section III-C, several components had to be conceptualized: Required (1) roles, (2) documents, and (3) sequences.

In general, the following roles were assigned:

- *Participant:* The recruited chatbot users belonging to the target group of potential candidates, who converse with the chatbot during their application process.
- *Wizard:* A researcher belonging to the research project, who operates the WOz framework by sending preformulated messages or creating ad-hoc responses as seemingly AI-based automated answers from a separate room/on remote based on the experimental study framework.
- *Moderator:* Another researcher also belonging to the research project, who accompanies the participant through the experiment giving an introduction, instructions, and guidance through the process.

The participants were provided a set of application documents (CV, internship certificate, master's certificate) to allow for a realistic application scenario while maintaining privacy and data protection requirements. The moderator guided the participants through the whole process and was also responsible for writing down the participants' answers to the introductory and the conclusive quantitative questionnaires himself for a consistent moderator-participant experience. The first quantitative questionnaire was conducted after the moderator's introduction in the form of a brief explanation of the experiment and the according procedure and prior to chatbot utilization for first participant classification concerning their demographics. It consisted of five questions regarding their professional situation, their study program as well as their experience with online applications, chatbots, and recruiting chatbots in specific.

In the main part of the WOz experiment, the participants accessed a job search portal with three predefined job ads; they had to choose between. Upon making a choice, they were able to make any kind of inquiry to the alleged chatbot prototype, positioned in an embedded chat window in the lower right-hand side of the job ad landing page. They had to gather all information they presumed necessary for taking up an application and then actually apply via a specially configured testing application platform provided by the cooperating industry partner of the authors. During this process, the participants were once again told to make use of the chatbot whenever it felt necessary in situations of upcoming questions. The utilization phase ended after information and document upload upon submission of the application. Eight checkpoints had been established for further encouragement of chatbot utilization in the form of active requests to formulate every possible question coming to mind, but this method proved unsuccessful in the initial experiments and was perceived as rather interrupting concerning the overall procedure. For this reason, the checkpoints were removed from the study design, and feedback collected this way was not further considered in the study.

Throughout the phase of chatbot use and application in the system, qualitative user feedback was yielded via a thinking aloud approach. The participants were encouraged to articulate all upcoming thoughts, perceptions, and feelings towards the chatbot and their interaction with it. The quantitative survey after the WOz experiment, contained ten UX items (concerning the interaction via the interface not focusing on the design), questions concerning the participants' satisfaction with the answer quality (completeness, competency, and speed), the perceived added value from the chatbot support in general, as well as for each application step of the application process. The quantitative survey concluded with questions on the perceived (dis-)advantages concerning (1) any previous recruiting chatbot usage and (2) the FAQ chatbot prototype presented in the WOz experiment.

2) Technical Infrastructure for the Wizard-of-Oz Study: According to [3], the only components necessary for WOz testing are the interface software and databases. Correspondingly, the WOz framework was technically set up via Rocket.Chat [32], a free open source chat platform allowing for back-end and front-end chat interfaces for the wizard and the experiment participant. Moreover, this chat server system provided functions for storing data, i.e., logging the messages with additional information for later analysis (e.g., time-stamps). Rocket.Chat as chat server was chosen as it comes as an installation option with the Ubuntu server operating system and due to its simple handling and configuration without the need for advanced programming expertise. The Live Chat feature of the platform was utilized as a communication channel for the participants. The chat server was installed on a dedicated Ubuntu server.

Apache webserver was installed and configured for setting up websites required for the study. By using a JavaScript code snippet as advised by Rocket.Chat, a chat window, was integrated into an HTML document, which was then accessible by the participants via a web browser. The HTML document was also used to embed an IFRAME with a job search platform presenting the job ads. The job ads were linked to made-up career websites with access to a test installation of the applicant tracking system BeeSite [33] (operated on the servers of the cooperation partner Milch & Zucker AG) where the participants entered their data and completed the application process.

The participant's front-end configuration is presented in Figure 2. While operating in the career portal as shown on the left-hand side, the chatbot was accessible throughout the whole process as an overlay in the lower right corner (depicted on the right-hand side). This way, all upcoming problems in the form of questions or irritations could be directed to the chatbot from the participants. Messages sent by the respondents in the chat window were sent to and stored in Rocket.Chat. The chats can be accessed via a certain interface and saved as JSON files. Via JavaScript, the JSON data, were then converted into CSV data for further analysis and handling via Microsoft Excel. The researcher acting as wizard utilized the Rocket.Chat administration interface to receive and process incoming inquiries while posing as a chatbot.

