Proceedings of the ASME 2025 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC2025
August 17–20, 2025, Anaheim, CA

IDETC2025-XXXX

LEVERAGING THE WIZARD OF OZ METHOD FOR STIMULATING DESIGN: BENEFITS AND CHALLENGES FROM AN APPLICATION CASE STUDY

Gaëlle Baudoux[1,*], Guillaume Gronier[2]

[1]University of California, Berkeley
Berkeley, CA
[2]Luxembourg Institute of Science and Technology
Luxembourg, Grand Duchy of Luxembourg

ABSTRACT

This paper presents a comprehensive understanding of the use of the Wizard of Oz (WOz) method for studying emerging design tools. This method involves hidden human "wizards" simulating the functionalities of an undeveloped technology without the user's knowledge, making it a valuable method for assessing technologies that are not yet mature or are too resource-demanding. Although WOz has been widely used in fields like human-robot interaction and autonomous vehicles, its application in design research is relatively novel. Through a case study aimed at stimulating design ideas generation, we applied the WOz method to simulate an AI design assistant based on designer's sketches during 17 one-and-a-half-hour design sessions. We evaluated the success of this WOz simulation, and the results showed exemplary performance by the wizards in both time management and content creation, along with a strongly positive designer perception of the simulation, and a rich data collection to inform research analysis. The WOz method offers two major advantages: versatility across domains and tasks, and cost-effectiveness. However, it also presents challenges, including the workload on wizards and ethical concerns. We conclude with three key axes of recommendations for implementing WOz in design settings: training wizards for feasibility, clearly defining simulated functionalities and experimental protocols, and ensuring system credibility. These insights will support future design research anticipating technological advancements to study emerging tools.

Keywords: Methods, Simulation, Wizard of Oz, Testing and evaluation, Design stimuli

1. INTRODUCTION

When studying the potential benefits and pitfalls of emergent design tools, researchers face the problem that technologies may not yet be mature or may be too resource-demanding. The Wizard of Oz (WOz) method is an experimental procedure for evaluating the use, impact on users, and human-machine interaction of an emerging technology prior to its development [1–3]. This method involves simulating the functionalities of the studied technology by replacing them with equivalent, hidden human (also called wizards) work, in such a way that the observed subjects believe they are using the said technology. This method is a fundamental pillar of technological innovation, with its origins dating back to the 1970s [4]. However, when this method is documented in academic literature, it is often briefly referenced in the method section of a paper that primarily focuses on the result achieved in the study, resulting in the implementation of the WOz method being rarely subjected to extensive detail or post-application reflection. Furthermore, while the method is widely used in certain domains of Human-Computer Interaction (HCI), it is rarely employed or documented in other technology-related domains, such as engineering and design.

In our research, we mobilize this simulation method to determine the benefits and pitfalls of visual stimuli for early-stage design. We tasked 17 professional architects with sketching their design of a residential building while being shown automatically generated (through WOz) cleaned-up 2D and 3D representations of their design, and appropriate inspiration images. Deriving from this case study application, we provide a critical analysis of the WOz method. The primary contribution of this paper is to:

- Present an original implementation of the WOz method to study design and the data to which it gave access.
- Identify the benefits and challenges of this WOz method.
- Provide recommendations for implementation best-practices in design research.

The paper begins with a review of relevant applications of

*Corresponding author: gbaudoux@berkeley.edu
Documentation for `asmeconf.cls`: Version 1.40, March 24, 2025.

the WOz method across various research fields (Sec. 2). We then describe the implementation of the WOz method in our case study, including the data collected and the approach used to assess the success of the implementation (Sec. 3). Subsequently, we present the results of the WOz method's evaluation (Sec. 4). Finally, we discuss the method's benefits and limitations and provide recommendations for a successful application in design research (Sec. 5).

2. BACKGROUND

This section starts by examining the main fields in which WOz is typically used with illustrative implementations, before reviewing in further detail its novel application in design research.

2.1. Main fields of WOz application

In its first applications, the WOz method was used to develop human-machine interfaces before the technological infrastructure was complete. Dow et al. [5] used WOz as a lightweight prototyping method, presenting users with sketches of new interface ideas, without any finished technological support, with the aim of evaluating existing interaction theories in an experimental approach. During the first experiments using WOz, it was mainly voice or multimodal commands that were simulated in HCI [6, 7]. In the following sections, we present several recent examples of WOz applications, depicting the primary fields where this method is commonly used nowadays, and how it is usually implemented.

Connected and Autonomous Vehicles (CAV). In Weschke et al. study used a WOz [8] to evaluate the passengers' perception of autonomous buses. A bus driver was hidden from passengers' view in a modified vehicle and received training in driving like autonomous vehicles, including driving smoothly, staying in lanes, and making few lane changes. In another study, Ranjbar et al. [9] tested how people with auditory and visual impairments perceive and accept autonomous vehicles. They placed several subjects in a car with a non-interactive driver. During the ride, the driver triggered vibrations on a device to inform passengers of things like arrival at destination. In this study, for ethical reasons, vulnerable users were informed about the WOz method in advance, despite the potential experimental bias.

Virtual and Augmented Reality (VR/AR). WOz is used in some VR/AR studies to test solutions before IT development. Vithanage et al. [10] explored a VR solution for sign language communication. The virtual assistant, wearing VR goggles, imitated a real person's sign language while the wizard interpreted the assistant's signs. Helgert et al. [11] used WOz to control a robot in a virtual environment. A control system was created for the virtual robot to respond naturally and quickly to the headset's commands. In another study, Sabeti et al. [12] equipped road construction workers with AR glasses to warn them of potential dangers, with the aim of increasing safety and improving reactions. These warnings were triggered by WOz method.

Human-Robot Interaction (HRI). WOz is also used in human-robot interaction, to study how people react to robots. Rodriguez-Dominguez et al. [13] investigated the use of robots in neurotherapy to stimulate patients with mild to moderate impairments. Patients interacted verbally with a robot, whose responses were generated remotely by a cognitive therapist. Therapists used pre-recorded phrases to start new dialogues with patients about their physical activity or diet. Sienkiewicz et al. [14] did a study of domestic robots, where participants had to teach robots to arrange objects in a room. The robot's learning behavior was simulated using the WOz method. The robot seemed to observe the participant, was turning its head and responding with *"I see"* and *"I understand."* Participants then rated their perception of the robot's intelligence. Jang, Park, and Bae [15] used WOz to recommend robot design for TV viewers. The simulated robot was wearing a TV screen and was controlled by an operator in a control room to take viewer requests and positions into account.

