Simulators' mimetic behaviours with automated simulation tools in wizard of oz study

Kuo-Kuang Fan[1] Xiang-Ming Zhong[2] Xiaolong Lou[3] Weidong Geng[3] Xiangdong LI[3]

[1] National Yunlin University of Science and Technology, Graduate School of Design

[2] National Yunlin University of Science and Technology, Graduate School of Design; Communication University of Zhejiang, Consortium of Internet and Society

[3] Zhejiang University, Dept. of digital media, college of computer science and technology; Alibaba-Zhejiang University Joint Institute of Frontier Technologies

Abstract

Wizard of oz (WoZ) employs human simulators to mimic intelligent systems that are beyond current technology levels or time- and cost-consuming to implement. Automated simulation tools are increasingly adopted in WoZ simulation. Conventional studies are usually focused on end users and little is known about the simulators' behaviours. To bridge this gap, we conducted a comparative study to investigate credibility, rigorousness, consistency, and efficiency in the performances of simulators as simulating smart spoken dialogue systems with automated simulation tools. The results reported two types of mimetic behaviours: the intentional and instinct response. Specifically, the simulators, regardless of experience levels, performed well as mimicking intentional responses, in which their responses were credible and efficient. However, the simulators exposed inconsistent and non-rigorous mimetic behaviours in instinctive responses. Furthermore, the experienced and amateur simulators showed significant differences in instinctive simulations in response speed, interpreting the end user inputs and making corresponding decisions. Nevertheless, the instinctive mimetic behavior had no significant effects on the perception of end user for the simulations. The implications for WoZ considering as an efficient design tool are discussed.

Keywords: Wizard of oz; Automated simulation tool; Spoken dialogue interface; Mimetic behaviour; Intentional and instinctive simulations

1. INTRODUCTION

The Wizard of Oz (WoZ) method employs human simulators (also called “wizards” as in the 1939 film of “The Wizard of OZ”) to mimic partial- or full-system components that are beyond current technology levels or are time- and cost- consuming to implement (Dahlbäck, Jonsson & Ahrenberg, 1993, Kelley, 1984). WoZ provides solution on a cost-effective when iterating system designs and it is useful in simulating spoken-language systems. This is because current human simulators still have better spoken-language abilities than computers and the human simulators adapt more easily to mimic different speaking agents (Thomason & Litman, 2013).

WoZ is quite flexible to implement, as it requires only a human simulator and a set of simulation tools to present pseudo-functionalities. For this reason, WoZ has been a popular iterative design tool in the field of speech communication and human-robot interaction. However, WoZ has also been criticized for inconsistent simulations in which is a condition usually caused by human simulators during the use of the simulation tools (Schlögl, Doherty & Luz, 2015, Sequeira, et al., 2016, Riek, 2012). On the one hand, inconsistency is not necessarily a weakness of WoZ while there is a notion to WoZ simulation. It emphasises the possibility for variation and adaptation and thus compares to inconsistency as a necessary feature (Schlögl, Doherty & Luz, 2013). For example, the inconsistency can offer certain advantages which allows for the exploration of unplanned dialogue strategies and unknown features. On the other hand, it comes to designing iterative process that focuses on rigorous testing when one form of simulation. The inconsistency can produce negative impacts on WoZ studies.

In general, the inconsistency is attributed to several perspectives. From the perspective of a simulator's personality, human simulators have individual preferences and emotional feeling which might inadvertently affect the simulator on the simulation decisions and mimetic behavior (Fraser & Gilbert, 1991). From the perspective of an ability of simulator, the interaction tasks become complicated which resulted in simulators become incapable of acting as rigorously for the computers during heavy calculations and repetitive operations (Grill, Polacek & Tscheligi, 2012). From the perspective of a behavior of simulator, experienced human simulators commonly used to be employed as a rigid emotionless system in WoZ studies which intended to deliver computer-like simulations and presented a convincing illusion to end users. However, because they are human, WoZ simulators make unintentional mistakes such as typos and misstatements not only during intense simulation tasks but also in everyday tasks (Bott & Laviola Jr, 2015). Considering the pressures in making instantaneous and precise responses during simulations, mimetic behaviors of a human simulator in WoZ studies involved innate risks of unpredictability and improvisation.

Reviewing the importance of human simulators in WoZ studies, previous studies have proposed measures to guide mimetic behaviors of simulators, including pre-study training (Ralph & Moussa, 2008), study scripts (Ashok, et al., 2014), separated simulators (Drummond & Litman, 2011), open-source platforms (Schlögl, Doherty & Luz, 2015), controlled recognition (Shiomi, et al., 2008), automated simulation tools particularly (Alce, Hermodsson & Wallergård, 2013). Currently, the simulation tools have become increasingly in integrated and automated (Pettersson & Wik, 2015). Such tools save simulators from large physical efforts and heavy cognition loads (Li & Bonner, 2014). For example, a simulator may need only to choose a simulation decision from a dropdown response list. Afterwards, the system automatically completes all the rest of the tasks: assembling the speech, voice synthesis, and speech outputs (Thomason & Litman, 2013). Previous studies have examined how the automated simulation tools help simulators gain operational efficiency and constrain improvisational operations. The results showed major improvements in the simulator on the overall simulation experience, response speed, and operational consistency (Adler, Iacobelli & Gutstein, 2016, Mavrikis & Gutierrez-Santos, 2010). However, despite the advantages of the automated simulation tools, one question is rarely concerned is how to apply these tools support the simulator responses?

Previous studies usually are focused on the behavior of the end users and the usability and effectiveness of simulation tools (Fabbrizio, Tur & Hakkani-Tür, 2005). However, the relations among these tools and human simulators remain a black box. That is, little is known about how these automated simulation tools are interacted and influence human mimetic behaviors of a simulator. This understanding would inspire other human-mediated design methods and related practices since it reflects how human simulators behave in a well-supported working environment.

In this paper we conducted a comparative exploratory study to investigate credibility, rigorousness, consistency, and efficiency of the simulators' mimetic behavior with automated simulation tools.

The major problem in this study is how the adoption of automated simulation tools in WoZ studies could support experienced and amateur mimetic behaviors of simulator. Owing to the wide adoption of WoZ, the significance of the study is self-explaining. The study values have three-fold. First, the study identifies two types of mimetic behaviors- the intentional and instinctive ones. In simulations with automated simulation tools, as the previous studies are mainly focused on the former type. Secondly, the study provides new understanding of the relations among human simulators and the use of automated simulation tools. The previous studies used to advocate that high-level simulation tools could comprehensively eliminate inconsistent simulations. Finally, the study provides the new understanding of how the automated simulation tools affect the simulators (including the amateur and experienced) while the previous studies prefer the experienced simulators. Furthermore, the study supplies the implications for the configuration as well as uses of human-involved design and evaluation methods which is of great inspiration for broad readership. Combining the above methodology, we believe that there are many new understandings and implications beyond current knowledge will be found in this study.

The rest of this paper is organized as follows. Section 2 reviews the state-of-the-art of WoZ studies. Specifically, the use of human simulators and the adoption of automated simulation tools in existing WoZ studies. Section 3 describes the methodological details of the study. Section 4 analyzes the data from the experiment. Section 5 summarizes the study results, and Section 6 discusses the study implications. Finally, Section 7 concludes the study findings.

2. Related Works

2.1 Roles of human simulators and mimetic behaviours

We initially structured the review along the simulator's core tasks and processes, which is a valid approach to elaborate on the important characteristics of WoZ. By walking through the procedural flows of existing WoZ studies, we extracted four major processes related to human simulator's mimetic behaviours, each of which addresses a respective simulation task. These processes include: (i) learning simulation scenarios and simulation tools; (ii) interpreting end users' inputs; (iii) making simulation decisions and formulating strategies; and (iv) using simulation tools to respond. These processes are distributed throughout WoZ studies with distinct impacts on simulation consistency and other aspects. Below, we followed the structure to review the use of human simulators and related problems in WoZ simulations.

Simulators are the experimenter who proposed the study and developed simulation tasks and scenarios. Therefore, they did not require the dedicated pre-study learning session because that was usually done in the system development (Grill, Polacek & Tscheligi, 2012). Later, to meet the extensive use of WoZ in various applications (e.g., augmented reality-based user interfaces, (Alce, Hermodsson & Wallergård, 2013)), amateur simulators are increasingly recruited with extra pre-study training. By pre-study training we mean the purposeful practices before the simulation. Both the amateur and the experienced simulators received pre-study training, it is ensured that the simulators know what and how do to accommodate end user actions (Chandler, Lo & Sinha, 2002). Up to present, few systematic evaluation of simulator in learning results are reported in existing WoZ studies while most research adopted subjective self-reporting as the indicator of learning outcomes (Li & Bonner, 2014).

Secondly, simulators act as the interpreter who aims to translate users' inputs in a subjective manner. In this role, the simulators perceive end user speech and other inputs such as body gestures and movements and use them to extract motivations and behavioral meanings of the end user (Schlögl, Doherty & Luz, 2015). Multimodal user interaction presents the simulators with additional challenges in interpreting an end user's inputs; they must spontaneously synchronize all information, for example, identifying the end user's inputs from a collaborative multi-user task (Vahdat, George & Serna, 2013). Alternatively, the multiple simulators can also be accepted, however, it introduces additional complexities in cooperative simulation management (Fitch, Bowman & Llaneras, 2014).

Thirdly, simulators act as the decision makers who determine which end user input requires a response. This third process involves emotional and persuasive impacts, thus causing simulators difficulty in maintaining objective throughout simulation studies (Adler, Iacobelli & Gutstein, 2016). For example, a simulator can hardly notice his/her reaction preferences and patterns (Deshmukh, et al., 2013). In addition, other factors may also influence the simulator in decision-making processes, including human-computer courtesy (Baum, 2014), the forms of uttered speech (Dahlbäck, Jonsson & Ahrenberg, 1993) and the interaction channels (i.e., speech and gestures for combined interactions have an effect on simulators' speech punctuation and word choices (Thomason & Litman, 2013)).

Finally, simulators act as the facilitator (or the operator) during the process of operating simulation tools. This process involves the most reliable role of simulator since the simulation tools usually are developed with acceptable usability, efficiency, and longevity of simulations (Biswas & Murray, 2013).

To sum up, human simulators play multiple roles in WoZ simulations, mainly as the simulation tool learner, user input interpreter, simulation decision maker, and simulation tool operator. Each of these roles has a distinct possibility of producing improvisational responses that likely lead to undesired simulations. Given the adoption of automated simulation tools, the risks of improvisational operations have been largely prevented whereas the risks are caused by the simulator's individual perception and cognition which are less addressed (Beuzekom, et al., 2010).

Despite the focus on the behavior of the end users, previous studies also provided some information on the simulator performances and challenges. For example, the human wizard's tasks are highly demanding in terms of response times, system behavior, and consistency (Schlögl, Doherty & Luz, 2013). The tasks of simulator are multidisciplinary which concerns visual language use, multi-task allocation, and simulation task automation (Li, Hong & Landay, 2007). Simulating multimodal applications requires greater cognitive load on the amateur simulators (Chandler, Lo & Sinha, 2002). The simulations can be classified in two types: the end-user response and the tool operation. The former is more likely to incur improvisational (or instinct) simulations, as the latter concerns only the intentional executive activities (Serrano & Nigay, 2009). Therefore, the simulator mimetic behaviors bear innate risks of inconsistency and non-rigorousness.

2.3 Automated simulation tools

Automated simulation tools (e.g., automated selection tools) are the implements that can help human simulators facilitate WoZ simulation operations in an automatic manner (Gandhe & Traum, 2014). Specifically, the automated simulation tools execute human simulator simulation in decisions by following a set of predefined operations (Fabbrizio, Tur & Hakkani-Tür, 2005). These tools do not help the simulators process the end user inputs or make autonomous decisions since those processes rely on the simulator individual background of knowledge, observations, understanding, and behavior predictions. For example, the OpenWoZ client user interface was intuitive and useful (Hoffman, 2016). But simulators are still required to acquire a solid knowledge of human-robot interaction before the simulation rather than the tools support the processes of simulation operation. It provides simulators with additional confidence that helps to maintain high level perceptions and judgements (Steinfeld, Jenkins & Scassellati, 2009). These simulated operations are particularly useful for amateur simulators, because using automated simulation tools often results in high usability and operational efficiency while minimising the requirements to produce simulation experiences. For example, Hudson, et al. used such a tool to develop a range of ambient sensors (Hudson, et al., 2003).

The automated simulation tools have advantages. These tools are learnable and physically undemanding which help simulators maintain cognition performances. Also, the tools provide equality across different simulation functions regardless of the function in complexities (Mok, et al., 2015). It helps to ensure simulation productivity and efficiency. Furthermore, the automated simulation tools have largely reduced pre-study training and simulation working loads, which extend the adaptability of human simulators in different simulation scenarios. The automated simulation tools help simulators facilitate multimodal responses (Thomason & Litman, 2013).

Previous research has explored the use of automated simulation tools in various contexts, placing a strong emphasis on the tool of effectiveness during use (Riek, 2012). Most methodological understandings are in the field of usability, generality, and efficiency for the end users (Habibovic, et al., 2016). For example, Katz, Basis and Shtub (2015) used an improved platform called the “Wizard of Oz telemedicine simulator” to present learners with a risk-free environment for transferring medical knowledge, and Taib and Ruiz (2007) took usability into account in deploying multimodal interfaces. In separate-wizard studies, research has focused on the development of distributed control panel designs rather than on how those designs affected information synchronisation among simulators (Li & Bonner, 2014).