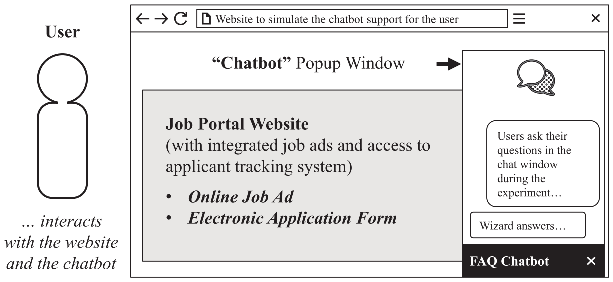
User
... interacts with the website and the chatbot
Website to simulate the chatbot support for the user
"Chatbot" Popup Window
Job Portal Website
(with integrated job ads and access to applicant tracking system)
• Online Job Ad
• Electronic Application Form
Users ask their questions in the chat window during the experiment…
Wizard answers…
FAQ Chatbot

Figure 2: Wizard-of-Oz User Front-end Configuration

As shown in Figure 3, a special cockpit was designed within a web application for the wizard to either (1) choose from the predefined answer related to a specific intent considered in the predefined intent set, (2) take a predefined answer and modify it according to the unexpected input, or (3) enter answers in real-time to create novel, individual content for distribution to the participant.

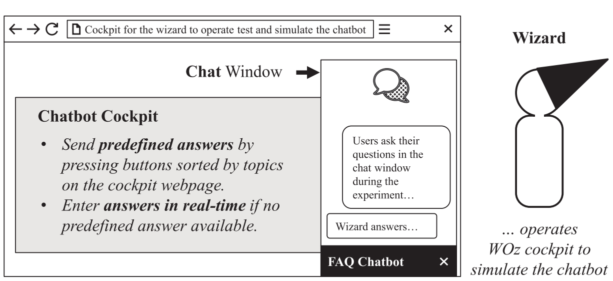
Cockpit for the wizard to operate test and simulate the chatbot
Chat Window
Chatbot Cockpit
• Send predefined answers by pressing buttons sorted by topics on the cockpit webpage.
• Enter answers in real-time if no predefined answer available.
Users ask their questions in the chat window during the experiment…
Wizard answers…
FAQ Chatbot
Wizard
... operates WOz cockpit to simulate the chatbot

Figure 3: Wizard-of-Oz Wizard Front-end Configuration

Figure 4 shows the framework as procedure embedded into the overall study design, including the different roles, documents, and processes. With the servers hosting the Rocket.Chat chat environment and the career portal as central parts, the users accessed the framework from front-end perspective (left-hand side) while the wizard operated in secret from the back-end perspective imitating the expected FAQ recruiting chatbot (right-hand side).

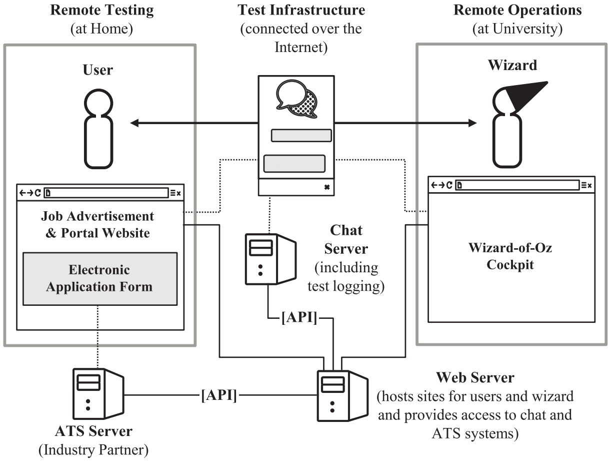
Remote Testing (at Home)
Test Infrastructure (connected over the Internet)
Remote Operations (at University)
User
Wizard
Job Advertisement & Portal Website
Electronic Application Form
Chat Server (including test logging)
Wizard-of-Oz Cockpit
[API]
ATS Server (Industry Partner)
[API]
Web Server (hosts sites for users and wizard and provides access to chat and ATS systems)