Artificial Intelligence (AI). The WOz method is useful for testing the user's perception of AI contributions in the context of large language models. In Vrins' study [16], subjects were asked to brainstorm ways to improve university students' living conditions with a robot. The robot's responses were fed by either GPT-4 or a human wizard in an observation room. The researchers then evaluated how people perceived the robot's contributions in these two conditions. Another study by Aljuneidi et al. [17] measured how fair citizens perceived an AI-managed context to be. During an ID renewal process, subjects encountered different price points based on "AI" judgments and explanations. Perceptions of fairness were gauged based on the explanations generated by a wizard controlling their sequencing.

This review highlights that the WOz method is particularly relevant in human-machine interactions in three situations:

- When the product is at an early stage of the design process in the form of a prototype. The focus is therefore on validating certain functionalities that have either not yet been developed, or only partially. The goal of the WOz implementation is to present potential users with a minimum viable simulated product.
- When the product has a low level of maturity at the experimental proof-of-concept stage (TRL 3). The aim is thus to carry out an analytical or experimental evaluation of the main characteristics of a concept.
- When the technology is expensive to develop but researchers want to study its impact on individuals or society.

2.2. WOz for studying design stimuli

To examine the application of the WOz method within the context of our research area, we conducted a systematic literature review following the PRISMA 2020 guidelines [18, 19]. We searched Google Scholar and Scopus databases for studies published between 1990 and 2024 included and presenting the following search terms: "stimul* AND design AND (Wizard of Oz OR WOz OR WoO) AND (creativ* OR idea*)". Of 22,300 records, we independently (2 researchers) reviewed the title, abstract, and keywords of the 270 most pertinent papers, as sorted by the database. We retained 31 studies as eligible. Among the various sub-fields of design represented in these studies, including graphic interface design, software design, product design,

creative activities, engineering, and architectural design, we narrowed the selection down to the papers most closely aligned with our own research area, i.e. engineering and architectural design. After reading these papers in their entirety, five were found to be studies that employed WOz protocols to stimulate creative, engineering, or architectural design activities (Table 1).

Throughout these studies, the WOz method is used in various design tasks. Idea generation is stimulated by an acting "AI" or "robot" with one or more of the following inputs: verbal prompts, inspiring visuals, keywords, and design information. Different types of stimuli, some visual, some verbal, are used with varying links to design subtopics: general calls for ideas, preset stimuli, and adaptations to live discussions. To interact between designers and wizards, all studies use verbal and visual modalities, with one even using gestural modalities.

While these studies present some shared set-ups, each of them presents distinct approaches and outcomes that help position our contribution. Khan et al. [20] and Xu et al. [21] both utilize external stimuli—images and videos or a preset gallery—to encourage idea generation, but these techniques primarily serve to prompt creativity rather than actively engage with the design process in a dynamic, evolving way. In contrast, Fu and Zhou [22] involve the WOz method more interactively, with a wizard offering verbal and visual input to directly influence participants' sketches. However, their approach remains prescriptive, with the AI offering predefined guidance rather than co-creating or reacting in real-time to the designers' evolving concepts. Similarly, Kim et al. [23] and Poser et al. [24] focus on providing external input (e.g., light settings or conversational feedback) but do so in more constrained or passive environments, such as VR or scripted interactions, limiting the potential for deep, continuous collaboration.

In contrast, our study expands the boundaries of application by implementing the WOz method to generate highly relevant, real-time visual stimuli—including CAD plans, 3D models, and inspirational images—based on the designers' ongoing sketches. This method not only reacts dynamically to the evolving design but also integrates a broader range of design elements, offering a richer and more flexible approach to design ideation. By engaging participants in a continuous loop of feedback and generation, our study introduces a novel application of WOz to design.

This literature review also shows that while the WOz method was applied to several fields of study in human-machine interaction, the specific field of application of design stimulation has been little investigated to date. The study we present in this paper aims to continue research in this area.

3. METHODS

3.1. Context of the case study application

Our broad research aims to gain insights on the instrumentation of early-phases design, where concepts are developed and refined, as it is uncommon to see tools being specifically developed for the specific activities of these phases. The AI design assistant under examination intends to stimulate idea generation by sending project-specific visual stimuli, aligned with the current design focus and derived from the designer sketches. These stimuli cover inspirational images and two- and three-dimensional visual representations of the project. Evaluating this technology through the WOz method allows us to investigate its usefulness and benefits for designers while also acquiring expertise in strategies for sketch recognition.

3.2. WOz implementation

The benefits and potential pitfalls of this AI design assistant are evaluated in a 90-minute design session, in which the designer conceptualizes a family housing solution on a sloping terrain while being presented with the visual stimuli at five-minute intervals.

In the Design Room (Fig. 1 on the right), the designer faces a digital sketch table, the printed design brief, paper site images, and three displays. From left to right, the displays show the session timer, a set of inspirational images, the 2D CAD-cleaned plans, a 3D model, and a discussion terminal with which the designer may command specific complementary images or representations. The researcher is also present in the room to facilitate the design session and to make preliminary observations.

In the Wizard Room (Fig. 1 on the left), the three modeling wizards and a coordinator face a control screen that displays the sketch being drawn in real time, the three stimuli displayed, the terminal, and a view of the design room. Each of the wizards interprets the sketch and then produces a specific sub-function of the simulated tool: (i) retrieving the appropriate inspirational images, (ii) producing the CAD plans, and (iii) modeling the 3D model. The coordinator helps interpret the sketch, triggers the sending of the produced visuals, and manages the terminal.

3.3. Designer and Wizard participants

Two wizard and coordinator teams were formed (based on calendar availability), with four people per team, for a total of eight people. All were engineering-architects students in their third to fifth year of a five-year program, both male and female (M=5, F=3), with an average age of 23 years old. They were pre-selected based on a task performance test and trained.

Seventeen designers (Table 2) participated in the study, with eight males (mean age = 36 years old) and nine females (mean age = 29 years old). The participants had a diverse range of professional backgrounds, including architecture and architectural engineering. They had between 2 and 30 years of professional experience, with an average of 10.5 years. They were recruited via email from local architecture agencies, and consented to taking part after being informed of the experimental context and procedure, although they were not aware that the tool was a WOz setup.