Previous studies have provided various implementations of automated simulation tools. For example, a WoZ sketch recognition system was implemented using structured graphs and symbol alphabets (Bott & Laviola Jr, 2015), and automated simulation modules were adopted to mimic language-technology applications (Schlögl, Doherty & Luz, 2015). However, given the outcomes of these studies, there is still a gap between the understanding of the relationships of automated simulation tools and the simulator’s mimetic behavior especially in the process of end-user response.

2.4 Requirements for the simulation by human simulators

WoZ configuration is relatively easy for the designer to use. However, WoZ has strict requirements for the simulator, which involve not only the simulator's experience and knowledge of the system but also the simulator's mimetic behavior throughout the use of automated simulation tools. These requirements serve for: (i) indicating the simulator's overall performances during WoZ simulations and (ii) measuring the mimetic behavior of simulators. Therefore, we reviewed the requirements based on a wide range of WoZ applications and studies and used these as criteria to reflect mimetic behavior of simulators in the following study.

The simulation using human simulators requires credibility (which refers to the level of how convincing the simulated technology or system is to end users). In previous studies, the simulators ensured the credibility by introducing intentional typos when mimicking a speech-recognition typewriter, so that the end users would believe that the system is real and had minor flaws (Gould, Conti & Hovanyecz, 1983). To some current simulators still follow this requirement, for example, by responding to the intentional speech recognition mistakes (Ashok, et al., 2014).

Credibility is fundamental to the WoZ study. It is because when the end users are aware of the simulated illusion, their attitudes and activities are affected, thus destroying the reliability and validity of the study. However, interestingly, the credibility is bound more to the end users than to the simulators (Schlögl, Doherty & Luz, 2015). For example, when a simulator provides an improvisational operation that is not sensed as such by the end users, the credibility of the simulation is not impaired. Even when a simulator provides the strict simulations, in case the end users have any doubts about it, the overall credibility of the simulation is low. The credibility is not an independent requirement; rather than it requires assistance from other measures when examining the simulator's mimetic behavior.

Simulations using human simulators require rigorousness. Rigorousness refers to the extent to which the simulator complies with the proposed system's input and output functions (Markopoulos, et al., 2008). It is challenged by increasing the cognition load in a simulation as well as the advancing technology levels that must be mimicked, such as the human-like robots (Hoffman, 2016). Basically, the simulators' mimetic behavior need to reflect the intended system in terms of reaction speed and other patterns (Bott & Laviola Jr, 2015). For example, a simulator must persist in responding with the same mistakes throughout a simulation.

Rigorousness is concerned with two behaviors of the simulator. From a behavior perspective, rigorousness involves how the simulator (as the desired computer system) uses simulation tools; from a perceptual perspective, it involves how the simulator (as an expert) interprets end users' inputs (Dahlbäck, Jonsson & Ahrenberg, 1993). Overall, rigorousness measures how accurately the simulator behaves in the desired role.

The simulations using human simulators require consistency. Consistency is the most important - and the most vulnerable - component in the WoZ study since it reveals the alignment of the responses across multiple simulators and across the longitudinal simulation processes of simulator (Li & Bonner, 2011). Consistency differentiates from the rigorousness: the former emphasizes the logical coherence across the simulator on overall simulation performances as well as across mimetic behavior, while the latter focuses on the accuracy with which the simulator mimics the desired system.

Consistency is linked to specific simulation tasks, for example, when simulating a hand-gesture recognition system, the simulator experiences less physical fatigue than when simulating a speech-recognition system due to the different response frequencies required (Forbes-Riley & Litman, 2011). Despite the efforts involved in various physical demands, a simulator's cognition load is consistently heavy. One reason for this is that simulators must not only be aware of their current response but must also consider their prior responses (Schlögl, Doherty & Luz, 2015). Another reason is when multiple simulators facilitate individual simulation threads (Alce, Hermodsson & Wallergård, 2013). This simulator separation reduces the complexity of simulation tasks, such as the 'NEIMO' project, which employed three simulators for speech recognition, face recognition, and mouse control, respectively (Salber & Coutaz, 1993). However, simulator separation also introduces challenges in coordinating multiple simulators' simulations.

The simulator-enabled simulations require efficiency. WoZ studies involve many pseudo-functionalities and simulation tools that are often undergoing iterative design-evaluation-design cycles. Therefore, the simulator's simulation needs to be well-organized and competent. The emergence of general WoZ platforms and longitudinal simulation tools will cause an increasing emphasis on simulation efficiency in future WoZ studies (Grill, Polacek & Tscheligi, 2012).

2.4 Lessons learned

The preceding review accentuates the lack of understanding of human simulators' mimetic behavior with the use of automated simulation tools in WoZ studies. Few previous studies have systematically investigated the influence of automated simulation tools on human simulators with various simulation experience. Additionally, previous studies usually adopted WoZ to probe advanced systems rather than focus on the simulators while the difficulties they are confronted during simulation. For instance, the simulator training and recruiting, study configurations and data collection techniques are good examples. In this regard, previous studies leave many questions unanswered: how the simulators can take advantage of automated simulation tools to provide credible, consistent, rigorous and efficient simulations regardless of simulation experience beforehand, and how the automated simulation tools would affect simulator behavior when facilitating end users' inputs in terms of simulation consistency and rigorousness.

To draw insights into these questions, we developed a set of automated simulation tools and put these in laboratory studies to investigate how the automated simulation tools affected simulators' behaviours of tool operations and end user response. The significance of the study is multi-fold. First, it uncovers the correlations between the simulator and the automated simulation tools. Also, it would be useful to understand the relationships between simulators and the end users. Secondly, the understanding of simulator's mimetic behaviours reveals how the spoken language interface-related features e.g. speech duration, pitch and energy would affect simulators' perception and responses. The features can indicate simulators' psychological and cognitive statuses. Finally, the study identifies both the strengths and the weaknesses of the simulators in iterative simulations.

2.5. Hypothesis development

Despite known benefits of automated simulation tools, state-of-the-art WoZ studies e.g. (Katz, Tepper & Shtub, 2017, Katz & Shtub, 2016) indicate that the problems of inconsistency and non-rigorousness remain in WoZ simulations. Taking account of the roles of human simulator and the use of automated simulation tools, we propose the study hypotheses as follows:

H1: Given the simulators' experience levels, the automated simulation tools cannot provide the same credible, consistent, rigorous, and efficient simulations.
H2: The automated simulation tools cannot fully support human simulators on end-user response and tool operation similarly with respect to simulation consistency and rigorousness.

3. Method

The objectives of this study were to investigate human simulators' mimetic behaviours when using automated simulation tools. Because it is difficult to conduct an exhaustive study that directly examines the simulator's experience and characteristics, we adopted a comparative exploratory study using two of the most common types of simulators (the amateur and the experienced simulators).

The simulators usually mimic a human-like spoken dialogue system in this study. Here, the 'human-like spoken dialogue system' is proposed for several reasons. One is that current speech-recognition technologies are approaching natural human language capabilities; thus, prototyping a pseudo-system that is slightly more advanced than current technologies appeared more convincing to end users. Another reason is that making simulators mimic a machine such as a computer does not capture the simulators on full perceptions and responses as human beings. In contrast, to successfully mimic a human-like spoken dialogue system, the simulators had to behave as rigorously as a computer but also to mimic the system as if it is a human. Thus, the simulators exposed activities that involved simulation task perception, user interpretation, and self-operational awareness. An extra advantage of human-like spoken dialogue interfaces is that they exert minimal influence on experienced and novice simulators on spoken-language capabilities. That is, the simulators of greatest differences lay in their prior experience with the simulation tools instead of their spoken-language abilities.

The automated simulation tools cannot be precisely manipulated in comparative studies. Thus, we set the human simulators who come with different simulation experience as the pseudo independent variable. In other words, we need to investigate the automated simulation tools by comparing their influence on the experienced and amateur mimetic behavior of simulators. The experience levels are configured by recruiting and training simulators with different simulation experience. The dependent variable is the mimetic behavior of simulators (or the simulation performance), which are being measured with the metrics of credibility, consistency, rigorousness, and efficiency. Other variables, such as the simulation tools, study tasks, study scenarios and procedures, are the same to all the simulators in the study. This section describes the methodological details of the study.

The study involves multiple experiments to measure the variables. It aims to investigate two groups of simulators (the experienced and the amateur), each group is measured with the credibility, rigorousness, consistency, efficiency, and overall engagement, and satisfaction. The study planned to recruit 60 simulators and 120 end users in total and each simulator needed to facilitate two individual end users. Therefore, the study can compare the mimetic behavior between- and within-simulators.

3.1. Participants

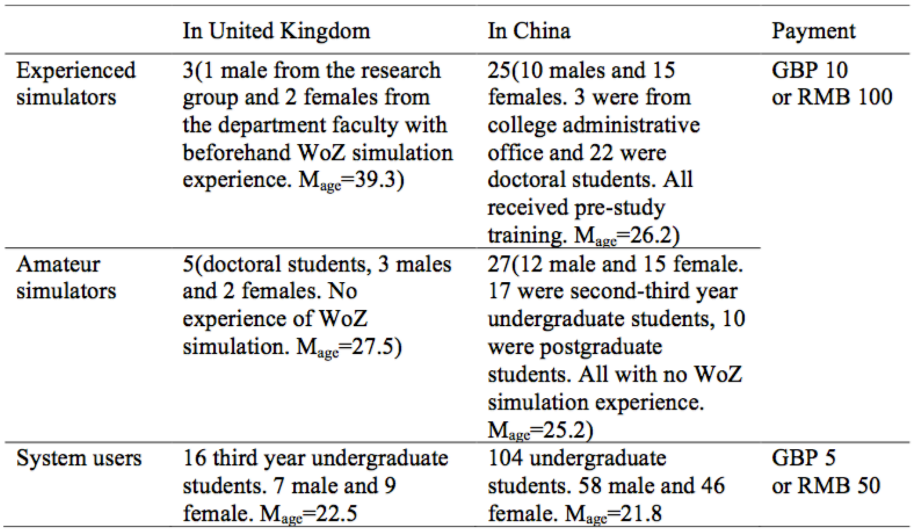
The study evaluated the mimetic behavior of human simulators with the automated simulation tools. The actual participants are the human simulators rather than the end users of the simulated system. Overall, the simulators could be categorized into three general groups according to their simulation experience levels from high to low: the experimenters, the trained experienced simulators, and the amateur simulators. The WoZ experimenters and the trained simulators are expected to have the same amount of simulation experience, therefore, the two participant groups are combined into the experienced simulator group. Finally, we shaped two main participant groups: the experienced and amateur simulators.

The scales to categories the simulators comprise (1) knowledge of the WoZ method and the role of simulator (wizard), (2) experience of simulating a specific WoZ system and (3) skills of a simulation tool operator. As such, the experienced simulators need to meet all scales simultaneously, while the amateur simulators need to meet the scale (3) without the other two scales. We provided extraordinary training and pre-study practice to the experienced simulators to ensure they met the scales (2) and (3), respectively. In contrast, the amateur simulators only received the pre-study practice to ensure they gained skills of tool operation.

Sixty participants are publicly recruited as the simulators. Specifically, eight of these simulators are based in the United Kingdom, where the first half of the study is conducted, and fifty-two participants are in China, where the second half of the study is conducted one year after the first study. The participants could use Chinese in the study while they all reported similar English language abilities as those in the first half study. To manage the cross-social-culture issue, we applied measures e.g. double-checking the simulators' dual language communication skills and proposing less cultural-dependent task scenarios e.g. weather and shopping.

The simulators are selected against the same criteria and the cultural background is not the necessary dimension of criteria. The bicultural backgrounds added diversities of the simulators. The details of the simulators are summarized in Table 1.

Table 1. The simulators and end-users in the study



	In United Kingdom	In China	Payment
Experienced simulators	3(1 male from the research group and 2 females from the department faculty with beforehand WoZ simulation experience. M_{age}=39.3)	25(10 males and 15 females. 3 were from college administrative office and 22 were doctoral students. All received pre-study training. M_{age}=26.2)	GBP 10 or RMB 100
Amateur simulators	5(doctoral students, 3 males and 2 females. No experience of WoZ simulation. M_{age}=27.5)	27(12 male and 15 female. 17 were second-third year undergraduate students, 10 were postgraduate students. All with no WoZ simulation experience. M_{age}=25.2)	
System users	16 third year undergraduate students. 7 male and 9 female. M_{age}=22.5	104 undergraduate students. 58 male and 46 female. M_{age}=21.8	GBP 5 or RMB 50

None of the simulators self-reported any visual or physical impairments that would possibly constrain WoZ simulations. Multiple scales are adopted to categories the simulators. It includes the recruitment of the simulators which are justified for multi-fold reasons:

(i) Most of mimetic behavior of the simulators are anticipated to occur during the processes of interpreting user input and simulation decision making, which are independent of the number of participants. In other words, end users' interactions are largely confined within a predictable range with respect to interactions and speeches. A larger number of end users would not significantly expand this range;

(ii) The simulation tools of simulators have predefined functionalities which means the mimetic behavior of simulators are limited to within a general range. The size of this range is not dependent on the number of simulators but on the tasks they perform;

(iii) The responses of simulators are followed study scripts, which also makes a larger number of simulators unnecessary;

(iv) Training and recruiting a large number of highly experienced WoZ experimenters is time-consuming, in contrast, a small number of carefully selected expert simulators can be equally representative for an exploratory study. The selection criteria includes the following. (1) Experienced simulators who had to have prior experience in mimicking natural spoken language systems. This requirement ensured that the simulators would have a good knowledge of simulation systems. (2) Experienced simulators must have had simulation experience in the prior half year. This requirement ensured that the simulators would have fresh simulation skills and experience. (3) Amateur simulators must not have been involved in any forms of WoZ simulations before. This requirement ensured that the amateur simulators had no previous experiences. (4) All simulators, including both amateur and experienced, need to be capable of intensive simulations such as responding over a long period. This requirement ensured that the simulators are capable of operations under stress. (5) All simulators received 5 min of pre-study practice time; any simulators who are unable to learn and operate the automated simulation tools during that period are removed. This ensured that the remaining simulators would have a similar level of familiarity with the simulation tool.