Figure 4: Wizard-of-Oz Framework of the Study

During the study, the setup had to be revised because of the outbreak of COVID-19. As a result, after an initial test with a first participant, the whole technical infrastructure had to be moved from a physical server of the laboratory in intranet (protected from access from the public internet by restrictive firewalls) of the university to (cloud) hosting providers in the public internet to allow all stakeholders in the form of the participants, the moderator, and the wizard to access the necessary interfaces remotely (without VPN access). For the moderator in the WOz experiment, who accompanied the experiment in a room together with the participants in presence mode, an adequate alternative had to be found to be able to perform this part of the experiment remotely. As a solution, Lookback [34] was identified. With this product, various user tests can be easily moderated and remotely performed. By means of Lookback, it was possible in the present experiment to guide the test participants to the already established test site (career portal incl. chat) and to accompany them during use. By integrating a video solution, the test participant and the moderator could stay in contact during the experiment.

IV. Preliminary Findings and Implications

A. *Metrics on Chatbot Interaction*

In total, eight users actively participated in the WOz experiments. One user did not use the chatbot as he did not required or considered support within the application process and was thus excluded from the analysis on the chatbot interaction in this and the next chapter. Another participant had to be excluded from this section analyzing the metrics, as changes in the setup of the WOz environment were required, as described in the previous section.

The remaining six participants interacted with the wizard in 79 chatbot sessions (cf. Table I). A session describes here a coherent sequence of interactions between chatbot (i.e., the wizard) and the user associated with a single user intent. The ratio of chatbot interactions per session (column c) varied between 1.00 and 1.43, with a mean value of 1.21. This evaluation shows, first of all that the activation for use and intensity of use varied greatly among the participants in the study. Moreover, it becomes evident that some respondents expected to get a prompt answer (comparable to a search request on a website) where others got more involved in an interactive dialog with the chatbot to get the intended information.

The wizard's response behavior in the experiment is also shown in Table I in columns (d) to (f). As described earlier, the wizard had three different response options to user queries in the experiment. More than half of all answers by the wizard (55; 63 percent) were given by predefined answers via the button option in the wizard cockpit (d), only in four cases (5 percent) the predefined answers were modified by the wizard (e). For about one-third of the user requests (28), there was no matching intent, and so the answer had to be formulated by the wizard (f). In a productive mode with a chatbot implemented based on the given concept, the questions with no matching intent could not have been answered. Two participants took advantage of the opportunity to be forwarded by the chatbot to a human contact person to answer a question, but only once each (human hand-over (g)).

TABLE I: Wizard-of-Oz Experiment Metrics (Absolute Values).

(#)	(a) Chat-bot ses-sions	(b) Chat-bot inter-actions	(c) Inter-actions per session	(d) Wizard answer via button	(e) Wizard answer edited	(f) Wizard answer free	(g) Human hand-over request
(1)	16	21	1.31	10	1	10	1
(2)	7	7	1.00	4	0	3	n.a.
(3)	12	17	1.42	6	0	8	1
(4)	13	14	1.08	8	2	2	n.a.
(5)	23	33	1.43	20	1	4	n.a.
(6)	8	8	1.00	7	0	1	n.a.
Sums	79	100	–	55	4	28	2
Means	13.2	16.7	1.21	9.2	0.7	4.7	0.3

As a next usage metric, the average response times required by the Wizard were recorded in this stage of the experiment (see Table II). Not surprisingly, the average response times of the wizard were lowest for the predefined answer buttons (column a), at an average of 20 seconds. Here, the wizard had to capture an incoming user request, then search the chatbot cockpit with keywords to find a matching intent/answer pair in the list and send the corresponding answer to the chat. The wizard in the experiment took noticeably longer (34 seconds on average) for the answer option (column b), where a slight adjustment of the predefined answers available in the intent/action list was made. Only marginally shorter, average response times were achieved for the option of free answers written by the wizard. Here, an average of 32 seconds passed between receiving the user inquiry and posting the answer in the chat (column c). Across the various response options and the participants in the experiment, the wizard took an average of 25 seconds to answer an user inquiry.

Overall, it can also be recognized that the response times were significantly longer than it could be expected from an automated answering system. However, the participants in the study were briefed in such a way that it is a test system with yet limited performance.

TABLE II: Wizard-of-Oz Experiment Mean Answer Times (In Seconds)

(#)	(a) Wizard answer via button	(b) Wizard answer edited	(c) Wizard answer free	(d) Overall
(1)	19	27	22	21
(2)	16	n.a.	23	19
(3)	26	n.a.	45	37
(4)	20	44	27	25
(5)	18	32	36	21
(6)	23	n.a.	39	25
Means	20	34	32	25

B. Quantitative User Experience Survey

After the experiment, the participants were asked for quantitative feedback in the form of a short user survey, e.g., with regard to satisfaction ratings on selected topics in the field of user experience.