3.4. Tasks

The wizards were assigned the sub-task of image retrieval, 2D modeling or 3D modeling, in which they demonstrated the highest level of proficiency at the selection test to ensure that the tool was simulated as performatively as possible. They maintained their specific sub-task throughout each design session.

Regarding the design task assigned to the designers, in contrast to most HCI research, which employs relatively simple tasks requiring minimal specific background knowledge and speaks to a majority of non-domain-qualified participants, our approach

TABLE 1: SYSTEMATIC LITERATURE REVIEW OF THE USAGE OF WOZ FOR DESIGN STIMULATION STUDIES

Study	Design task	Software function	Stimuli	Interaction modality
Khan et al., 2016 [20]	Generate their most innovative concepts for a Japanese Zen garden.	A "robot", (in WOz), prompts participants to generate ideas by inquiring, *"Can you think of another way to do that?"*. Additionally, the "robot" retrieves pertinent images and video clips from the Internet, which it presents to the participants to facilitate the generation of creative concepts.	Verbal impulse calling for more ideas, and displayed inspiring visuals retrieved according to the design topic.	Verbal and Visual
Xu et al., 2020 [21]	Engage in an oral brainstorming session within a group context, with the objective of addressing a certain design prompt (not explicitly delineated in the paper).	2-by-2 condition setting where the "AI" provides the designers with a preset picture gallery or a preset word gallery. This is done passively (every 30 seconds) or actively (as frequently as necessary).	Displayed inspiring visuals or keywords retrieved from a preset gallery.	Verbal and Visual
Fu and Zhou, 2020 [22]	Design a lamp in three distinct phases: analysis, sketching, and production.	The AI collaborator agent (the Wizard) served as a source of guidance and suggestions, offering input both verbally and through the incorporation of ideas into the designer's sketch.	Additional information given orally or in a sketch on the design choice implemented.	Verbal and Visual (sketched)
Kim et al., 2022 [23]	Design eight different solutions for the lighting of a car interior in a VR setting with a headset.	The wizard modeled the light setting according to the instructions provided by the designer's head gesture. Additionally, the wizard offered advice on the selected lighting characteristics and their impact, such as the assertion that *"typically, orange-toned lighting creates a tranquil ambience"*.	Additional information given orally on the design choice implemented.	Gestural, Visual, and Verbal
Poser et al., 2022 [24]	[non-specified]	The WOz conversational agent interacted with participants through scripted responses and facilitated their idea generation by providing them with insights from cycles designed previously.	Displayed inspiring visuals retrieved according to the design topic.	Verbal and Visual

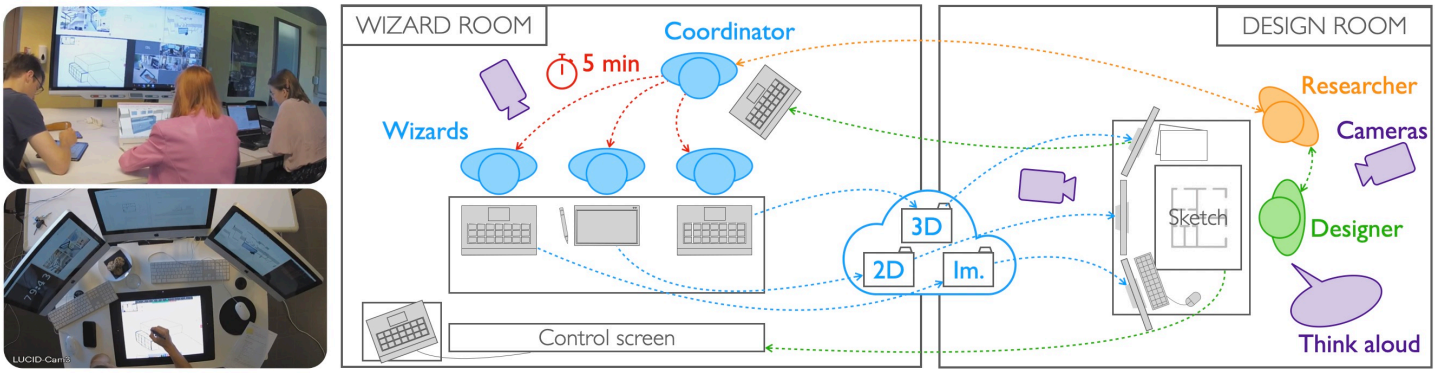

FIGURE 1: EXPERIMENTAL SET-UP OF THE WOZ METHOD FOR DESIGN STIMULATION

involves the use of a realistic design brief. We ensured that the design brief was sufficiently challenging to maintain the engagement of the designers and provided a reason to use stimuli instrumentation, while still guaranteeing a satisfactory level of

TABLE 2: DESIGNERS POPULATION DESCRIPTION

P	Gend.	Age	Background	Activity	Seniority
1	M	52	Eng. Arch.	Agency	Senior
2	F	24	Eng. Arch.	Agency/Res.	Junior
3	M	25	Eng. Arch.	Agency	Junior
4	F	34	Arch.	Agency/Res.	Interm.
5	F	30	Eng. Arch.	Agency	Interm.
6	M	30	Eng. Arch.	Agency	Interm.
7	M	48	Eng. Arch.	Agency	Senior
8	M	30	Arch.	Research	Junior
9	M	28	Eng. Arch.	Agency	Junior
10	F	31	Eng. Arch.	Agency	Interm.
11	F	24	Eng. Arch.	Agency	Junior
12	F	40	Eng. Arch.	Agency	Senior
13	M	33	Eng. Arch.	Agency/Res.	Interm.
14	M	45	Eng. Arch.	Agency	Senior
15	F	25	Eng. Arch.	Agency/Res.	Junior
16	F	27	Eng. Arch.	Agency	Interm.
17	F	27	Eng. Arch.	Agency	Interm.

design achievement by the conclusion of the 90-minute design session. Furthermore, the brief was designed so the wizards' production task remained within a manageable scope.