As described in Table 1, one hundred and twenty undergraduate students are recruited from the local university to play the end users of the simulated system. They are not informed about or involved in any means of system simulation but only to use the system and then provide feedbacks. These end users needed the basic knowledge and experience of spoken language applications such as the Siri. The requirements for the end users on the previous experience included (1) the familiarity to the form of human-system interaction via natural spoken language and (2) using the spoken language-enabled system to complete a mundane task such as weather information enquiries. The end users signed a consent form before the study, but they are not informed about the simulation until the end of the study.

In addition, an instructor is presented during the study, passing and collecting task sheets and questionnaires and organising post-study semi-formal interviews. The instructor, unless explicitly requested by the end users, does not interfere in the simulators' and the end users’ tasks. The instructor is allowed only to provide technical support, without any forms of task completion guidance.

Another three human-computer interaction researchers are publicly recruited to inspect usability of the simulation tool before it is used in the formal studies. The researchers walk through the simulation tool’s functionalities (e.g. text typing and operation clicking) and reported potential usability problems. The simulation tool has improved through iterative designs until no more major usability problems are reported.

3.2. Apparatus

The study implemented a distributed WoZ simulation system that consisted of two components in separate rooms: the automated simulation tools and the simulated applications (Figure 1). The two components run on the separate host computers but are connected through intranet protocols and ports. The distributed structure helped to circumvent the risk that end users would sense the simulators.

The automated simulation tools consisted of a set of control panels displayed on a 22-inch Wacom Cintiq touch-sensitive monitor (Figure 2). The simulation tools and the monitor supported multimodal inputs, including pen, keyboard, mouse, and speech recognition. The simulators, including the amateur and the experienced, are instructed to use the pen and mouse to select the preset operations and use the keyboard to type additional operation messages. And the speech recognition is the alternative to the keyboard as it allowed the simulators to type the response message by speaking. The automated simulation tools have speech synthesis capability for delivering the acoustical responses along with the display of text messages.

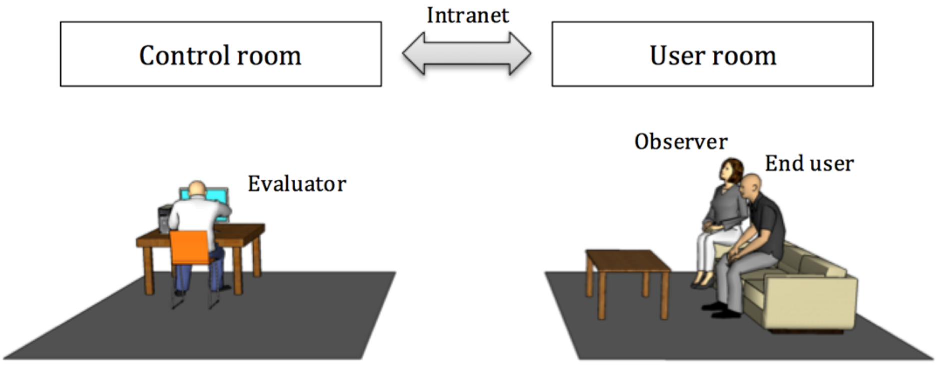
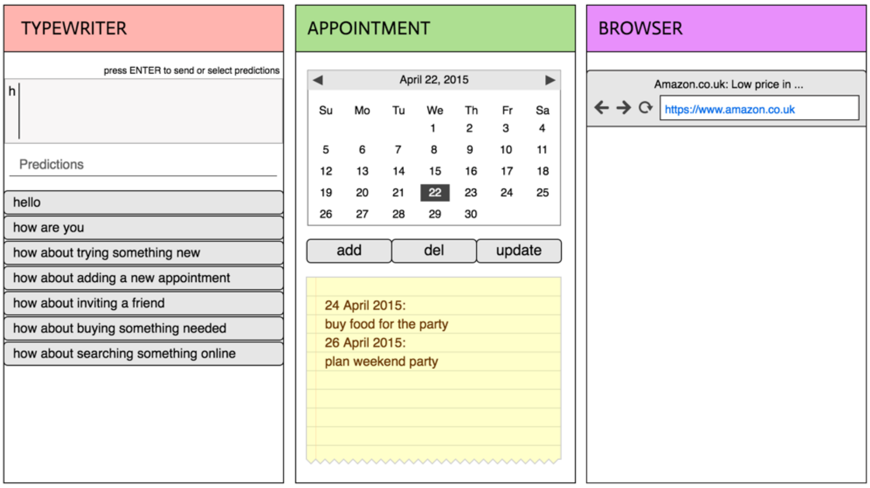


Fig 1. Distributed simulation system of WoZ simulation

Fig 2. Multi-touch and pen-input capable monitor with mouse and keyboard

The automated simulation tools comprise a predictive typewriter, a calendar integrated with an appointment manager, and a customised web browser (Figure 3). These components are specifically optimized for automatic operations. First, multimodal input methods are enabled. Each component is capable of at least two input modalities. For example, the typewriter accept the keyboard input as well as the spoken language speech. The calendar accepted pen and mouse input; and the browser accepted natural language speech and all the other modalities. Second, automated operations are integrated after the simulator selected a simulation decision. For example, when the simulator said the word ‘balloons’, the typewriter transform the speech into text and showed a list of dynamic candidate simulation decisions such as ‘make a new appointment to buy a balloon’. After a decision is selected, the simulation tools automatically performed tasks such as launching the calendar and displaying the appointment interface to the end user. Finally, simulators are able to customize the tools layouts for their preferences. A simulator can drag these components around the monitor.



hello
how are you
how about trying something new
how about adding a new appointment
how about inviting a friend
how about buying something needed
how about searching something online

Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

24 April 2015:
buy food for the party
26 April 2015:
plan weekend party

Fig 3. Components of the automated simulation tools

The simulated application is a smart spoken dialogue system projected on a coffee table by a ceiling-mounted projector (Figure 4). A webcam affixed to the front side of the projector sends a live surveillance video stream to the simulator in the other room. The simulated application is capable of speech recognition and displayed simple graphics and texts according to the response of simulators.

The application took place in a simulated domestic scenario with one long sofa and a coffee table in front of the sofa (Figure 1). The simulated application accepts end users speeches and displays visual results on the coffee table. The application is introduced to end users as a system with human-like speech-recognition capabilities.

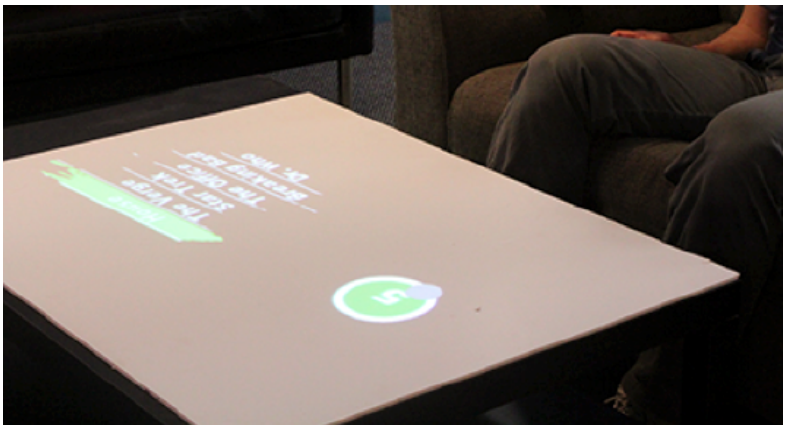


Fig 4. Simulated application projected on the coffee table

3.3. Procedures

The study has consisted of one hundred and twenty simulations in total; each simulator (60 simulators, including 28 experienced and 32 amateur simulators) simulated the spoken-language dialogue system twice, each time with a single different end user (correspondingly, 120 end users in total). The "one simulator and two end users" approach aims to reflect the simulator's performances when simulation experience increased. By contrast, arbitrary comparisons between the individual simulators are impractical because the simulators had various perception and motion abilities in simulation tasks.

Pre-study configurations are conducted spontaneously in two separate rooms. In the control room, the experimenter gave the simulators a brief introduction to the simulation tools and simulated applications and guided them as they walked through the study scripts, which provided a list of simulations corresponding to specific user inputs. The experienced simulators in both the first and second half of the study received an extensive pre-study training. The training lasted approximately one hour, consisting of simulation tool introductions and rehearsals. This training ensured that all experienced simulators had a high level understanding and skills of the automated simulation tools.

Subsequently, each simulator underwent the practice session with the automated simulation tools. During the practice session, the simulators needed to learn to customise the layouts of the simulation tool, use the pen, keyboard, mouse and speech recognition to operate the simulation tool, and use the interaction methods to control the simulated applications. The practice contents can be done multiple times within the time limit.

The practice is for two purposes: practicing the simulation tool and inspecting its usability. The simulators are asked to qualitatively report potential usability problems during the learning of simulation tool. The inspection measured the overall usability of the simulation tool and it aims to prevent the usability problem of the selected WoZ automated simulation tools influencing the study results. Also, it ensures that the simulators' performance is not affected by any usability problems that are caused by the designs of simulation tools. The practice session involved no real end users, as it is intended only to familiarize the simulators with the system. In the end user's room, each user is seated on the sofa and read through an introduction to the pseudo-system under the instructor's supervision. The instructor played a 5-min video clip demonstrating the use of the system before the practice session officially started. The overall configuration flows of both rooms are illustrated in Figure 5.

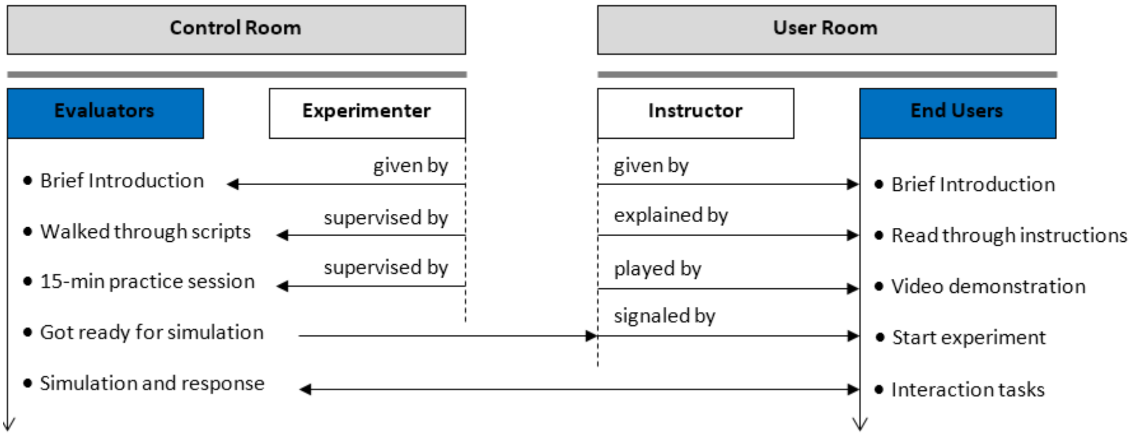


Fig 5. Procedural flows of pre-study configurations in both rooms

The formal procedures after the pre-study configurations are as follows.

First, the simulator is required to finish the practice session before the end user's learning session began; thus, they would be prepared for the simulation. When the simulator is ready, he/she used the simulation tools to signal the application to display the splash interface.

Second, the instructor noted the interface and then guided the end user to start the tasks. The tasks consisted of:

(i) it asks the spoken dialogue system about the weather over the next weekend,

(ii) it asks for the spoken dialogue system's assistance in making a shopping list for a weekend party,
(iii) it asks the spoken dialogue system to invite friends to the party.

The three tasks had incremental complexity levels. As previous studies indicated, complex simulation is more cognitively demanding, including longer response time and higher chances of improvisational behavior (Deshmukh, et al., 2013). In this study, the influence of the complexity is mostly on the simulator's end user response rather than on his/her automated simulation tools operations.

These are open tasks that are intended not only to encourage the end users to try different interactions but also to remove unnecessary pressures on the end users. During the tasks, the end users talk to the spoken dialogue system which inhabits in the coffee table and asked the system for technical assistance to complete the tasks. At the same time, in the other room, the simulators are instructed to act as a human-like spoken dialogue system by interpreting the end users' speech and deciding how the system should be responding to the end users. The simulators and the end users took part in the study in random orders. The simulators have a 15min breaking after each simulation study.

Need to note, the simulators and the end users recruited in China are allowed to use Chinese in the stud, and the system commands and related spoken dialogue responses are translated accordingly. This measure ensured the naturalness of the interaction between the simulator-mimetic system and the end users. Furthermore, it circumvents unnecessary risks possibly incurred by language barriers. The study data are translated back to English in the analysis. Also, previous studies such as (Fraser & Gilbert, 1991) had proved the effectiveness of WoZ simulations in multilinguistic context.