Table III shows a summary of the result of this survey. The table shows the survey results for the seven participants that interacted with the simulated chatbot. It can be seen that the satisfaction with regard to the answer completeness is rather indifferent and moderate. Three of the seven participants were rather satisfied, another three partly satisfied, and even one not satisfied at all. The quality of the answers and thus the perceived competence of the chatbot is another important evaluation criteria in the WOz experiment. As shown in Table III, two of the seven participants stated that they considered the answers they received as rather competent. The remaining five participants were moderately satisfied only. Not surprisingly, the satisfaction rating with the chatbot performance, i.e., the speed of the chatbot answers, turns out to be recognizably poor, which is of course also due to the character of the WOz project: Since the chatbot is only simulated by a human, the person needs time to record and process the questions and to write the answers. This finding does not seem to have influenced the perception of the general added value of chatbots. This is probably due to the fact that the test persons were aware of the test situation and performance limitations. Six of the seven participants consider the tested use case as relevant and the general added value in applicant support of such a FAQ chatbot as (very) high.

TABLE III: User Experience Evaluation of the Chatbot Prototype

(Absolute Values; N = 7)	com-pletely satisfied	rather satisfied	moder-ately satisfied	rather not satisfied	not at all satisfied
Answer Completeness	0	3	3	1	0
Competence	0	2	5	0	0
Speed and Performance	0	1	1	2	3
	very high	rather high	moderate	rather low	very low
General Added Value	2	4	0	1	0

Beyond the general added value, the test persons were also asked to evaluate the added value of the presented FAQ chatbot in the individual process phases of the application. As shown in Figure 5, the three areas to which the participants attribute the greatest added value are answering questions about the job advertisement, questions about registration and general questions on the application process, and the further procedure after the application.

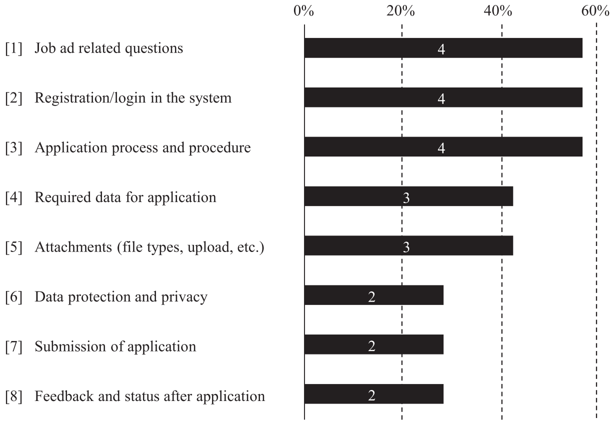

Figure 5: User Assessment on the Added Value of the FAQ Recruiting Chatbot Prototype (Multiple Answers Possible).

The last part of the quantitative survey contained questions on the user experience and usability of the chatbot as perceived by the participants in the WOz experiment. As shown in Figure 6, all participants perceived the FAQ chatbot simulated in the WOz scenario as easy to use. Furthermore, all but one participant agreed that potential applicants quickly learn to use such a chatbot. Four of the seven participants could still imagine using such a FAQ chatbot on a regular basis, following the example of the one used in the experiment during the application process. The other answers regarding the perceived technical complexity, quick learnability, or data security show that chatbot systems like the one presented can be quickly adopted and easily mastered. However, feedback on the integration of the chatbot and inconsistencies in the answers indicate a potential for improvement.

C. Qualitative User Feedback

In the following, some important observations and findings from the thinking aloud approach will be summarized. The most apparent problem was the long latency times caused by the human wizard simulating the chatbot. However, only one of the participants suspected that the delay might result from a human acting as a counterpart in this experiment. Due to the delay, some participants started to adapt their asking behavior by reformulating inquiries or reducing the number of questions to prevent further waiting frustration. The response times in the WOz experiment are, therefore, clearly too long. In future experiments, response times must be reduced. This could be done by optimizing the wizard cockpit, integrating a recommender system to pick answers (instead of searching for answers in the cockpit), or the integration of language-to-text interface to avoid typing in text for free answers. It should be noted, however, that this finding is more a problem of simulation than of the actual chatbot concept.