3.5. Data collection and coding method

Given the high complexity of this type of experiment and studied phenomenon, a wide range of data collection instruments were set up. Designers verbalized their thought process during the design session using a **think-aloud** protocol. We obtained a detailed account of the design solution, its rationale, and the designer's comments on the visual stimuli. Before the design session, designers did a training think-aloud exercise on a simplified design task. During the session, a visual cue reminded them to verbalize their thought processes. If needed, the researcher also provided a reminder. The collected think-aloud are divided into 10-second segments, with each segment transcribed. From this temporal transcription, we proceed to a double-blind extraction of keywords describing the stimuli commentary, analogies and design attributes. We also **video recorded** the design room, wizards' room, control screen displaying the sketch and associated stimuli, and the designer's gaze and interactions with the experimental environment from multiple perspectives to document the design process. Camera recordings are subjected to analysis at 10-second intervals. For each segment, we double-blindly code the focus of the designer's gaze and the moment visuals appear on screen. During the experiment, the researcher also observed and noted stages of the design process and important moments. After the session, designers were interviewed in an **auto-confrontation** protocol [25] and shown these moments on camera, then asked to clarify their reasoning or actions. The transcripts from the auto-confrontation interviews are then subjected to qualitative analysis in order to elucidate specific instances or patterns of behavior that have been identified. Semi-structured interviews were also conducted with designers using a **questionnaire** with a four-point Likert scale to assess the usefulness and interference of different stimulus types (inspirational images, cleaned plans, 3D models): *"Did you miss any content or receive it at the wrong time? Did your representations affect your design activity?"*. The **wizards** were also interviewed using an auto-confrontation protocol to express their strategies for sketch recognition and overall workflow. Transcribed auto-confrontations with the wizards are analyzed following a coding procedure developed by Lejeune [26] where each elucidated action of the wizards is conceptualized as a "tag". Ultimately, a comprehensive diagram is generated, delineating the sequence of actions and sub-actions undertaken by the wizards to accomplish the sketch recognition task, as elucidated in the interviews.

3.6. Success assessment method

We assess the success of the WOz simulation based on the following criteria:

Data and knowledge accessed. We evaluate the quantity of data collected through the simulation and the extent of knowledge it provided access to.

Relevance of the stimulation strategy. We first assess the usefulness and level of disruption caused by the stimuli produced by the wizards. Semi-structured interviews conducted at the end of the design sessions were used to evaluate these aspects for each type of stimulus provided. Designers were asked to rate the visuals received on two Likert scales: (1) usefulness, ranging from 1 (not useful) to 4 (very useful), and (2) perturbation level, ranging from 1 (non-perturbing) to 4 (perturbing).

Wizards' efficiency. The effectiveness of the implementation is measured by the wizards' ability to complete their tasks efficiently and the rate at which stimuli were produced. Additionally, we examine how well the wizards understood and anticipated the designers' intentions, and how effectively they selected or generated the most appropriate visuals in response.

Designer's perception. To evaluate the success of the simulation, we also consider the designers' perceptions, both in terms of their ability to discern that humans, rather than a real tool, were involved, and their impressions of the tool's performance. The transcription of the think-aloud highlights five dimensions of designers' perception of the AI design assistant that we will investigate further: (1) the proportion of positive comments on the received stimuli, (2) the relevance of the timing of the stimuli, (3) the relevance of the style of the inspirational images, (4) the alignment of the visuals with the designer's expectations, and (5) any perceived wizards' misunderstandings.

4. RESULTS

4.1. General assessment and knowledge accessed

This WOz method allows us to instrument each design session with less than 35 minutes of setup time (which is short for an experiment of this scale) and can be easily repeated (as for 17 times for the present case-study). The designers each successfully completed their preliminary design task before the end of the session time, achieving a satisfactory design solution. Together, they designed 1,459 building attributes, meaning that their design activities achieved a rate of 0.95 new elements per minute. The

wizards achieved a production of 852 stimuli, from which the designers extracted a total of 225 design ideas.

The experimental setup has provided enough quality data to publish several studies on various topics, characterizing:

- The mechanisms underlying complex sketch recognition, to gain insights into a future intelligent sketch-based design tool, based on the wizards' strategies and workflow to go from the perceived sketched line to a mental model of the designed object [27].
- The benefits and pitfalls of automating analogical reasoning through a sketch-based image generator for design, focusing in particular on the uses of the inspirational images stimuli [28].
- The multi-instrumented reflective interactions appearing between the designer and the visuals sent as stimuli, and their implication for the preliminary design activity [29].
- The human-machine co-creation exchanges supported by this sketch-based communication modality, to discuss the benefits and challenges of AI image generators for design ideation [30].

4.2. Relevance of the stimulation strategy

In the semi-structured interviews, the designers indicated that they were generally satisfied with the design process in this WOz setting. As for the usefulness, most of the designers described 3D modeling as an interesting way to represent the volumes and proportions of the various elements of the building. The reference pictures (M=2.56, SD=1.16) and the cleaned plans (M=2.53, SD=0.99) were rated as slightly less useful than the 3D models (M=2.84, SD=1.11) (Fig. 2). However, the large standard deviations indicate a wide range of opinions among the designers. These results suggest that the perceived usefulness of the stimuli received was more influenced by the reception context, regardless of the quality of content provided by the wizard.

As for the perturbation, because the CAD plans presented the wizards with a slightly more complex set of challenges, designers spotted minor discrepancies between their drawings and the wizards' productions due to misunderstandings or delayed updates and therefore put more time and effort into their drawings to improve machine understanding, thus disrupting their activity. This issue does not arise with inspiration pictures and 3D models, as they are not as directly representative of the project. This is a phenomenon reflected to some extent in the perturbation evaluation, with Figure 3 showing that the plans (M=1.38, SD=0.60) were considered slightly more disruptive than the reference pictures (M=1.28, SD=0.56) and 3D models (M=1.25, SD=0.56). In general, however, these perturbation scores were very low, indicating that the participants were very satisfied with the wizards' production.

During the semi-structured interviews, designers furthermore indicated that co-designing with the technology proved beneficial for both idea generation and idea communication with fictitious clients. They also highlighted its non-locking functionality and time-saving capabilities. Designers identified potential avenues for enhancement, suggesting that the stimuli be aligned with the level of detail of the design at the time of receipt and that specifically commanded images be transmitted with greater expediency.