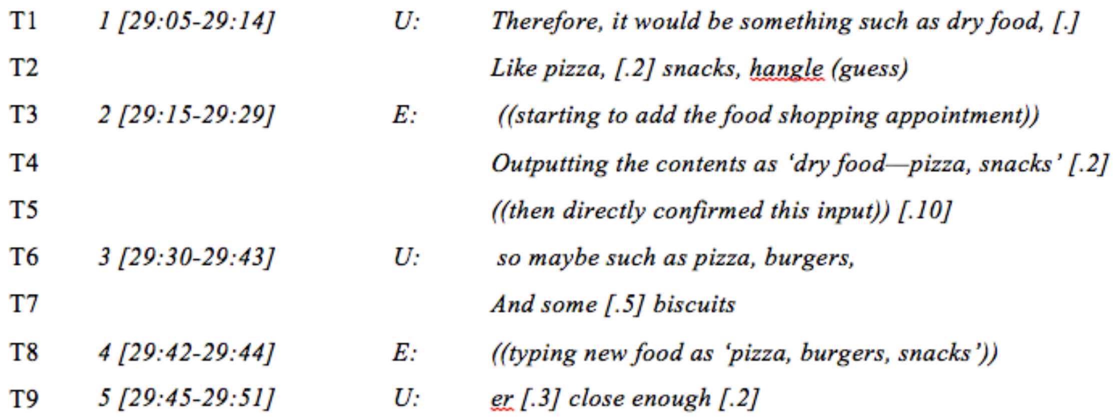
Finally, after completing all tasks, the end users are instructed to fill a 5-point Likert-scale questionnaire (Appendix A) and then the study instructor conducted semi-formal interviews (the questions are provided in Appendix B). All the simulators and the end users followed the similar procedural flows. In each study, they include the pre-study configurations – which is done for only once at the beginning of the study, simulations tasks, post-study questionnaires and interviews, took approximately 7-min (M simulation time=6.54, SD simulation time=2.10).

4. Analysis

The study collected 120 questionnaires, interviews, end-user surveillance videos, and simulator surveillance videos, respectively. The simulators does not take part in the questionnaires or interviews because they are often unaware of their mimetic behavior during the simulation (Alce, Hermodsson & Wallergård, 2013). To measure the simulators' mimetic behavior, we adopt the four WoZ simulation requirements – credibility, consistency, rigorousness, and efficiency – as the criteria, as these requirements are the direct indicators of simulators' mimetic behavior (Steinfeld, Jenkins & Scassellati, 2009).

The study data are processed as the following. Firstly, the hard copy questionnaires are transferred to a database for the later statistical analysis, and the interview feedbacks are manually annotated by the experimenters and organize in the form of keywords. Secondly, both the end users and the simulators surveillance video footages are processed by a speech-to-text programme that implemented the Google cloud speech API, so to transcribe all the 'simulator - end user' conversational interactions into texts ordered by ascending video time stamps. And then, ten experimenters manually inspected the transcribed texts, corrected recognition errors (unrecognizable speeches are marked as 'guess'), and inserted additional annotations into the texts (see the contents in double brackets). Finally, 87730 transactions (each transaction is annotated with a number, as shown in the following example) are collected. The study analysis used descriptive statistics and non-parametric tests e.g. normality distribution test and Mann-Whitney test to depict the overall study results and the differences between the experienced and amateur simulator groups, respectively.

A transaction is seemed as an individual line of simulators' and end users' spoken dialogues as follows, each of which lasts for several seconds and has a specific intent. The structures and marks of the transactions followed the conventional conversation analysis formats such as in (Matthews & Heinemann, 2012, Seedhouse, 2004). One of the transactions captured in the study is shown below.



T1	*1 [29:05-29:14]*	*U:*	*Therefore, it would be something such as dry food, [.]*
T2			*Like pizza, [.2] snacks, hangle (guess)*
T3	*2 [29:15-29:29]*	*E:*	*((starting to add the food shopping appointment))*
T4			*Outputting the contents as 'dry food—pizza, snacks' [.2]*
T5			*((then directly confirmed this input)) [.10]*
T6	*3 [29:30-29:43]*	*U:*	*so maybe such as pizza, burgers,*
T7			*And some [.5] biscuits*
T8	*4 [29:42-29:44]*	*E:*	*((typing new food as 'pizza, burgers, snacks'))*
T9	*5 [29:45-29:51]*	*U:*	*er [.3] close enough [.2]*

T-transaction, U-end user, E-simulator (evaluator)

The simulators for surveillance videos are not transcribed because the previous transcriptions already include the simulators' speeches. These videos are used to validate the findings from the end users on the conversation analysis.

The above study data are analyzed with multiple methods. The questionnaires are quantitatively analysed through statistical analysis. The interview notes are qualitatively analysed to extract the end users' feedbacks. The speech transcriptions are analysed using the conversational analysis method, which is a proven method of conversation analysis (Matthews & Heinemann, 2012, Seedhouse, 2004). The formats and symbols in the speech transcriptions mostly complied with this method. In addition, the simulators' surveillance videos are analysed by the expert-walkthrough method. The requirements for expert-walkthrough include: (1) comparing the responses of simulators with the end users on the inputs and (2) analysing the behavior of the simulators according to the study script and (3) validating the simulators' performances against the end users' qualitative feedbacks. Table 2 summarises the criteria for the simulators' mimetic behavior.

Table 2. The criteria for the simulators' mimetic behaviour

Criteria	Measures	Methods
Credibility of simulation	•End users' awareness of simulation •End users' feedbacks on how convincing of the system functionalities	Questionnaires, Interviews

Rigour of simulation	•Simulators' mean response speed(behave) •Simulators' response styles •Overall conformity with presumed system defined by study scripts	Conversation analysis, Expert-walkthrough
Consistency of simulation	•Simulators' behaviours in long- and short-term simulations •Simulators' behaviours in predicting end user input time and patterns	Conversation analysis, Questionnaires, Expert-walkthrough
Efficiency of simulation	•Simulators' and end users' overall response speed throughout simulations •Response rhythms made by the simulators	Conversation analysis, Questionnaires, Interviews
Overall engagement and satisfaction in simulation	•Overall engagement perceived by end users •Overall satisfaction perceived by end users	Questionnaires, Interviews

4.1. The credibility of the simulations

Credibility is the foundation of the WoZ simulations because it has a direct influence on overall reliability and validity of the study. Therefore, we firstly analysed the credibility of the simulators' responses as it is perceived by the end users. Given the analysis results, the simulations that did not have sufficient credibility are invalid and therefore removed from the next analysis.

The credibility cannot be directly measured by analyzing the simulator responses because the simulators on the self-report could be unwittingly biased. Oppositely, the study adopted questionnaires and semi-formal interviews to collect the end users on the feedbacks on the overall simulation credibility. Two questions (Appendix A, Questions 1 and 2) and one interview question (Appendix B, Question 1) are used to reflect the credibility in the study.

On questionnaire A Question 1, 81.7% (98 out of 120) end users gave the pseudo-system a 5-rating and the rest gave it a 4-rating (M=4.30, SD=0.46). On questionnaire A Question 2, 70.80% (85 out of 120) end users gave a 5-rating and the rest gave a 4-rating (M=4.18, SD=0.39). The results indicate very positive results of the credibility.

The results of the questionnaire A Question 1 and 2 are not in normality distributions (one-sample K-S test denied normality distribution, pQ1=0.00, pQ2=0.00). Therefore, we use non-parametric Mann-Whitney test to compare the results between the amateur and the experienced simulators. As Figure 6 shows, the results reported no significant differences between the two simulator groups (Question 1: U=1088.00, Z=-5.52, p=0.10. Question 2: U=672.00, Z=-7.48, p=0.07), indicating that the simulators' experience levels had no influence on the end users' perceptions of the credibility.

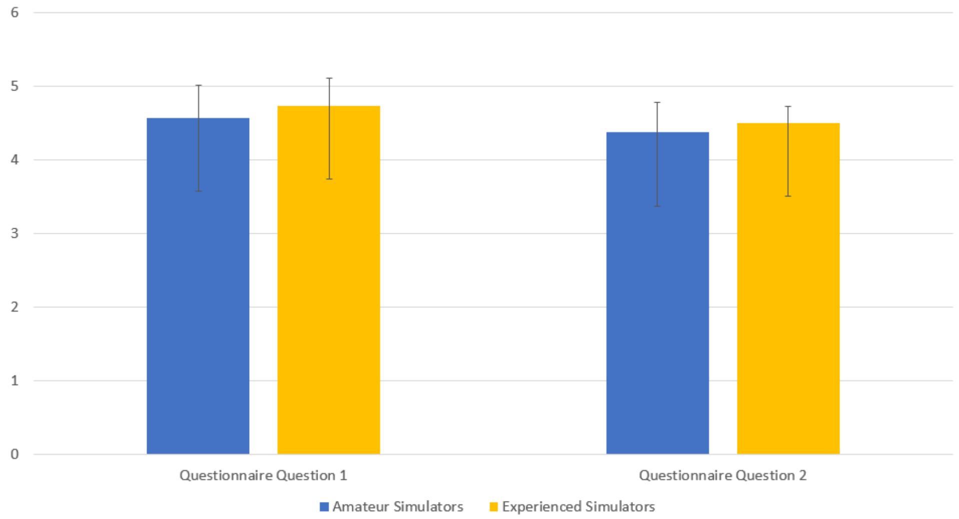


Fig 6. The results of credibility between the experienced and the amateur simulators (no significant differences in Question 1 or Question 2)

The results of the credibility between the two simulator groups are generally high while most end users gave the highest rating. For those who gave low ratings, the semi-formal interviews revealed some reasons. First, none of the end users doubted the simulated system. Secondly, some end users give low ratings because they thought the spoken dialogue system should have had more advanced functions such as "[the end user] using the system to control smart home devices". Overall, the results showed no clues that the end users are aware of the simulation, therefore, the reliability and validity of the study are ensured.

4.2. Rigorousness of the simulations

The rigorousness of the simulators of responses is measured by the simulators of response speed and styles. The former refers to the time between the simulator receiving the end user's input and giving a response, and the latter refers to the simulators' utterance styles. We calculate the speed based on the time stamps in the transactions, as the simulators' every individual response is logged with the time. Furthermore, we analysed the questionnaire results to understand how the rigorousness of simulation is perceived by the end users. In the meantime, we adopted expert walkthrough method to identify the differences between the simulators of responses and the study scripts. The expert walkthrough examined the accordance of the simulators interpreting and responding to the end users' inputs.

First, we analyze the simulator speed of simulation (seconds per simulation, a simulation is defined as a collection of transactions between the simulator receiving the end user inputs and giving a response) between the simulator groups which is calculated using the following equation:

$$S=\frac{\sum T_{(0\ldots n)}}{N}, \qquad (1)$$

where S is the mean speed, N is the total number of simulations which might have multiple transactions, T is the duration of individual simulations, and n is the number of simulations that are manually counted by the experimenters.

We compared the amateur and the responses of experienced simulators with the first and the second end users, so to understand whether the current experience of simulator levels will affect the simulations. The simulators' average response speed is M=1.52, SD=0.64. To be more specific, the amateur simulators' response speed was like the overall speed (M=1.89, SD=0.44) and the experienced simulators' response speed was faster than the overall speed (M=1.14, SD=0.61). In Figure 7 shows, the results reported no significant differences between the amateur and the experience simulators when simulating with the first end user, (tested with no normality distribution, non-parametric Mann-whitney test U=50.40 p=0.15). Similarly, there were no significant differences between the amateur and the experienced simulators in the simulations with the second end user (tested with no normality distribution, non-parametric Mann-Whitney test U=14.00 p=0.38).

However, the simulators' response speed had significant differences between the simulations with the first and the second end users (tested with no normality distribution, non-parametric Mann-Whitney test U=69.2 p=0.032). This result indicated that the simulators' simulation efficiency had significant improvement when they gain more simulation experience. According to the results, we presumed that the simulators' interpreting the end users' input and making response decisions were improved when their simulation experience grew.

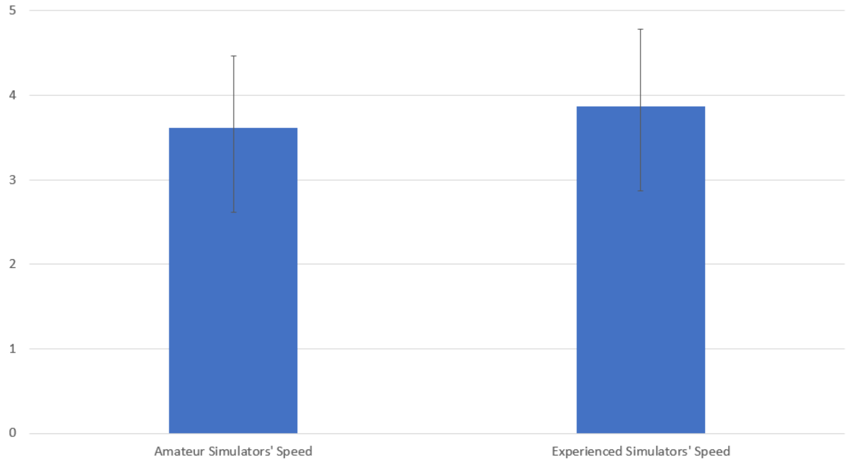


Fig 7. The results of response speed between the two simulator groups in the first and the second simulations (the simulations with the first and the second end users had significant differences)

To further understand whether the inherent simulation experience may influence the mimetic behavior of simulator, we analyze the two simulators of groups response speeds by measuring the end users questionnaire feedbacks. The questionnaire results (see Appendix A, Question 3) show that the end users' perception of different simulator groups' response speeds. The overall (M=3.75, SD=0.86, provided only for reference), the amateur (M=3.62, SD=0.85), and the experienced simulators' response speeds (M=3.87, SD=0.91) were compared (see the results in Figure 8). Independent-samples T-test reported no significant differences between these groups (tested with normality distribution, p＝0.28). The results indicated, despite the significant differences in the simulators of response speeds between the first and second simulations, the end users are not sensitive to the speed changes.