A more substantive observation concerns the complexity of the questions. The solution intended for implementation of the chatbot does not support the identification of multiple intents in a single user prompt. For a realistic scenario, the chatbot did answer questions one at a time. Ignoring question portions led to misunderstandings and confusion among individual participants. In a later implementation, solutions must be found to identify such problems and provide users with appropriate feedback to simplify questions. Another frequent remark was the perceived superficiality of several chatbot answers, which overall did not satisfy the users but rather frustrated them and was perceived as inept in some instances. This indicates a need for improving the intent set used in the experiment, as well as the corresponding answers. For chatbot implementation, response quality and relevance to a user might be improved by integrating the usage context. For example, responses could be personalized by processing information about the applicant already entered in the applicant management system. It has also become clear that a chatbot system must be able to distinguish between requests that can be answered with standardized information or with advice from a human contact person. One participant suggested human hand-overs for important questions and leaving the chatbot for rather simple inquiries. As discussed in the previous section, however, it is not to be expected and occurs rather rarely that the users of a chatbot themselves request such an offered option for a hand-over to second-level support.

Although there was little to criticize by the users regarding the usability of the chatbot, and there is little scope for design, individual possibilities for improvement were identified. A typing indicator was not implemented in the WOz front-end and was reported missing by the participants. Such an indicator can show that the request is being processed on the other side and shall be included in the future version of the WOz setup, as well as in the real chatbot if significant processing time would occur. There were also several helpful remarks regarding the positioning of the chatbot: Some participants felt

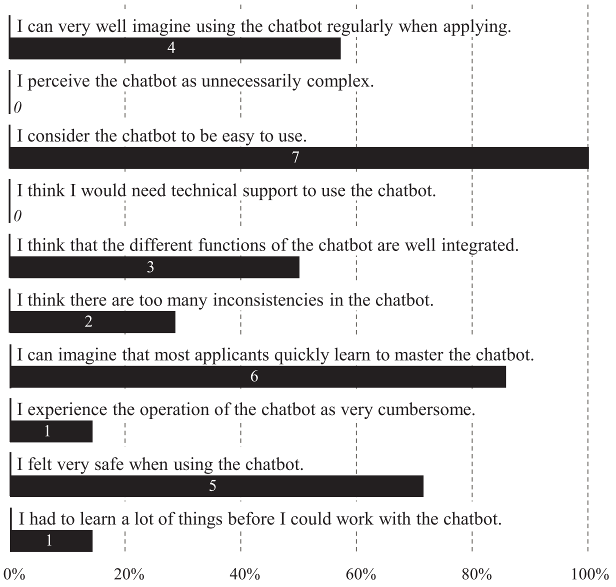

Figure 6: User Experience Rating of the FAQ Recruiting Chatbot Prototype (Sum of Response Options ”Totally Agree” and ”Rather Agree”).

it was partly hidden behind the banner informing about the website's cookies, and one participant did not recognize the chatbot button when retracted into the small starting button at the beginning of the experiment.

In total, it can be said that from candidate-sided user perspective, there are several aspects in the WOz framework that need enhancing and further development prior to continuation of the experiment, as well as consideration in the envisioned chatbot prototype. Consistent with Maulsby et al. [31], the authors learned a lot about required improvements of the chatbot concept by posing as wizards during the experiment. That way, not only the feedback of the participants can be integrated, but also the researchers' perspective can be considered through their role as back-end chatbot operators. About the content and scope of the chatbot, we learned that several topics and questions, for example, concerning the career portal, have not yet been considered and need to be included for a more comprehensive intent set of the envisioned chatbot.

V. Conclusion and Limitation

This study has demonstrated that WOz experiments can be used in the early phases of system development to validate concepts for chatbots. Appropriate infrastructures are to be implemented with a manageable amount of resources based on existing open-source web and chat server solutions. In such WOz setups, participants can be credibly convinced to interact with a real chatbot. However, the time needed to select appropriate answers is problematic if the restrictions of the chosen chatbot design are to be maintained in the experiment, and the wizard should not simply answer freehand. The experiment has also shown that users do not automatically accept support offered by a chatbot and do not necessarily enter into a more comprehensive dialogue with such a system.