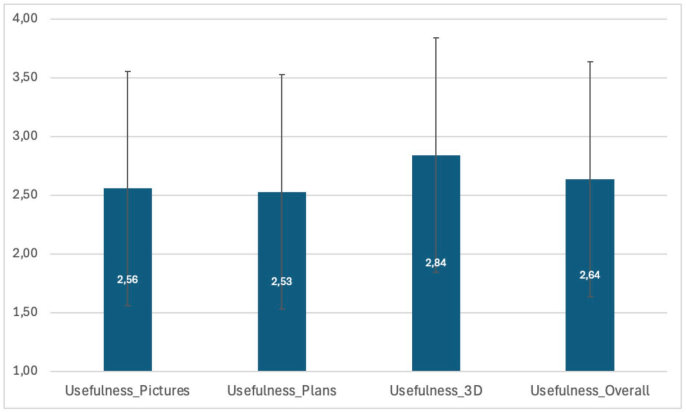

FIGURE 2: PERCEIVED USEFULNESS FOR THE PICTURES, THE PLANS, AND THE 3D MODELS SENT BY THE WIZARD (1= NOT USEFUL ; 4 = VERY USEFUL)

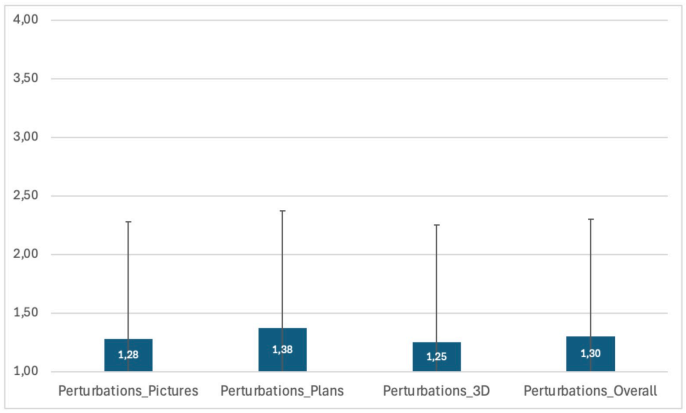

FIGURE 3: PERCEIVED PERTURBATIONS FOR THE PICTURES, THE PLANS, AND THE 3D MODELS SENT BY THE WIZARD (1 = NON-PERTURBING; 4 = PERTURBING)

4.3. Wizards' efficiency

To assess the wizards' ability to complete their assigned task, we evaluate the interval between two completed productions and sending of stimuli, and compare it to the 5-minutes prompted rate. Over the course of the 17 design sessions, the wizards transmitted the stimulus with an average time interval of 5 minutes and 14 seconds (SD=25sec). The target rate of stimulus delivery, which was instructed to be 5 minutes, has thus been successfully met.

Additionally, we analyze the wizards' ability to comprehend and predict the designers' intentions and produce the most appropriate visual stimuli. Wizards in auto-confrontation interviews said they anticipated participants' expectations using their design domain experience and identifying design paths and logic that the successive designers followed. Thanks to this knowledge, the wizards know designers' design method, which item

will be designed first and in which order following elements are usually designed. Wizards also tailor their production to designers' specific activities. They "get under the skin" of designers to understand their behaviors and intentions. Wizards adapt each designer's practices to meet their expectations. The wizards moreover say they separate the designers' intentions from errors or non-intentional sketches. Sometimes, designers commit to a sketch that serves as thinking support but does not align with their vision. The wizards must recognize this and not model these sketched reflections. They admit that mistakes happen in their work, but they are quick to correct them to not lose the designer's trust. Finally, the wizards collaborate among their team to ensure they interpret the designer's behavior correctly, even when it is difficult in this dynamic situation of simulation.

4.4. Designers' perception

First of all, none of the participants were able to discern the human input in the interaction with the technology. However, they showed varying degrees of suspicion. Two individuals expressed suspicious admiration for the tool's ability to understand their drawings, yet didn't question further the veracity of the design set-up as they lacked awareness of the WOz method and had a robust relationship of trust with the researcher. Two other participants, upon being informed at the end of the experiment that the setup was a Wizard of Oz, expressed their previous suspicion through sentiments akin to *"I had also perceived this software to be remarkably sophisticated"*. The remaining 13 participants were flabbergasted by the implausible configuration underlying their perception of a real AI tool.

Prior to disclosing the WOz method, participants expressed their impression of the sketch-based AI tool as impressive and sophisticated in semi-structured interviews. They also perceived the tool as highly innovative. Furthermore, participants indicated that the tool surpassed their expectations in terms of its capacity to understand the sketched design and simulate the potential visual outcomes. However, they also noted occasional delays in reactivity covering the subsequent versions of the building and identified areas for improvement regarding the various stimuli presented.

Although two distinct teams of wizards were involved in the design sessions, no biases or discrepancies were observed in the performance or feedback provided by participants. As the training process and behaviors' instructions were standardized across both team to ensure consistency in the execution of the experiment, the variation in wizard teams did not affect the overall results.

Investigating further the designer's perception of the simulation, we analyzed the think-aloud data with a thinner granulometry. This revealed five dimensions reflecting designer's perception. We present their positive and negative perceptions along each dimension:

Proportion of positive comments on the received stimuli. From the think-aloud data, we observed that 24 percent of the visuals sent were commented on positively by the designer receiving them. We also record 68 explicit positive "software"-related comments following the reception of stimuli. These positive comments includes designers stating *"He understood that I was sketching a section; that was impressive"*; sometimes, specifically on inspirational image stimuli *"Yes, the staircase is a nice idea, it gives inspiration"*; and sometimes leading to design evaluations like *"Oh yeah, the facade is refined, that's cool; and I think it'll look good from the street"*.

Relevance of the timing of the stimuli. Overall, the designers considered the system's reaction time to be satisfactory (e.g. *"Sending is smooth; the images arrive at just the right moment"*). The time at which visuals were spontaneously proposed by the "AI" was always considered relevant (e.g. *"Five minutes is fine in general, but for the images we ordered, it took a little too long"*). On the other hand, the time needed to receive the explicitly requested visuals was considered too long.

Relevance of the style of the inspirational images. Some designers felt that the images had been taken from the Pinterest image database and were therefore too typical and unsatisfactory. One expressed *"Ah... that's from Pinterest, isn't it? Could we change it? I don't like this style, it's too typical"*. This factor is of comparatively minor significance, as it was mentioned by designers on a mere four occasions.

Alignment of the visuals with the designer's expectations. Generally, the images were judged to be relevant, corresponding both to what was expected and to the design brief. Designers stated *"The pictures are pretty close, even if it's not what I expected"* or *"Yes, it matched what I was imagining but pushing the idea further"*. Additionally, 27 percent of the 852 visuals were utilized to integrate new ideas into the building design. The diversity of the images (plans, interior and exterior perspectives, etc.) was also generally appreciated. Designers said *"I'm really happy, nothing was missing"*. Some designers felt that the images sometimes focused too much on design details, and should focus more on functional zoning plans (e.g. *"I think the images are too detail-oriented, so it's a bit out of sync with the general zoning I was working on in the very beginning"*).