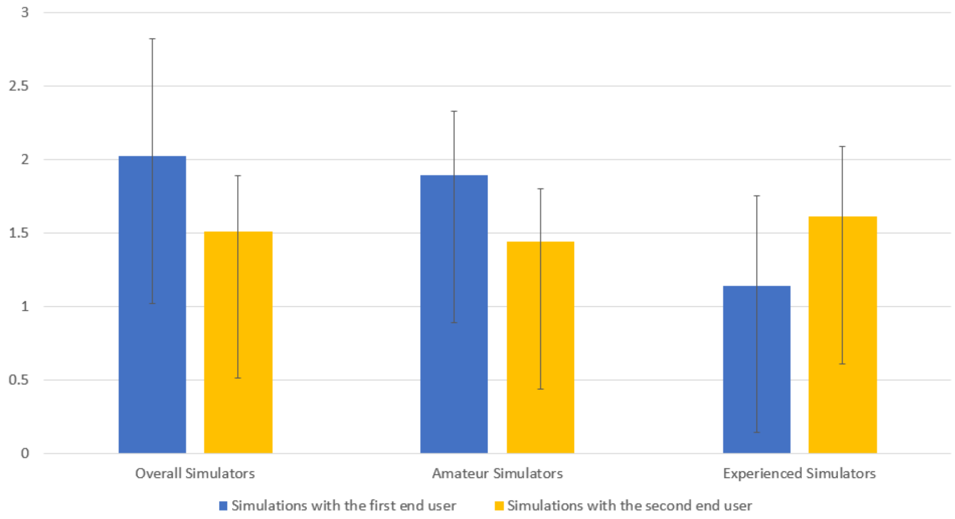


Fig 8. The results of the simulators' response speeds as perceived by the end users (no significant differences were reported between the amateur and experienced simulator groups)

Second, we analyze the simulator response styles among the simulator groups. The response styles of simulators specifically refer to three aspects in the study: the response speech rhythms, the preferred speech utterances, and the response speech pitches. The instructor explicitly explains these aspects when the end users answer the questionnaire Question 4. The questionnaire results, as shown in Figure 9, reveal no significant differences in any of the aspects of response styles (tested with normality distribution, independent-samples T-test p=0.33).

The interview analysis (Appendix B, Question 2) shed light on the understanding of the above results. Since the end users only participated in the study once, they cannot compare different simulators' responses. In addition, the end users felt that "the system's speeches quite adaptive and easy to understand". This quote indicates that: the end users understood and adapted to the system's speech styles and the system adapts to the end users of input speeches.

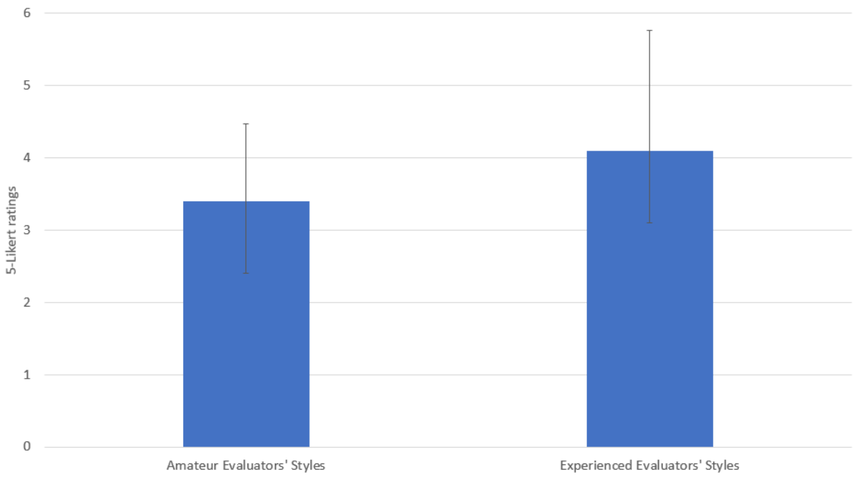


Fig 9. The results of the simulators' response styles (no significant differences between the amateur and the experienced simulators)

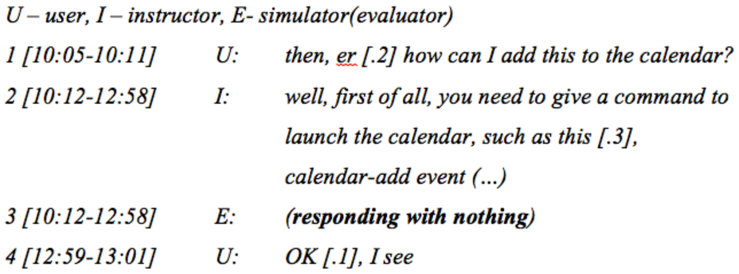
Finally, to understand whether the simulators would adapt to the end users of speech utterances, we develop a program to automatically take random transaction samples out of individual simulators' responses. Each sample is 30sec long. By adaptive mimetic behavior, we mean that the simulators responded with the same speech phrases and utterances as directly derive from the end user input. We removed the transactions that do not contain essential dialogues, and finally extracted 2551 transactions. We then ask seven experts (experimenters) to re-visit these transactions to validate if the simulators really leaned the end users speech for utterances in the simulations. The expert examination found several examples of non-rigorous mimetic behavior and the details are reported below:

(1) When facing unexpected user inputs such as 'OK, could you show me an example of what I am supposed to shop for', both the amateur and experienced simulators tend to respond with 'Sorry, I do not understand, please say that again'. The study scripts allowed such responses when the simulators encountered difficulties in understanding the end users' input. Within the selected data, 278 transactions included such "unrecognised input" responses (Nexperienced simulator=202, Namateur simulator=76). Independent-sample T-test reported significant differences in the numbers of the “unrecognised input” response between the amateur and the experienced simulators (tested with normality distribution, p=0.038).

(2) When the instructor is asked to demonstrate the use of speeches during the simulation, the simulators did not respond. For example, when the instructor explained how to add a shopping list appointment using speeches such as “to create an appointment, for example, you can use speech commands such as 'new appointment', or you could use 'make a new appointment'”, the simulators are expected to respond un-discriminatively, but the selected transcriptions showed that no simulators responded to the instructor's speeches.

(3) The simulators adapted to the end users on the interactions. The experts analyze the selected transcriptions and noted several changes in the simulator of responses. These changes include both speech utterances and response rhythms. For example, when the end users spoke slowly, the simulators respond in a more flexible way, and vice versa. The experts find that all (28 in total) experienced simulators and 21 (out of 32) amateur simulators exhibited such instinctive adaptations. However, in this regard, no significant differences are found between the amateur and the experienced simulators (tested with normality distribution, independent-sample t-test p=0.61).
In the following parts, we provide two examples of the simulators' instinctive adaptations. The non-rigorous behavior are highlighted in boldface.

Example 1:



U – user, I – instructor, E- simulator(evaluator)

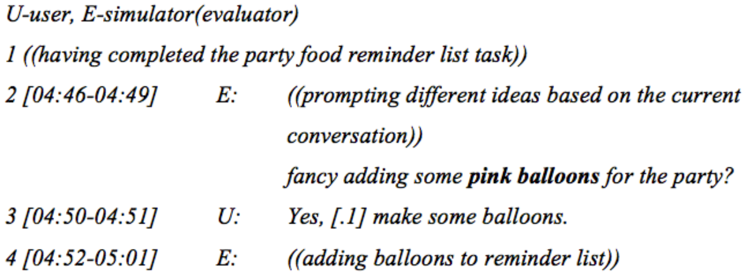
1 [10:05-10:11] U: then, er [.2] how can I add this to the calendar?

2 [10:12-12:58] I: well, first of all, you need to give a command to
launch the calendar, such as this [.3],
calendar-add event (...)

*3 [10:12-12:58] E: **(responding with nothing)***

4 [12:59-13:01] U: OK [.1], I see

Example 2:



U-user, E-simulator(evaluator)

1 ((having completed the party food reminder list task))

2 [04:46-04:49] E: ((prompting different ideas based on the current
conversation))
*fancy adding some **pink balloons** for the party?*

3 [04:50-04:51] U: Yes, [.1] make some balloons.

4 [04:52-05:01] E: ((adding balloons to reminder list))

(4) Therefore, the simulators made different decisions after making unintentional mistakes. The experienced simulators tended to return an unrecognized response when the end users attempted to repeat the speeches. In contrast, the amateur simulators appeared more likely to provide correct responses after the end users on the retrials. Given the limited transcription samples, no direct evidence is captured to prove any significant difference between the simulator groups.

Overall, the above analysis indicates that the amateur and experienced simulations of simulators include non-rigorous mimetic behavior with automated simulation tools and that, in several aspects, these behavior are significantly different. However, from the end user perspective, the non-rigorous mimetic behavior are not easily noticed, therefore, they have no significant influence on the end users perceptions.

4.3. Consistency of the simulations

By consistency in this study we mean how strictly the simulators could mimic the desired system. To understand the consistency of the simulators on the mimetic behavior, we analyze the simulations from individual simulators and simulations across multiple simulators. The former is measured through the questionnaires and interview analysis and the latter is measured through the conversation analysis. In addition, the expert walkthrough is adopted to validate the results.

First, we analyze how consistent the simulations of simulators are between the first and second end users. The questionnaire analysis (Question 5: Moverall=3.95, SDoverall=1.12, provided for reference; Question 6: Moverall=3.90, SD overall=1.05) showed some deviations across the simulations of simulators (see detail data in Figure 10) but reveal no significant difference consistency between the first and second simulations (Question 5: independent-sample T-test Pfirst and second simulation=0.48; Question 6: independent-sample T-test Pfirst and second simulation=0.09). The interview analysis (Appendix B, Question 3) show no hard evidence that supported any form of inconsistency, because all the end users are satisfied with the system. To sum up, we assume that the end users do not sense inconsistencies in the individual simulators' simulations.

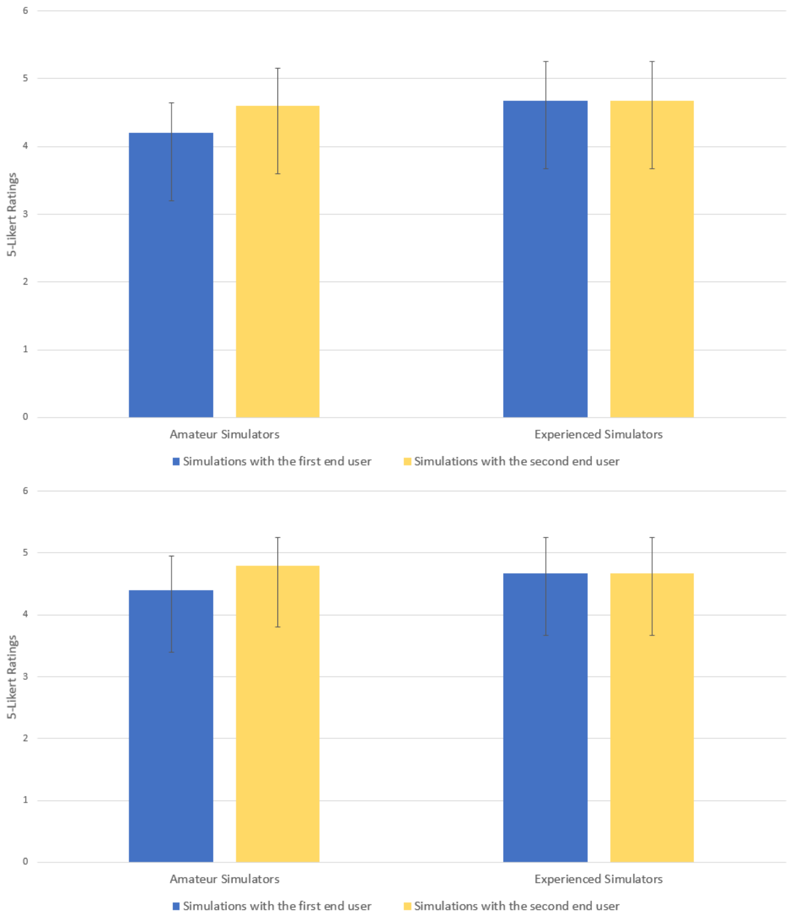


Fig 10. Top: results of simulation consistency perceived by the end users (questionnaire A Question 5); Bottom: results of simulation consistency perceived by the end users (questionnaire A Question 6) (no significant differences were reported between the first and second simulations and between the amateur and experienced simulator groups)

Secondly, we analyze how consistent the simulators' simulations are across multiple simulators, namely, the amateur simulators and the experienced simulators. Given the simulators on the non-rigorous responses (as mentioned in Section 4.2), we extract 1530 transactions from the simulation transcriptions. All these transactions included the "unrecognizable input" response and those that are incurred by technical failures are removed, finally, 1031 transactions are selected. The experienced simulator group responded more often with "unrecognizable input" responses than did the amateur simulator group (Nby experienced simulators=625, Nby amateur simulators=406); however, independent-sample T-test reported no significant differences in unrecognizable response numbers between the amateur and experienced simulator groups (p=0.09).

To determine the overall consistency of simulations across all the simulators, we examine how consistently the simulators interpreted the end users of speeches and how consistent their response decisions are. Due to the difficulties of analyzing all the transactions, we used two experts to traverse the simulators' surveillance videos and extracted short clips that contained predictive interpretations. Predictive interpretation means that the simulator understood the end users on the intentions and prepared a response beforehand. We counted the number of these clips across all the simulators in their simulations with the first and second end users, respectively (Figure 11), and found significant differences in predictive interpretation frequency between the amateur and experienced simulator groups (independent-samples T-test p=0.044). This confirmed the influence of simulator experience on the process of interpreting user input.