The findings of the study indicate that the implementation of FAQ chatbots in application processes is seen by the participants as easy to master and valuable. However, it is important during implementation that the chatbot is actively promoted and indicated on the respective website. When interacting with a user, the chatbot must not only provide suitable answers for questions but also need to point out necessary simplifications in case of complex inquiries or even take the initiative to offer a hand-over to second level support by human experts.

WOz experiments can also provide important insights into the required content and scope of the chatbot concept. While most user questions could be handled by predefined answers from a given intent set that reflected the current status of the chatbot concept, more than 30 percent of the user inquiries in the experiment had to be answered freehand by the wizard showing a need to extend the intent set to the topics not covered yet. This is supported by the findings of the quantitative survey on the answer completeness that was not fully satisfying and thus needs to be improved. The perceived superficiality of the chatbot answers is another quality-related problem of the chatbot concept identified in the experiment that indicates further improvements of the intent/answers sets.

Certain limitations need to be taken into considerations: Our findings are based on a WOz study with a very small sample of eight participants only. However, in early phases of development and in studies focusing more on general feasibility and usability than generalizable results, small groups of test persons are quite common [35]. More critical, however, are the statements in our study about the added value or usefulness of the presented solution, which must definitely be verified by surveys with more participants. Another inaccuracy with regard to the implementation of the chatbot concept is if the intent matching performance of the wizard can be achieved by today's chatbot platforms available in the market. While the coverage of user inquiries and the responses of the chatbot were realistically limited by the intent set, intent matching may still vary considerably in a later implementation, which may influence user satisfaction as well. In general, for WOz experiments, maintaining consistent wizard behavior and the incapability to simulate errors or suboptimal system performance are limiting aspects of studies of this kind [6].

In future research, the authors can profit from the insights of this pre-study by optimizing the chatbot infrastructure or utilizing a hybrid approach, as suggested by [28]. Such a hybrid approach could be implemented, for example, by integrating a functional chatbot prototype into the WOz framework and limit the scope of human intervention to areas where the chatbot does not respond appropriately to inquiries. Other future studies might look into the field of speech-based dialogue systems in the form of voice assistants, predicted to be the even more efficiency enhancing and generally next logical step after establishment of text-based chatbot solutions [10].

Acknowledgment

The study was carried out as part of the research project CATS (Chatbots in Applicant Tracking Systems) of the Rhein-Main University of Applied Sciences. This project (HA project no. 642/18-65) is funded in the framework of Hessen Modell-Projekte, financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

References

[1] L. Schildknecht, J. Eißer, and S. Böhm, “Motivators and barriers of chatbot usage in recruiting: An empirical study on the job candidates' perspective in germany,” *Journal of E-Technology*, vol. 9, no. 4, pp. 109–123, Nov. 2018.

[2] U. Gnewuch, J. Feine, S. Morana, and A. Maedche, “Soziotechnische gestaltung von chatbots (socio-technical design of chatbots),” in *Cognitive Computing*, Springer Fachmedien Wiesbaden, 2020, pp. 169–189.

[3] D. Jurafsky and J. H. Martin, *Dialog systems and chatbots*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 2018.

[4] M. F. McTear, “The rise of the conversational interface: A new kid on the block?” In, J. F. Quesada, Francisco-Jesús, M. Mateos, and T. L. Soto, Eds., Berlin: Springer International Publishing, 2017, pp. 38–49.

[5] J. Nielsen, “The usability engineering life cycle,” *Computer*, vol. 25, no. 3, pp. 12–22, Mar. 1992.

[6] S. Schlögl, G. Doherty, and S. Luz, “Wizard of oz experimentation for language technology applications: Challenges and tools,” *Interacting with Computers*, vol. 27, no. 6, pp. 592–619, May 2014.

[7] S. Böhm *et al.*, “Intent identification and analysis for user-centered chatbot design: A case study on the example of recruiting chatbots in germany,” in *The Thirteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2020*, (in press), 2020.

[8] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies," in *Proceedings of the 1st international conference on Intelligent user interfaces - IUI '93*, ACM Press, 1993, pp. 193–200.

[9] B. Hmoud and V. Laszlo, "Will artificial intelligence take over human resources recruitment and selection," *Network Intelligence Studies*, vol. 7, no. 13, pp. 21–30, 2019.

[10] L. Dudler, "Wenn bots übernehmen – chatbots im recruiting (when bots take over – chatbots in recruiting)," in *Digitalisierung im Recruiting*, T. Verhoeven, Ed., Wiesbaden: Springer Gabler, 2020, pp. 101–111.