Perceived wizards' misunderstandings. The first form of misunderstanding occurs when the designer errs in their design and creates an impossible proposition. As part of the software simulation, the wizards are not allowed to rectify the designer's mistake, resulting in a blockage in the modeling process, which sometimes has a cascading effect. For example, Participant 9 at some point in the design session erroneously depicts the ground floor as 10 meters wide instead of 12 meters, and then divides it into three rooms. The wizards proceed to model the plan in accordance with the dimensions provided. However, they are unable to establish a connection with the neighboring houses, despite the fact that the designer's project was supposed to be adjacent. The second form of misunderstanding arises when the wizards do not fully update the models before sending them off, often due to time constraints. Regardless of form, a total of 45 instances of misunderstanding were identified in a corpus of 854 visual stimuli produced. This low occurrence, amounting to 0.05 percent, aligns with the positive designer's perception.

5. DISCUSSION

This section begins by synthesizing the results of our case study, which demonstrated the successful implementation of the

WOz method in simulating emergent design tools. We then discuss the method's benefits and limitations in the context of design studies. Finally, we propose actionable best practice recommendations for future researchers interested in applying this method to design applications.

5.1. Synthesis of the case-study implementation success

The results section demonstrated the successful implementation of the WOz method in this design case study. The new knowledge derived from the experiment confirmed that the method generated sufficient data to document the tool's usage and interactions effectively (see 4.1). Evaluating the perceived usefulness and perturbation of the received stimuli revealed that all types of visuals were considered non-disruptive. Designers' assessments of usefulness were more in link with the context and timing of the stimuli rather than the wizard's production (see 4.2). In terms of the wizards' efficiency, they successfully adhered to the required rate of stimuli production while maintaining highly project appropriate content. They showed a strong understanding of the designers' intentions and behaviors (see 4.3). Additionally, designers were unable to distinguish human input within the WOz setup and expressed high levels of satisfaction with the tool's capabilities. Although some designers noted occasional delays, slight mismatches in the desired level of detail of the representation, or misunderstandings, they overwhelmingly viewed the tool positively, emphasizing the relevance of the timing and of the content of the stimuli (see 4.4). Overall, these findings confirm that the wizards performed well and that the WOz method effectively simulated the emerging technology under study.

5.2. WOz method benefits for emergent design tools

Our case study reaffirms the benefits of the WOz method that have been previously discussed in the literature, while also highlighting specific advantages in the design context.

Range of applications. The WOz method is versatile, supporting a variety of human-technology interactions across multiple domains, including design. As we saw in our application, designer-wizard(s) interactions can be visual (e.g., gestures, images, sketches) or verbal (typed or spoken), and they can vary depending on the simulated function. Our case study involved a combination of visual and verbal inputs from the participants, along with typed commands, while the system provided both verbal and visual responses. This adaptability allows the WOz method to be applicable to diverse design tasks, as demonstrated by its use in human-robot interaction and conversational agents. The flexibility to incorporate various input and output forms makes the WOz method highly adaptable to the needs of emergent design tools.

Cost associated. One of the primary advantages of the WOz method is its ability to simulate complex phenomena, such as design under stimuli, in a cost-effective manner. Traditional methods, such as interviews or focus groups are limited in that they cannot provide a realistic simulation of the design environment, thus participants may struggle to envision themselves using the emergent tool. Moreover, creating prototypes for such early-stage explorations can be prohibitively expensive. The WOz method allows for the creation of high-quality, realistic visual stimuli for designers at various stages of the design process, something that static image libraries or database systems cannot offer. In this way, WOz aligns with the philosophy of the prospective ergonomics approach [31], anticipating future user needs and artifacts. It is human-centered, future-oriented, and focused on fostering creativity and innovation.

5.3. WOz method limitations for emergent design tools

Despite the advantages, our case study also revealed several limitations inherent in the WOz method.

Wizards' work. While the WOz method effectively simulates tool functions, it places a significant workload on the wizards. To ensure the simulation's success, wizards were preselected based on their expertise in architectural design (3–5 years of study), meaning they were proficient in reading architectural floor plans, CAD drawings, and 3D modeling. Wizard candidates took a test, and those who passed were assigned their most relevant sub-function (either reference image, 2D CAD plans, or 3D model) and received 1.5 hours of training. Despite this preparation, the workload for the wizards was considerable. They had to maintain concentration for 90 minutes while producing images that were consistently relevant to the design subject. To manage this, the wizards developed strategies for prioritizing tasks—starting with essential elements that define the main design constraints, followed by secondary design elements that specify design characteristics, and finally non-essential elements that serve aesthetic or communicative purposes—and ensuring that they modeled design elements as decisions were made. These strategies were crucial for the success of the simulation, as demonstrated by the consistent 5-minute rhythm of image production. However, the level of effort required from the wizards should not be underestimated.

Ethical aspects. The primary ethical concern with the WOz method lies in the concealment of the human involvement in the simulation. Although most designers admired the stratagem, a few were disappointed upon learning that the technology was simulated by humans rather than an actual automated system. This revelation led to a slight erosion of trust between the participating designers and the researchers. The rationale for concealing the human element is to prevent bias—if designers knew they were interacting with a human rather than a tool, they might alter their behavior or expectations. However, this raises ethical concerns, as the deception could affect participants' trust. Further research is needed to understand the potential effects of revealing the WOz setup on participant behavior and whether this disclosure would lead to biased results.

Associated logistic. While the WOz method reduces the costs of developing technology, it is not without its own expenses. Setting up the WOz system required extensive planning, protocol design, and a variety of tools and technologies. Preparatory work was essential to ensure a smooth implementation. Thus, while the WOz method can reduce development costs, the engineering

and design efforts involved in its preparation are substantial as observed in our case-study and as confirmed by other researchers [32]. Therefore, while it offers cost-effectiveness in simulating technologies, it requires careful consideration of the resources needed for setup and execution.