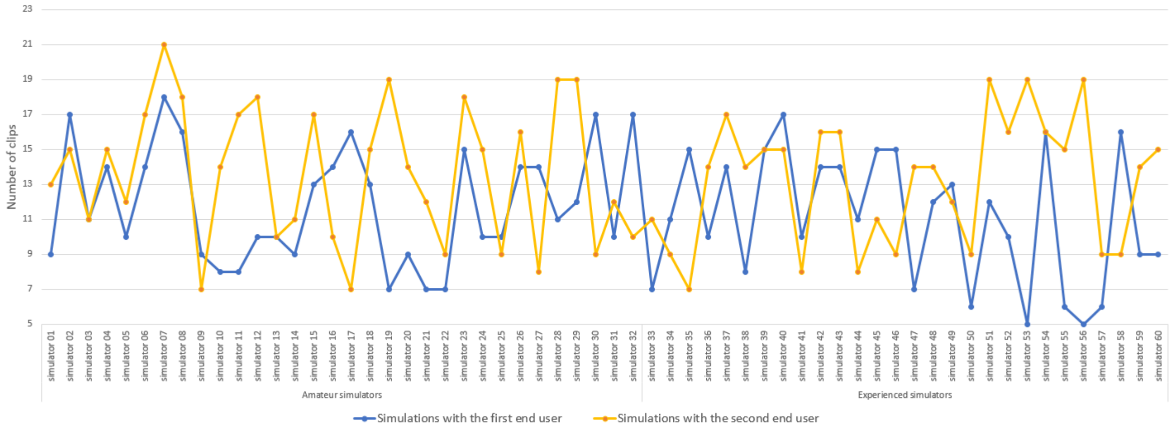


Fig 11. Results of the simulators on the predictive interpretations (significant differences between the amateur and the experienced simulators)

The expert walkthrough analysis reveals several reasons for the results. Given the support of automated simulation tools, simulator responses are faster than those of the end users; therefore, the simulators had time to make predictive interpretations. In addition, during end user pauses, experienced simulators appear more likely to prepare subsequent operations.

Overall, the analysis shows that the mimetic behavior of individual simulators (e.g., interpreting user input) are consistent, regardless of the increase of simulation experience in the study. In contrast, the analysis results also show significant differences in simulation consistency (predictive interpretation and frequency of "unrecognised input" responses) across the two simulator groups. However, the inconsistencies in the simulations of simulator are not sensed by the end users, as indicated by the questionnaires and interviews. Given these results, we can not only confirm the inconsistency of simulation in this study but also report that such inconsistency did not influence the end users.

4.4. The efficiency of the simulations

Efficiency is an aspect of the consistency of a simulation because the simulators might have different qualities in their ability to simulate the expected responses. This aspect reflects the simulator overall productivity with the simulation tools. To measure the efficiency within and between the simulators, we analyze the questionnaires and interviews. In addition, we use an expert-walkthrough (one expert in this case) to examine the simulators on the surveillance videos to validate the analysis results.

To measure the individual simulation of simulator efficiencies (with the first and the second end user, respectively), we analyze the questionnaires with independent-sample T-test (Question 7: M=3.62, SD=1.20; Question 8: M=3.29, SD=1.07); the results are shown in Figure 12. Neither Question 7 nor Question 8 reported significant differences in simulation efficiency between simulations with the first and second end users (independent-sample T-tests, PQuestion7=0.71, PQuestion8=0.54).

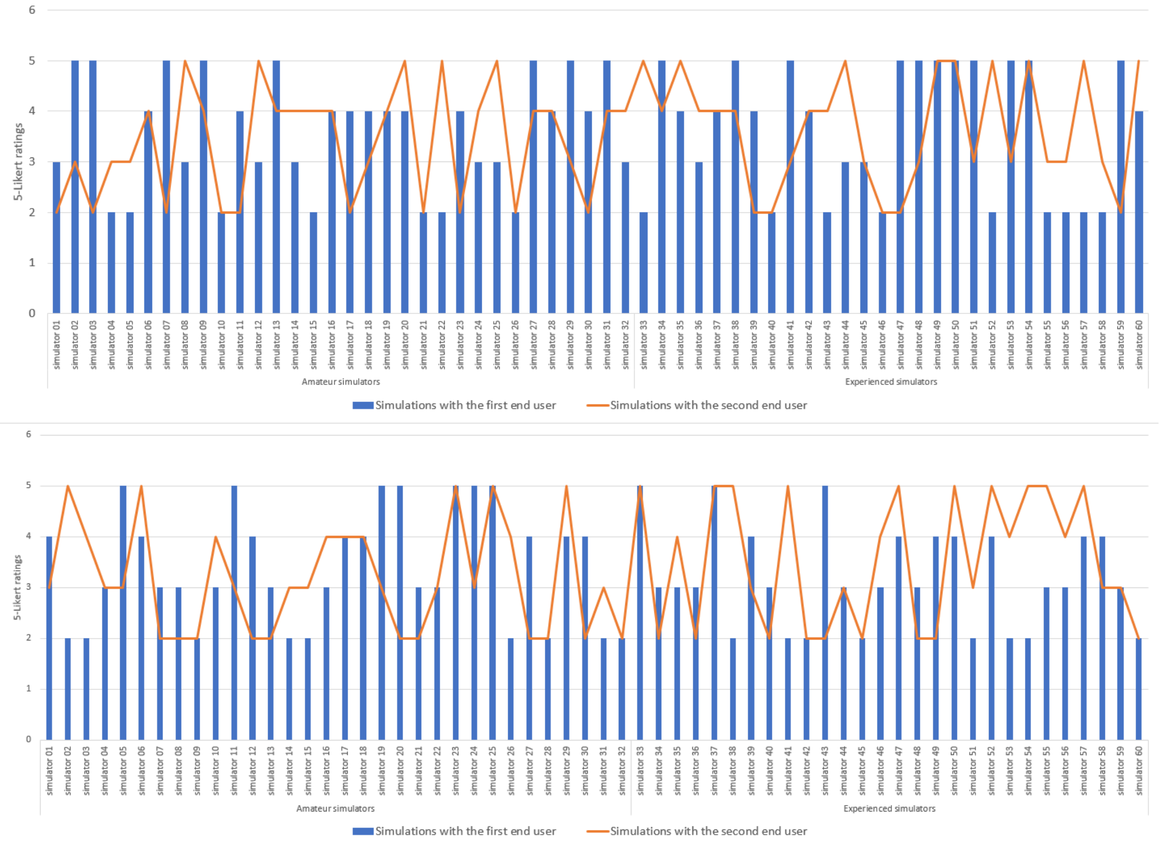


Fig 12. Results of individual simulators' simulation efficiency with the first and the second end users (top: Question 7, bottom: Question 8, no significant differences were reported in these questions)

To obtain a quantitative understanding of the simulators' simulation efficiencies in the simulations with the first and second end users, we refer to the analysis results discussed in Section 4.2 (the simulators' response speeds) and found that both types of simulator groups gained efficiency during the simulations. However, we note that such efficiency improvements are subtle, because no evidence is captured in the study to demonstrate that the end users sensed the changes in efficiency.

As mentioned in Section 4.1: the Credibility of the simulation, the study found no significant differences in overall simulation speed between the two simulator groups. It indicates that the efficiency between the simulator groups are similar. Moreover, the interview data analysis (Appendix B, Question 4) reported no obvious differences in system responses. The responded participants that they 'feel the overall system responses are instant and acceptable' and that they 'understand that the system would have different response speeds with different speeches'. Taken together, we can claim that there is no significant difference among the simulator groups.

4.5 Engagement and satisfaction

We analyze the overall engagement and satisfaction of the end users in the study. This served two purposes. First, the analysis data are useful in validating other measures, especially the credibility and consistency of the simulations. Secondly, the analysis examined the end users' interaction status which ensured the overall reliability and validity of the study.

To reflect end user engagement with the simulated system, we designed one specific question to measure the overall usability of the simulation system as perceived by end users (Appendix A, Question 9). Due to the multiple facets of usability in different contexts (Hertzum, 2010), we adopt the extended usability definition based on the widely-accepted usability definition of ISO 9241 (Bevan, 2001). The usability definition comprises five main characteristics: effectiveness, efficiency, error tolerance, ease of learning and overall user engagement. Because the first four characteristics are already well addressed through the WoZ structure, the characteristic of engaging is adopted as the major indicator of usability. The analysis results show that the end users generally perceived highly positive engagement in the study (MQuestion9=4.54, SDQuestion9=0.55; MQuestion10=4.36, SDQuestion10=0.92). The analysis results showed no significant differences between the simulations of simulators with the first and second end users or between the amateur and experienced simulator groups (independent-sample T-tests, P first and second simulation=0.48, P amateur and experienced simulator groups=0.36).

Additionally, we analyze the end user overall satisfaction with the study. The instructor's observations of the end users report good task completion, as all the end users successfully accomplish the study tasks without encountering serious difficulties. The questionnaire results (Appendix A, Question 10) showed high satisfaction ratings (M=4.58, SD=1.30), but do not report significant differences between the simulators' simulations with the first and second end users or between the amateur and experienced simulator groups (independent-sample T-tests, Pfirst and second simulation=0.24, Pamateur and experienced simulator groups=0.065).

During the interview analysis, we capture the end users' comments concerning their overall experiences using the spoken dialogue system (Appendix B, Question 5). A few end users complain about the uncertain versatility of speech vocabularies because these end uses felt 'unsure about whether [the end users'] speech could be successfully recognized and responded to, especially after some speeches are rejected'. In addition, the end users mention the overall use of the system as being 'a little bit strange when talking to a desk (the coffee table) as if it's a human being'.

5. Results

The analysis findings are classified according to the four major requirements for WoZ simulations, plus the additional aspect of overall user engagement and satisfaction. Table 3 provides a summary of these findings.

Table 3. Summary of the results

Measures	Criteria of measures	Results
Credibility of simulation	•Awareness of simulation •Perception of system functionalities(convince)	•High convince level perceived by end users •No significant difference in end users' perceived credibility •No significant difference in amateur and experienced evaluator groups' perceived credibility •Experienced levels have no significant influence on the overall credibility
Rigour of simulation	•Evaluators response speed and styles •Evaluators' user input interpretation accuracy •Simulator's dependency on evaluator's inherent experience •Evaluator's accordance with presumed system as defined in study scripts (likeness of presumed system)	•No significant differences in the evaluators' gained experience between the amateur and experienced evaluator groups •No significant differences in the simulation with the first and second end users between the overall, amateur, and experienced evaluators •No significant differences in evaluator's inherent experience between the first and second simulation •No significant differences in evaluator's response styles between the amateur and experienced groups •Identified evaluator's adaptive simulation behaviours •Identified evaluator's instinctive(unintentional) response •Identified non-rigorous response in both amateur and experienced groups
Consistency of simulation	•Long-term simulation consistency •Short-term simulation consistency •Prediction activities, including times and patterns	•Prediction times: significant differences between evaluators •Prediction patterns: found different patterns between evaluators
Efficiency of simulation	•Response speed •System reaction rhythms perceived by users	•No significant differences in simulation speed between the simulation with the first and second end users •No significant differences in simulation rhythms and response speed between the amateur and experienced evaluator groups
Engagement and satisfaction	•End user's perceived engagement of interaction •End user's perceived satisfaction during the interaction	•High level satisfaction reported •No significant differences in engagement between the amateur and experienced evaluator groups •High level satisfaction reported •No significant differences in satisfaction between the amateur and experienced evaluator groups

Of these findings, credibility of the simulation is the foundation of this WoZ study. Rigour of simulation provides evidence to H2 and consistency of simulation provides a direct answer to H1. In addition, the efficiency of simulation helps to reflect the effects of automated simulation tools on simulators' overall simulation productivity, and engagement and satisfaction help to validate the overall quality of task interactions. The tests of H1 and H2 are explained as follows.

H1 is supported because from the end user of aspect, no significant differences are reported throughout all the five criteria. However, from the simulator aspect, there is significant differences in consistency of simulation. Specifically, prediction time and the patterns between amateur and experienced simulators are different.

H2 is supported because non-rigorous mimetic behaviours such as adaptive response styles and instinctive responses were reported from the hard evidence gathered during the study. In this regard, there is a significant difference between amateur and experienced simulators.

6. Discussion

Modern Wizard of Oz studies include many measures that support human responses of simulators. These measures are effective in hindering most improvisational operations and being of help to maintain consistent mimetic behavior. This study confirmed the advantages of adopting automated simulation tools in WoZ simulations. However, the most important contribution of this study is not to repeat the merits of the automated simulation tools in the context of spoken dialogue system simulations; instead, the intent of this study is to explore how automated simulation tools affected the mimetic behavior of simulators. The study results add new understandings to the existing knowledge since WoZ studies are conventionally believed to be more reliable. It is equipped with carefully designed automated simulation tools facilitate by experienced simulators. However, one important study value is that both amateur and experienced simulators can provide non-rigorous and inconsistent responses, yet such responses caused little harmful to the overall simulation.

The automated simulation tools cannot prevent the instinctive mimetic behavior. This study provides evidence that the human simulators are not strictly capable of impersonating advanced computer systems even when they are equipped with automated simulation tools. This evidence is useful for a wide range of designers and researchers who employ novice or experienced or hybrid simulators to present pseudo-functionalities. Additionally, the results call for more future research in simulators' non-rigorous mimetic behavior during simulation responses. For instances, showing how these instinctive mimetic behavior occur for what circumstances they occur, and whether pre-study training or other measures can be implemented to reduce the chance of instinctive mimetic behavior. Future research in this direction could lead to improvements in the reliability and validity of WoZ studies when they are used as tools to iterate design.