[11] A. Følstad and P. Bae Brandtzæg, "Chatbots and the new world of HCI," *Interactions*, vol. 24, no. 4, pp. 38–42, Jun. 2017.

[12] T. K. Landauer, *The Trouble with Computers*. Cambridge, Massachusetts: The MIT Press, 1996.

[13] N. Tavanapour and E. A. Bittner, "Automated facilitation for idea platforms: Design and evaluation of a chatbot prototype," *Thirty ninth International Conference on Information Systems*, pp. 1–9, 2018, San Francisco.

[14] Botsociety, *Design chatbots and voice experiences*, 2020. [Online]. Available: https://botsociety.io/ [retrieved: 07/16/2020].

[15] S. A. Abdul-Kader and D. J. Woods, "Survey on chatbot design techniques in speech conversation systems," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 72–80, 2015.

[16] S. Ghose and J. Joyti Barua, "Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor," in *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, IEEE, May 2013, pp. 1–5.

[17] F. L. Baum, *The Wonderful Wizard of Oz*. Chicago: George M. Hill, 1900.

[18] J. F. Kelley, "Wizard of oz (woz): A yellow brick journey," *Journal of Usability Studies*, vol. 13, no. 3, pp. 119–124, 2018.

[19] R. Eynon and C. Davies, "Supporting older adults in using technology for lifelong learning," *Proceedings of the 8th International Conferenceon Networked Learning*, pp. 66–73, 2012.

[20] J. Eißer and S. Böhm, "Hedonic motivation of chatbot usage: Wizard-of-oz study based on face analysis and user self-assessment," in *The Tenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2017*, 2017, pp. 59–66.

[21] L. El Asri *et al.*, "Frames: A corpus for adding memory to goal-oriented dialogue systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, 2017, pp. 207–219.

[22] F. Guerin, "Learning like a baby: A survey of artificial intelligence approaches," *The Knowledge Engineering Review*, vol. 26, no. 2, pp. 209–236, May 2011.

[23] W. R. Kearns *et al.*, "A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems: 1-9*, ACM, Apr. 2020.

[24] L. 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, Aug. 2012.

[25] S. Quarteroni and S. Manandhar, "A chatbot-based interactive question answering system," in *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, 2007, pp. 83–90.

[26] M. X. Zhou, C. Wang, G. Mark, H. Yang, and K. Xu, "Building real-world chatbot interviewers: Lessons from a wizard-of-oz field study," in *Joint Proceedings of the ACM IUI 2019 Workshops*, 2019, pp. 1–6.

[27] R. Kocielnik, D. Avrahami, J. Marlow, D. Lu, and G. Hsieh, "Designing for workplace reflection," in *Proceedings of the 2018 Designing Interactive Systems Conference*, ACM Press, 2018, pp. 881–894.

[28] J.-W. Ahn *et al.*, "Wizard's apprentice: Testing of an advanced conversational intelligent tutor," in *Tutoring and Intelligent Tutoring Systems*. Nova Science Publishing, 2018, ch. 12, pp. 321–340.

[29] S. Meurer, S. Böhm, and J. Eißer, "Chatbots in applicant tracking systems: Preliminary findings on application scenarios and a functional prototype," in *In Böhm, S., and Suntrayuth, S. (Eds.): Proceedings of the Third International Workshop on Entrepreneurship in Electronic and Mobile Business*, (in press), 2019, pp. 209–232.

[30] W. Albert and T. Tullis, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Waltham, Massachusetts: Morgan Kaufmann, 2013.

[31] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through wizard of oz," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, 1993, pp. 277–284.

[32] Rocket.Chat, *The ultimate communication hub*, 2020. [Online]. Available: https://rocket.chat/ [retrieved: 07/16/2020].

[33] M. Z̃. AG, *Beesite recruiting edition – job posting applicant management talent pools*, 2020. [Online]. Available: https://www.milchundzucker.com/products/beesite-recruiting-edition-job-posting-applicant-management-talent-pools/ [retrieved: 07/16/2020].

[34] Lookback, *Talk to your users: See how they're using your app or website.* 2020. [Online]. Available: https://lookback.io/ [retrieved: 07/16/2020].

[35] J. Nielsen, *How many test users in a usability study?* 2012. [Online]. Available: https://www.nngroup.com/articles/how-many-test-users/ [retrieved: 07/16/2020].

ISBN: 978-1-61208-829-7