5.4. Best practice recommendation for design applications

During the design of our experimental protocol, the most detailed set of recommendations for applying the WOz method came from Fraser, who, at the conclusion of his use of the method, noted that "WOz simulations are not a panacea: they are only useful if certain pre-conditions are met (...), namely: 1. it must be possible to simulate the future system, given human limitations; 2. it must be possible to specify the future system's behavior; 3. it must be possible to make the simulation convincing" [33, p. 82]. Fraser's three key criteria remain fundamental. Based on the characteristics of our WOz implementation that were identified as crucial for its success, we formulate a series of more detailed and actionable recommendations:

1. **To make the simulation possible given human limitations - Training and standardize for wizards:** Effective training of the wizards is essential. But moreover, our findings suggest that a structured, multi-phase training program is required for optimal performance. In our case, wizards' performance improved only after multiple training sessions (i.e. two pilot experiments plus the first two participants)—indicating that a substantial number of sessions are necessary for developing the skills required to handle real-time design tasks effectively. Future implementations should consider strategies to mitigate fatigue effects. Moreover, wizards must be trained to read the designer's workflow to adapt their responses accordingly. In addition, wizards should be equipped with strategies to handle the situations where designers present ambiguous or exploratory inputs (e.g., non-final sketches, verbal brainstorming). Ensuring that wizards can distinguish between exploratory ideas and final decisions helps maintain the flow of the design process and reduces the risk of introducing errors or confusion.

2. **To specify the system's behavior - Developing a precise experimental protocol:** The success of the simulation relies heavily on a detailed and well-structured experimental protocol. Our experience demonstrated that while wizards were given comprehensive guidelines for producing relevant design stimuli, they still had to adapt to unforeseen design circumstances, showing the need for a dynamic protocol that accounts for variations in designer behavior and that allows flexibility for the wizards to make real-time adjustments while maintaining consistency in the simulation's outcome. Moreover, defining the inputs, outputs, and triggers for each simulated function is vital to ensure smooth interactions between the wizards and participants. Finally, when working with a team of wizards, it is essential to implement effective communication and collaboration tools within the protocol to facilitate coordination and ensure smooth interactions.

3. **To make the simulation convincing - Aligning wizard actions with designers expectations:** A key element to ensure the illusion of the technology is preserved is aligning wizard actions with participants' mental models of how the system works. Our study found that deception was easier because participants were unfamiliar with the specific AI capabilities being simulated. However, common software behaviors are already well known to users, which necessitated precise actions to avoid revealing the illusion. Careful attention must be paid to visual aspects such as component's apparition timing, placement, and selection, which should align seamlessly with user expectations. Moreover, rather than strictly adhering to a "hands-off" simulation model, wizards should have the flexibility to provide light, non-intrusive correction strategies when the designer makes a mistake or exhibits a misunderstanding in their design approach. This flexibility ensures that the design process remains productive, preventing unnecessary blockages while still maintaining the illusion of an autonomous system. In addition, the system must respond promptly to designers' inputs to avoid unnecessary delays that can disrupt the design process. Timely responses enhance the perception of the tool's competence and utility, ensuring that designers feel supported throughout the interaction. Finally, the researchers should define how the experiment and its goals will be communicated to participants to maintain the credibility of the simulation.

6. CONCLUSION

The Wizard of Oz (WOz) method has been employed for decades to simulate human-machine interactions without the need for technological development. In this paper we presented a novel application of the WOz method in design research, specifically for generating support visuals to stimulate the design process. We contribute to the existing literature by expanding the method's application beyond its traditional domains, such as human-robot interaction and autonomous vehicles, into the field of design research and by providing a detailed documentation of this application. Through a critical evaluation of the method's performance, we identify both the benefits and limitations of using WOz in design contexts. Additionally, we offer actionable recommendations for researchers seeking to employ WOz in studies of design tools and methodologies. Overall, our findings provide valuable insights that can guide future research studying the use of emerging technologies in design.

ACKNOWLEDGMENTS

We would like to thank the LUCID-ULiège and the Co-Design Lab of UC Berkeley for their financial support in covering the expenses associated with this experiment. Additionally, we are grateful to LUCID-ULiège for hosting the experiment in their Digital studios lab space and providing the necessary facilities and resources.

REFERENCES

[1] Dahlbäck, Nils, Jönsson, Arne and Ahrenberg, Lars. "Wizard of Oz studies: why and how." *Proceedings of the 1st international conference on Intelligent user interfaces*: pp. 193–200. 1993.

[2] Browne, Jacob T. "Wizard of oz prototyping for machine learning experiences." *Extended abstracts of the 2019 CHI conference on human factors in computing systems*: pp. 1–6. 2019.

[3] Rietz, Finn, Sutherland, Alexander, Bensch, Suna, Wermter, Stefan and Hellström, Thomas. "WoZ4U: an open-source wizard-of-oz interface for easy, efficient and robust HRI experiments." *Frontiers in Robotics and AI* Vol. 8 (2021): p. 668057.

[4] Alevêque, Guillaume. "Intelligence et artifice. Le Magicien d'Oz ou la simulation de l'interaction humain-machine." *Techniques & Culture. Revue semestrielle d'anthropologie des techniques* (2019).

[5] Dow, Steven, MacIntyre, Blair, Lee, Jaemin, Oezbek, Christopher, Bolter, Jay David and Gandy, Maribeth. "Wizard of Oz support throughout an iterative design process." *IEEE Pervasive Computing* Vol. 4 No. 4 (2005): pp. 18–26.

[6] Taib, Ronnie and Ruiz, Natalie. "Wizard of Oz for multimodal interfaces design: Deployment considerations." *Human-Computer Interaction. Interaction Design and Usability: 12th International Conference, HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part I 12*: pp. 232–241. 2007. Springer.

[7] Bradley, Jay, Benyon, David, Mival, Oli and Webb, Nick. "Wizard of Oz experiments and companion dialogues." *Proceedings of HCI 2010*. 2010. BCS Learning & Development.

[8] Weschke, Jan, Bahamonde-Birke, Francisco J, Gade, Kay and Kazagli, Evanthia. "Asking the Wizard-of-Oz: How experiencing autonomous buses affects preferences towards their use for feeder trips in public transport." *Transportation research part C: emerging technologies* Vol. 133 (2021): p. 103454.

[9] Ranjbar, Parivash, Krishnakumari, Pournami Krishnan, Andersson, Jonas and Klingegård, Maria. "Vibrotactile guidance for trips with autonomous vehicles for persons with blindness, deafblindness, and deafness." *Transportation research interdisciplinary perspectives* Vol. 15 (2022): p. 100630.