The evidence revealed by this study is also useful in the broader context of design methods that employ human beings to mimic system components. For example, experimenters might be concerned with unnoticeable non-rigorous sorting behavior in a card sorting design, although critical psychological and cognition studies will be required to identify related effects. Similarly, the design methods that involve humans such as cognition walkthroughs and heuristic evaluations are the same non-rigorous behavior might also be encountered regardless of the designers' backgrounds and expertise. In this regard, this study yields generalizable evidence of simulator roles in human-participatory design activities. This information is helpful to both designers and evaluators in future emerging automated-interaction user interfaces and tools through intentional human experimenters and instinctive behavior which are reflected in iterative design processes. More implications are discussed below.

This study focused on the mimetic behavior of simulators rather than on end user task performances. Consequently, the automated simulation tools—the major medium for simulations—received extraordinary attention. Several roles are extracted from this new understanding. For example, the simulation tools functioned first as an efficient toolkit to support the simulators' responses. Automated tools save simulators from having to perform trivial operations and improve the speed of sequential responses. In addition, the tools act as filters that reduce occurrences of the most inconsistent operations. The simulators are forced to follow modularised procedures to return responses. Finally, the tools are acted as a "translator" to transform the instinctive behavior of simulators into intentional operations. Considering the above mentioned, we can confirm the effectiveness of increasing the automation levels of simulation tools. However, despite their important roles in WoZ simulations, simulation tools by themselves cannot eliminate the simulator non-rigorous behavior. Moreover, pre-study training does not seem to have a strong impact in this regard. We foresee an increasing adoption of integrated and automated simulation tools in future WoZ studies that will gradually change current WoZ study structures and conditions. For example, when using such tools, simulators no longer need to rush to type a response as they did in three decades ago rather than simulators can respond only a few clicks and their responses are more concise and consistent. Given those changes, simulators are likely to become interpreters instead of operators—that is, the simulators surveille end users' activities, decode their input and determine the response strategies rather than being preoccupied by operational minutiae.

Finally, and most importantly, we classify mimetic behavior of the simulators into two types: the intentional and instinctive. The former reflects the simulators' interactions with the automated simulation tools, which are mostly convincing, consistent, and efficient. This result provides new evidence that supports the findings in previous WoZ studies. The latter reveals the simulators in non-rigorous mimetic behavior, which are instinctive and unwitting. This result adds new understanding to the current knowledge of simulators' mimetic behavior with automated simulation tools. Moreover, it clarifies the correlations between such behavior and simulation performances. Given the increasing desire for speech interaction to mimic human emotion and exhibits personality, the findings indicate that the instinctive response would also influence the evaluation of mimetic behavior in WoZ method.

This exploratory study has several limits worth mentioning. First, this study is not a strict exploratory study because it does not intend to follow the typical avenues of such exploratory studies. Specifically, the authors consider this study as more of an empirical and exploratory hybrid study. The main reason it is termed an exploratory study is due to its main contributions: the study explores a specific topic and raises questions for future research. In this regard, admittedly, the understanding of how automated simulation tools affect the simulators' mimetic behavior is incomplete because more research is required to clarify the rationales and the related influence factors. In addition, quantitative analysis is used in this study, which serves as a convincing foundation for the questions raised in this study.

Secondly, this exploratory study is clearly not an exhaustive traversal of the mimetic behavior of all simulators to reveal related influence factors. Instead, it started with the problem that simulators provide inconsistent responses—a problem that has been of common concern in massive WoZ simulations. Given the observations concerning of the simulators in interactions with automated simulation tools, this study then hypothesised potential relationships between automated simulation tools and human simulators' mimetic behavior. In addition, we conducted partial empirical experiments to validate the hypothesises. As mentioned, for example, there are several reasons for selecting a small sample group in these experiments, as listed below.

1) There are practical difficulties in recruiting as many simulators as possible to reach a sense of 'significant' participant numbers. The philosophy behind the WoZ method is to use fewer simulators to mimic more system components; however, this introduces challenges in recruiting numerous simulators who have experience with other specific system simulations.

2) Despite these practical difficulties, we gave careful methodological consideration to the sample size before conducting the study. It is a significant challenge to traverse every single mimetic behavior. An alternative approach is to use an exploratory study which observes simulator mimetic behaviors throughout specialized simulations and then extracts features that violate conventional criteria. Moreover, the results from the strict experimental procedures and statistical analyses should be reliable, because these findings indeed exist in simulators' mimetic behavior. Furthermore, the study results are self-explanatory since the given instinctive non-rigorous mimetic behavior are largely simulation task- and tool- independent. From this base, we believe it is not unreasonable to recruit a carefully selected group of simulators, extract typical features from their mimetic behavior, and then classify these features into a complete taxonomy. In addition, we have had special awareness of the potential impacts of cross-social-culture on the study results, as the first and the second half of the study are conducted in different countries. We revisited the study results and found few clues of such impacts. For instances, the simulators and end users comply with the study scripts and their spoken dialogues are relatively short, which can prevent misunderstandings during long speech interpretation.

3) The study included some measures to constrain the simulators' activities, thus making these activities not as free as in open tasks. For example, the study scripts defined what the simulators could and could not do. Considering these restrictions, including a larger number of simulators would be unlikely to enrich the varieties of simulated interactions.

Third, this study uses undergraduate students as end users. The students have good learning abilities and are open to novel interactive systems. These are positive impacts. Changing the end users to elderly people, for example, might have possible effects on the results. However, the end user's interactions are not measured in any way in this study. It is only their feedback concerning the performances of the simulated system is considered.

Fourth, this study is not intended to provide a direct solution to the problems of inconsistent and non-rigorous simulations. On the other hand, the study contributions are intended to benefit a wide range of future designers using WoZ simulations. An enhanced understanding of automated simulation tool-related determinants and impacts are expected to be helpful because human-simulator-based simulations may remain useful for a variety of purposes and in a variety of scenarios.

Finally, because information and communication technologies (ICT) are improving rapidly as well as new possibilities are likely to arise in which multiple simulators may work collaboratively and remotely to mimic an interface. Although this study does not address the topic of multiple simulators, a better understanding of the behavior revealed here is likely to be applicable to multi-simulator studies as well.

7. Conclusion

In this work, an exploratory study is conducted to investigate the mimetic behaviors of a simulator with automated simulation tools. It raises generalizing considerations on the role of human simulators when mimicking rigid computer systems. The study compares the performances of the experienced simulators with those of amateur simulators from the perspectives of credibility, consistency, rigor, and efficiency. The study found significant differences between the two groups of simulators in simulation consistency and rigor, but no significant differences in the other metrics. Furthermore, the study reveals two different characteristics on mimetic behaviors of simulator: the intentional and the instinctive. The latter characteristic, as explained in the Discussion section, is less affected by WoZ measures such as intense simulation training or advanced simulation tools. As a part of study values, these findings give implications that are generalizing to wizard-of-oz researchers and users. Moreover, the study discusses the influence of such non-rigorous mimetic behavior.

Acknowledgements

The work was supported by the National Key Research & Development Program under Grant 2016YFB1001304.

References

N. Dahlbäck, A. Jonsson and L. Ahrenberg (1993). Wizard of Oz studies: why and how. Proceedings of the 1st international conference on Intelligent user interfaces. Orlando, Florida, United States. January. ACM. 193-200.

J. F. Kelley (1984). An iterative design methodology for user-friendly natural language office information applications. Vol. 2, pp. 26-41: ACM.

J. Thomason and D. J. Litman (2013). Differences in User Responses to a Wizard-of-Oz versus Automated System. HLT-NAACL. 796-801.

S. Schlögl, G. Doherty and S. Luz (2015). Wizard of Oz experimentation for language technology applications: Challenges and tools. Interacting with Computers, 27(6), 592-615.

P. Sequeira, T. Ribeiro, E. Di Tullio, S. Petisca, F. S. Melo, G. Castellano and A. Paiva (2016). Discovering social interaction strategies for robots from restricted-perception Wizard-of-Oz studies. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 197-204. IEEE.

L. D. Riek (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. Journal of Human-Robot Interaction, 1(1).

S. Schlögl, G. Doherty and S. Luz (2013). Managing Consistency in Wizard of Oz Studies: A Challenge of Prototyping Natural Language Interactions.

N. M. Fraser and G. N. Gilbert (1991). Simulating speech systems. Computer Speech and Language, 5, 81-99.

T. Grill, O. Polacek and M. Tscheligi (2012). Conwiz: A tool supporting contextual wizard of oz simulation. Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia. 21. ACM.

J. N. Bott and J. J. Laviola Jr (2015). The WOZ Recognizer: A Wizard of Oz Sketch Recognition System. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(3), 15.

M. Ralph and M. A. Moussa (2008). Toward a natural language interface for transferring grasping skills to robots. Robotics, IEEE Transactions on, 24(2), 468-475.

V. Ashok, Y. Borodin, S. Stoyanchev, Y. Puzis and I. Ramakrishnan (2014). Wizard-of-Oz evaluation of speech-driven web browsing interface for people with vision impairments. Proceedings of the 11th Web for All Conference. 12. ACM.

J. Drummond and D. Litman (2011). Examining the impacts of dialogue content and system automation on affect models in a spoken tutorial dialogue system. Proceedings of the SIGDIAL 2011 Conference. 312-318. Association for Computational Linguistics.

M. Shiomi, D. Sakamoto, T. Kanda, C. T. Ishi, H. Ishiguro and N. Hagita (2008). A semi-autonomous communication robot—A field trial at a train station. Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on. 303-310. IEEE.

G. Alce, K. Hermodsson and M. Wallergård (2013). WozARd: a wizard of oz tool for mobile AR. Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services. 600-605. ACM.

J. S. Pettersson and M. Wik (2015). The longevity of general purpose Wizard-of-Oz tools. Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction. 422-426. ACM.

X. Li and J. V. H. Bonner (2014). Using wizard-of-oz method to build multipurpose platform for domestic ambient media research and applications. Multimedia Tools and Applications, 72(2), 1011-1026.

R. F. Adler, F. Iacobelli and Y. Gutstein (2016). Are you convinced? A Wizard of Oz study to test emotional vs. rational persuasion strategies in dialogues. Computers in Human Behavior, 57, 75-81.

M. Mavrikis and S. Gutierrez-Santos (2010). Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. Computers & Education, 54(3), 641-651.

G. D. Fabbrizio, G. Tur and D. Hakkani-Tür (2005). Automated wizard-of-oz for spoken dialogue systems. Ninth European Conference on Speech Communication and Technology.

C. D. Chandler, G. Lo and A. K. Sinha (2002). Multimodal theater: Extending low fidelity paper prototyping to multimodal applications. CHI'02 Extended Abstracts on Human Factors in Computing Systems. 874-875. ACM.

M. Vahdat, S. George and A. Serna (2013). Wizard of Oz in designing a collaborative learning serious game on tabletops. International Journal of Information and Education Technology, 3(3), 325.

G. M. Fitch, D. S. Bowman and R. E. Llaneras (2014). Distracted driver performance to multiple alerts in a multiple-conflict scenario. Human Factors: The Journal of the Human Factors and Ergonomics Society, 0018720814531785.

A. Deshmukh, S. Janarthanam, H. Hastie, S. Bhargava and R. Aylett (2013). WoZ Pilot Experiment for Empathic Robotic Tutors: Opportunities and Challenges. International Conference on Social Robotics: Proceedings of workshop on Embodied Communication of Goals and Intentions.

L. F. Baum (2014). The Wizard of Oz-With Audio. Oxford University Press.

M. Biswas and J. Murray (2013). Building a long term human-robot relationship: how emotional interaction plays a key role in attachment.

Y. Li, J. I. Hong and J. A. Landay (2007). Design Challenges and Principles for Wizard of Oz Testing of Location-Enhanced Applications. Pervasive Computing, IEEE, 6(2), 70-75.

M. Serrano and L. Nigay (2009). Temporal aspects of CARE-based multimodal fusion: from a fusion mechanism to composition components and WoZ components. Proceedings of the 2009 international conference on Multimodal interfaces. 177-184. ACM.

S. Gandhe and D. Traum (2014). SAWDUST: a Semi-Automated Wizard Dialogue Utterance Selection Tool for domain-independent large-domain dialogue. Proc. SIGDIAL 2014 Conference. 251-253. Citeseer.

G. Hoffman (2016). OpenWoZ: A Runtime-Configurable Wizard-of-Oz Framework for Human-Robot Interaction. 2016 AAAI Spring Symposium Series.

A. Steinfeld, O. C. Jenkins and B. Scassellati (2009). The oz of wizard: simulating the human for interaction research. Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on. 101-107. IEEE.

S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee and J. Yang (2003). Predicting human interruptibility with sensors: a Wizard of Oz feasibility study. Proceedings of the SIGCHI conference on Human factors in computing systems. Ft. Lauderdale, Florida, USA. ACM.

B. K.-J. Mok, D. Sirkin, S. Sibi, D. B. Miller and W. Ju (2015). Understanding Driver-Automated Vehicle Interactions Through Wizard of Oz Design Improvisation. 8th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.

A. Habibovic, J. Andersson, M. Nilsson, V. M. Lundgren and J. Nilsson (2016). Evaluating interactions with non-existing automated vehicles: three Wizard of Oz approaches. Intelligent Vehicles Symposium (IV), 2016 IEEE. 32-37. IEEE.

A. Katz, F. Basis and A. Shtub (2015). Using Wizard of Oz technology for telemedicine. HS, 4(3), 224-235.

R. Taib and N. Ruiz (2007). Wizard of Oz for multimodal interfaces design: deployment considerations. Human-Computer Interaction. Interaction Design and Usability Vol. 4550/2007, pp. 232-241. Berlin/Haidelberg: Springer Berlin/Haidelberg.