[10] Vithanage, Shashindi, Dey, Arindam, Korte, Jessica, of Electrical, Institute and Electronics Engineers, issuing body., author. *'Auslan Alexa': A case study of VR Wizard of Oz prototyping for requirements elicitation with Deaf participants*. 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), IEEE, Piscataway, NJ : (2023-3).

[11] Helgert, Andre, Straßmann, Carolin and Eimler, Sabrina C. "Unlocking Potentials of Virtual Reality as a Research Tool in Human-Robot Interaction: A Wizard-of-Oz Approach." *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*: pp. 535–539. 2024.

[12] Sabeti, Sepehr, Ardecani, Fatemeh Banani and Shoghli, Omidreza. "Augmented reality safety warnings in roadway work zones: Evaluating the effect of modality on worker reaction times." *Transportation Research Part C: Emerging Technologies* Vol. 169 (2024): p. 104867.

[13] Rodríguez-Domínguez, María Trinidad, Bazago-Dómine, María Isabel, Jiménez-Palomares, María, Pérez-González, Gerardo, Núñez, Pedro, Santano-Mogena, Esperanza and Garrido-Ardila, Elisa María. "Interaction Assessment of a Social-Care Robot in Day center Patients with Mild to Moderate Cognitive Impairment: A Pilot Study." *International Journal of Social Robotics* Vol. 16 No. 3 (2024): pp. 513–528.

[14] Sienkiewicz, Barbara, Sejnova, Gabriela, Gajewski, Paul, Vavrecka, Michal and Indurkhya, Bipin. "How language of interaction affects the user perception of a robot." *International Conference on Social Robotics*: pp. 308–321. 2023. Springer.

[15] Jang, Hyeji, Park, Daehee and Bae, Sowoon. "Intelligent Moving Display Robots: Human-Centric Design Guidelines for Human–Robot Interaction Designers." *International Journal of Human–Computer Interaction* (2024): pp. 1–15.

[16] Vrins, Anita, Pruss, Ethel, Ceccato, Caterina, Prinsen, Jos, De Rooij, Alwin, Alimardani, Maryam and De Wit, Jan. "Wizard-of-Oz vs. GPT-4: a comparative study of perceived social intelligence in HRI brainstorming." *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*: pp. 1090–1094. 2024.

[17] Aljuneidi, Saja, Heuten, Wilko, Tepe, Markus and Boll, Susanne. "Did that AI just charge me a fine? Citizens' perceptions of AI-based discretion in public administration." *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*: pp. 57–67. 2023.

[18] Page, Matthew J, Moher, David, Bossuyt, Patrick M, Boutron, Isabelle, Hoffmann, Tammy C, Mulrow, Cynthia D, Shamseer, Larissa, Tetzlaff, Jennifer M, Akl, Elie A, Brennan, Sue E et al. "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews." *bmj* Vol. 372 (2021).

[19] Page, Matthew J, McKenzie, Joanne E, Bossuyt, Patrick M, Boutron, Isabelle, Hoffmann, Tammy C, Mulrow, Cynthia D, Shamseer, Larissa, Tetzlaff, Jennifer M, Akl, Elie A, Brennan, Sue E et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." *bmj* Vol. 372 (2021).

[20] Kahn, Peter H, Kanda, Takayuki, Ishiguro, Hiroshi, Gill, Brian T, Shen, Solace, Ruckert, Jolina H and Gary, Heather E. "Human creativity can be facilitated through interacting with a social robot." *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*: pp. 173–180. 2016. IEEE.

[21] Xu, Jingwen, Chao, Chi-Ju and Fu, Zhiyong. "Research on intelligent design tools to stimulate creative thinking." *International Conference on Human-Computer Interaction*: pp. 661–672. 2020. Springer.

[22] Fu, Zhiyong and Zhou, Yuyao. "Research on human–AI co-creation based on reflective design practice." *CCF Transactions on Pervasive Computing and Interaction* Vol. 2 (2020): pp. 33–41.

[23] Kim, Taesu, Shunayeva, Aigerim, Lee, Gyunpyo and Suk, Hyeon-Jeong. "Sketching in-vehicle ambient lighting in virtual reality with the Wizard-of-Oz method." *Digital Creativity* Vol. 33 No. 1 (2022): pp. 49–63.

[24] Poser, Mathis, Küstermann, Gerrit C, Tavanapour, Navid and Bittner, Eva AC. "Design and evaluation of a conversational agent for facilitating idea generation in organizational innovation processes." *Information Systems Frontiers* Vol. 24 No. 3 (2022): pp. 771–796.

[25] Mollo, Vanina and Falzon, Pierre. "Auto-and allo-confrontation as tools for reflective activities." *Applied ergonomics* Vol. 35 No. 6 (2004): pp. 531–540.

[26] Lejeune, Christophe. *Manuel d'analyse qualitative*. De Boeck Supérieur (2019).

[27] Baudoux, Gaëlle and Goucher-Lambert, Kosa. "Understanding Complex Sketch Recognition Strategies for Intelligent Sketch-Based Design Tools." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 88407: p. V006T06A017. 2024. American Society of Mechanical Engineers.

[28] Baudoux, Gaëlle and Goucher-Lambert, Kosa. "Automating Analogical Reasoning: A Wizard of Oz study on the benefits and pitfalls of a sketch-based AI image generator for design." *International Conference on-Design Computing and Cognition*: pp. 21–37. 2024. Springer.

[29] Baudoux, Gaëlle and Safin, Stéphane. "Study of computer multi-instrumented reflexive conversation activity in preliminary architectural design." *International Journal of Architectural Computing* (2025): p. 14780771241310207.

[30] Baudoux, Gaëlle. "The benefits and challenges of artificial intelligence image generators for architectural ideation: Study of an alternative human-machine co-creation exchange based on sketch recognition." *International Journal of Architectural Computing* Vol. 22 No. 2 (2024): pp. 201–215.

[31] Robert, Jean-Marc and Brangier, Eric. "Prospective ergonomics for the design of future things." *Ergonomics* (2024): pp. 1–18.

[32] Bernsen, Niels Ole, Dybkjær, Hans and Dybkjær, Laila. "Wizard of oz prototyping: How and when." *Proc. CCI Working Papers Cognit. Sci./HCI, Roskilde, Denmark* (1994): p. 67.

[33] Fraser, Norman M and Gilbert, G Nigel. "Simulating speech systems." *Computer Speech & Language* Vol. 5 No. 1 (1991): pp. 81–99.