J. D. Gould, J. Conti and T. Hovanyecz (1983). Composing letters with a simulated listening typewriter. Vol. 26, pp. 295-308: ACM.

P. Markopoulos, J. C. Read, S. MacFarlane and J. H sniemi (2008). The Wizard of Oz Method. Evaluating Children's Interactive Products pp. 218-233. Burlington: Morgan Kaufmann.

X. Li and J. Bonner (2011). Improving control panel consistency of wizard-of-oz design and evaluation studies. the 17th International Conference on Automation & Computing. 163-168. Huddersfield, UK.

K. Forbes-Riley and D. Litman (2011). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. Computer Speech & Language, 25(1), 105-126.

D. Salber and J. Coutaz (1993). Applying the Wizard of Oz technique to the study of multimodal systems. Human-Computer Interaction Vol. 753/1993, pp. 219-230. Berlin/Heidelberg: Springer Berlin/Heidelberg.

K. Hornbak and E. Frokjar (2008). A study of the evaluator effect in usability testing. Human-Computer Interaction, 23, 251-277.

B. Matthews and T. Heinemann (2012). Analysing conversation: Studying design as social action. Design

Studies, 33(6), 649-672.

P. Seedhouse (2004). Conversation analysis methodology. Language Learning, 54(S1), 1-54.

M. Hertzum (2010). Images of usability. Intl. Journal of Human–Computer Interaction, 26(6), 567-600.

N. Bevan (2001). International standards for HCI and usability. International journal of human-computer studies, 55(4), 533-552.

M.van Beuzekom, F.Boer, S.Akerboom and P. Hudson(2010). Patient safety: latent risk factors. British journal of anaesthesia, 105(1), 52-59.

在 WOZ 研究中使用自動仿真工具模擬仿真器的行為

范國光[1] 鐘祥銘[2] 婁小龍[3] 耿衛東[3] 厲向東[3]

[1]國立雲林科技大學設計學研究所

[2]浙江傳媒學院互聯網與社會研究院

[3]浙江大學計算機科學與技術學院、阿里巴巴與浙江大學前沿技術聯合研究所

摘要

「綠野仙蹤」(WOZ)使用「人」作為模擬器來模擬智能系統，這些系統超出了當前的技術水平或實現所需的時間和成本。在 WOZ 仿真中，越來越多的採用自動仿真工具。傳統的研究通常集中在終端用戶身上，對模擬器的行為所知甚少。為了彌合這一差距，我們進行了一項比較研究，以研究作為使用自動模擬工具模擬智能語音對話系統的模擬器性能的可信度、嚴謹性、一致性和效率。結果表現為兩種模仿行為：故意反應和本能反應。具體來說，不管經驗水平如何，模擬器都能很好的模擬故意的反應，在這種反應中，它們的反應是可信和有效的。然而，模擬器暴露了不一致和非嚴格的模仿行為的本能反應。此外，經驗豐富的和業餘的模擬器在響應速度、解釋最終用戶輸入和做出相應決策方面的本能模擬表現出顯著差異。本能模仿行為對模擬的最終用戶感知沒有顯著影響，這也反應出，討論 WOZ 作為一種高效設計工具的意義。

關鍵詞：綠野仙蹤；自動仿真工具；語音對話界面；模仿行為；故意和本能模擬

「設計學刊」投稿規範

一、投稿規定

投稿類別：

- 學術論文類(ResearchPapers)：具原創性之特點，在理論與方法上有可靠之系統化推演過程，或有實證的演繹歸納、其目的、方法、結論有明確交待者，或以既有研究之評論及分析比較為主，其觀點在知識推廣上，其資料在系統整理上，對提升國內設計學術研究有所助益者。本刊接受中文或英文投稿。

投稿內容不得有侵犯他人著作權或商業宣傳之行為，法律責任由作者自行負責。投稿論文需經本刊編輯委員會推薦之各領域專門審查者評審，通過後經編輯委員會正式決議通過後方得登載。

二、格式規定

- 為便於論文的編排與出版，所有稿件一律請以學刊所提供的範本檔案進行編輯與投稿。
- 中英文論文以不超過 15, 000 字、且不得超過出版論文 15 頁(含圖表)為原則，無論中文或英文論文皆需附 300 字以內中文摘要及 150 字以內英文摘要，論文頁數以雙數為主。
- 為進行雙匿名審查，稿件內請勿列出作者姓名，包含中英文之作者姓名及任職單位等相關資訊。如需引用作者先前已發表之其他著述，請以第三人稱敘述。
- 正文欲加強說明時所採之註釋，以(註 1)、(註 2)…. 為之，並將註釋依編號次序排於正文之後。註釋內文獻引用法與下述參考文獻規定相同。
- 文獻引用 APA 格式為之。APA 是美國心理協會(AmericanPsychologicalAssociation)所發行的出版手冊（PublicationManual）中，有關投稿該協會旗下所屬期刊（目前約為三十種）時必須遵守的規定。

  與原來所採用之編號系統不同的是，APA 格式的論文，在文章中是直接標上參考文獻的作者姓名與年代。針對不同的描述方式，而在文章中有下列兩種不同的標註方法：

1. 如果您在文章中要直接引用作者的姓名，請在其名字後直接加上該參考文獻的發表年份；例如：杜瑞澤（2002）提出了綠色設計（GreenDesign）的概念...

2. 如果您是直接引用研究的結果或論點，而沒有在句子中提及作者的姓名，請在該引用的字句旁，以()標註上文獻的來源；如：由於科技進步產業發達，人類在享受多樣的消費商品與商品使用的便利性，但也造成環境的高度污染，近幾年來人們開始對環境有所反思，逐漸的對綠色環保意識的重視，許多企業紛紛開始思考，如何在商品與環境之間，取得平衡點，因此「綠色創新設計」的觀念逐漸受各界重視。（杜瑞澤、陳炫助、管倖生，2015）。

目前 APAStyle 最新的版本是第六版，裡頭也有許多針對新的網路資料所制訂的標註規則。由於 APA 並沒有特別針對中文論文制訂其寫作格式，因此，本期刊所採用的，是由張保隆與謝寶煖（2006）兩位教授所撰寫的書籍[學術論文寫作：APA 規範　（華泰文化出版）]為之，請各位作者參考該書籍準備投稿論文。

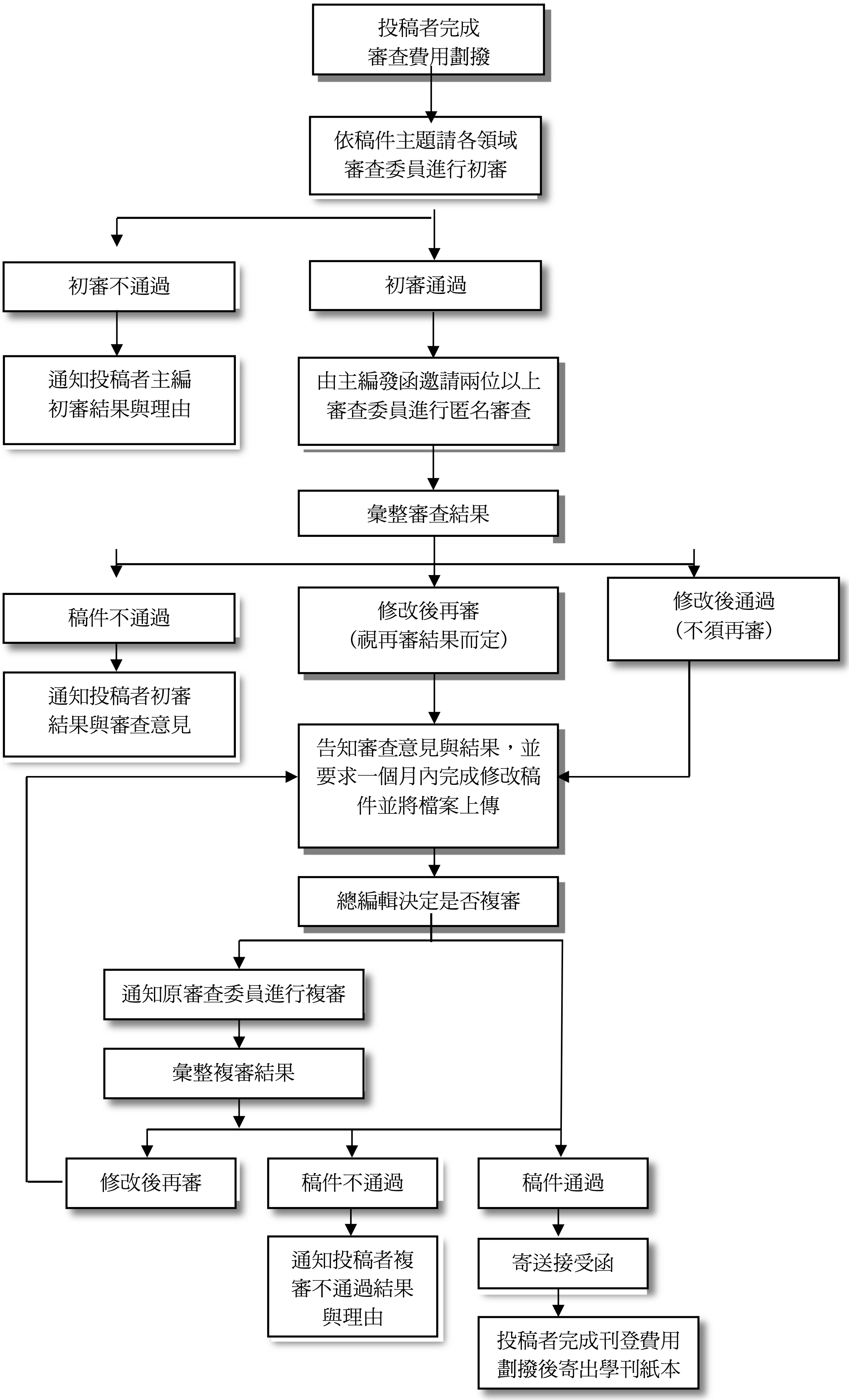
- 文章章節之編序以一、二、三…. 為章，以 2-1，2-2. . 為節，以 2-1. 1、2-2. 2、2…. 為小節來標示。小節以下依 1、2、3. . 及(1)、(2)、(3)等層級標示之。
- 為便於期刊編輯與印刷出版，投稿論文請使用所提供的範本檔案進行編排。
- 論文所採單位以國際標準制(SI 制)為主，所有數字皆以圖 6、200km、19 人、0. 98 等阿拉伯數字表之。
- 圖表製作必須清晰，圖表中所有字體以打字體完稿，並附有明顯的編號、標題及出典說明，否則不予受理。表之標題附於表上，圖之標題附於圖下。圖表編號皆以表 3、圖 9 等阿拉伯數字體表之。照片編號亦以圖號系列編列之，而不另以照片 1、2 編列。

- 圖表製作必須清晰，並清楚標示出圖表的詳細出處（包含書本中的第幾頁）外，還應該在投稿前取得其授權，以避免將來論文在網路與紙本上出版後，引起不必要的爭議。且論文通過後，作者需簽署著作權同意書，使得刊登。

三、線上投稿程序

- 請依照本學刊官網的指引，下載所需檔案並寄出，即可以將您的稿件傳遞到編輯委員會的信箱。作者資訊需包含論文之所有作者，有兩個以上作者時，依對論文貢獻程度順序排列，並註明各作者之服務機關。
- 自 106 年 9 月起，請於投稿後將審查費用新台幣 2600 元劃撥至設計學刊專用劃撥帳戶(帳戶：杜瑞澤，帳號：22823427)，確認收到款項後即開始審查。
- 當您將稿件上傳並完成劃撥程序後，編輯單位會先初步審查，如：論文格式是否正確；投稿者（包含共同作者）的資料（姓名、服務單位 Affiliation 等），之後將在最短的時間內通知主編開始稿件的審查作業，原則上於收件 3 個月內（接獲投稿費日起算）將提供審查結果。
- 當稿件通過審查並接受刊登時，我們將針對每一個稿件收取刊登費（每頁新台幣 150 元）。屆時，請在指定的時間內將匯票以掛號寄到編輯辦公室，始得登載。如作者要求輸出彩色版面，除需支付全額彩色刊登費用外，並於刊登前告知學刊編輯部。

設計學刊稿件審查流程


投稿者完成審查費用劃撥
依稿件主題請各領域審查委員進行初審
初審不通過
初審通過
通知投稿者主編初審結果與理由
由主編發函邀請兩位以上審查委員進行匿名審查
彙整審查結果
稿件不通過
修改後再審（視再審結果而定）
修改後通過（不須再審）
通知投稿者初審結果與審查意見
告知審查意見與結果，並要求一個月內完成修改稿件並將檔案上傳
總編輯決定是否複審
通知原審查委員進行複審
彙整複審結果
修改後再審
稿件不通過
稿件通過
通知投稿者複審不通過結果與理由
寄送接受函
投稿者完成刊登費用劃撥後寄出學刊紙本

設計學刊
Design

Journal of Design Studies

Publish by School of Design
National Yunlin University of Science and Technology

發 行 人：黃世輝
總 編 輯：黃世輝
執行編輯：林芳穗
編輯助理：石泓純、李芸菲
發 行 者：雲林科技大學設計學院
出 版 者：雲林科技大學設計學院

出版日期：2019 年 6 月
地　　址：（64002）雲林縣斗六市大學路 3 段 123 號
訂購金額：零售每本 450 元
訂購方式：請洽設計學刊編輯部 jds@yuntech.edu.tw

ISSN:2415 5632

JDS

Journal of Design Studies

National Yunlin University of Science and Technology School of Design