ExperienceDNA

A framework to conduct and analyse user tests in VR using the Wizard-of-Oz methodology

Jamil Joundi[1§[0002−3437−1972]], Klaas Bombeke[1§[0003−2056−1246]], Niels Van Kets[2§[0001−5495−2240]], Wouter Durnez[1§[0001−8045−8801]], Jonas De Bruyne[1§[0000−0002−6077−6084]], Glenn Van Wallendael[2§[0001−9530−3466]], Peter Lambert[2§[0001−5313−4158]], Jelle Saldien[1[0003−2557−3764]], and Lieven De Marez[1[0001−7716−4079]]

[1] imec-mict-UGent, Department of Communication Sciences, Ghent University, Miriam Makebaplein 1, 9000 Gent, Belgium
[2] imec-IDLab-UGent, Department of Electronics and Information Systems, Ghent University, Technologiepark-Zwijnaarde 126, 9052 Zwijnaarde, Belgium

Abstract. It is often challenging to measure participants' reactions during user tests where a Wizard-of-Oz (WoZ) method is applied. This method is applied by the observers to validate functionalities of concepts that are difficult to build. In this study, a new technique is developed using virtual reality (VR) to improve the measurement of the participant's reactions during such a test. In this system, called ExperienceDNA, VR user tests can be monitored and controlled through a desktop interface. In addition, physiological trackers (eye tracking and heart rate monitoring) are used to measure what the participant is looking at and to gauge their preferences. Moreover, the use of VR allows for quick adaptations to the virtual environment the participant is confronted with. In this way, highly versatile tests can be conducted while minimising the initial setup effort. Our approach has been validated by performing a pilot test on a predefined use case. The qualitative feedback collected from analysing batches of data from the pilot test is presented in the results section. In conclusion, this paper covers the development, description and evaluation of the ExperienceDNA framework, as well as some ideas for future improvements to this framework.

Keywords: virtual reality · user testing · design review · virtual training · WoZ testing

1 Introduction

In an ever-digitizing society, we face a new wave of smart products and services [20]. Interfaces are shifting, and the key differentiator is the end user's experience. However, in this smart and ubiquitous technological environment, experiences are increasingly determined by a complex interplay of interactions people are not always aware of [22]. People not only interact with each other and

technological objects anymore, but also with a broad diversity of content, contexts and platforms [25]. This interplay applies to many aspects of our lives: smart cities tell us how to optimize energy consumption, smart cars guide us to avoid traffic jams while playing our favorite music, smart homes reduce our heating bill by learning our habits, and smart factories help reduce the cognitive workload on their operators [24]. Interestingly, this shift from manifest to latent interactions, where interactions between human and computer become less prominent, has also led to new theoretical frameworks in the fields of product design, human-computer interactions and quality of experience research. Geerts and colleagues [5], for example, proposed an integrated Quality of Experience (QoE) framework consisting of four main components: the user, the (ICT) product, the use process and the context. These components can be measured distinctively in order to bring subtle aspects of QoE to the surface that could otherwise be overlooked. The more recent Human-Computer-Context Interaction (HCCI) framework of Van Hove and colleagues [25] defined the experience on five relevant interaction levels instead of using distinct components. These interaction levels include user-object, user-user, user-content, user-platform and user-context interactions, all of which should be considered during every stage of the user research or new product development processes to optimize the user experience. One of the advantages of this theoretical framework is that interactions can occur in two directions: user-object interactions, for example, can be about the user manipulating an object (e.g. switching on the vacuum cleaner), but also about the object having an effect on the user (e.g. an alarm informs the user that the cleaner got stuck).

However, in order to shape and guide a truly user-centric design process for these future smart products, we do not only need new theoretical frameworks, but also new methodologies and tools to disentangle this complexity of interactions. Doing so will allow us to isolate, simulate and assess the impact of each single determinant of the end-user's experience. In this sense, we like to think of user experiences as strands of DNA. One strand of this 'experience DNA' represents an ideal scenario, where users interact with the product or service exactly as the creators envisioned it. Another strand shows how the experience actually unfolds. In a perfect world, both strands bind perfectly. In untested products, however, things tend to be slightly different. QoE is often not what is expected, and it is difficult to pinpoint where things went wrong [10]. Perhaps the product is not as intuitive as we hoped, leading to some unwanted confusion about its 'smartness'. Possibly, end users do not interface with the product at the right time, in the appropriately 'smart' way. Hence, our objective was to build a concrete framework or methodology to detect these 'genetic malfunctions' – points of pain in the experience – in an early stage of the design process, allowing us to redesign the solution, and create the best possible experience.

When conceptualizing our ExperienceDNA framework, three main requirements were identified. First, our framework requires a highly immersive product-testing experience, allowing researchers to seamlessly simulate various contextual factors. Second, researchers should be able to steer the interactions happening during this experience. Third, and finally, our framework demands the capa-

bility to objectively capture all occurring interactions, as well as measure the cognitive-affective state and behavior of the test subjects.

In order to meet the first requirement – the capability to simulate immersive experiences and contexts – we turned to virtual reality (VR). Although it would also be possible to make product testing experiences more immersive without VR (e.g. putting furniture in a video dome or using augmented reality glasses), we believe that VR, at this point in time, has several advantages over other options. The accessibility of VR technology has increased in recent years, not only in terms of hardware costs (i.e., the HMD and supporting computer with appropriate GPU), but also in terms of software, with 3D software platforms (e.g. Unity) inviting non-experts to create VR environments themselves. The accessibility of such 3D engines is further complemented by the availability of vast libraries containing pre-made assets (e.g. [14,8]) giving users access to realistic models with minimal design efforts. To maximize the valorization potential of our tool and methodology, cost and accessibility (and relatedly, scalability) need to be considered, as such factors weigh heavily on a company's decision to adopt such a solution. Indeed, advanced prototyping is inherently risky given the costs associated with developing a first functional, real-world product, particularly when the core concepts are futuristic or unproven. As such, using VR as a testing platform yields a second advantage: the possibility to simulate products and their intended functionalities, regardless of their (potentially very low) technological readiness level (TRL). A final advantage of using VR to create immersive product testing experiences relates to another requirement described below. The latest generations of VR HMDs are progressively equipped with built-in sensors (e.g. eyetracking, facial expression recognition) that can be used to obtain objective measurements of the test subjects while they are engaged in the experience. Nevertheless, using VR during the product design [19,6,4] or automotive prototyping process [13] is not new. However, in these cases VR was mainly used as a visualisation tool (for products such as car interiors, shoes and accessories) but was not integrated as a tool facilitating the possible HCCI interactions.

Next to the capability to simulate experiences in immersive contexts, our approach requires the ability to give the researcher control over the interactions happening during the experience. Doing so not only makes it possible to evaluate all would-be functionalities of product, but also allows designers to simulate suboptimal use circumstances. This way, the product is effectively subjected to a virtual *stress test*, which may lead to the identification of previously hidden points-of-pain. This second requirement led us to a specific methodology in innovation research, the Wizard-of-Oz (WoZ) test protocol. When applying the WoZ methodology, users interact with an interface or system as if it were functional, even though its actions and responses are prompted by the researcher rather than the system itself. In other words, the researcher is pulling the strings, whereas the user thinks the system is automatically reacting to his or her actions [3,11,28]. The benefits of using a WoZ protocol include a shorter development time,

less resources and more freedom in conducting the user test since adaptations of the test can be materialised very quickly.

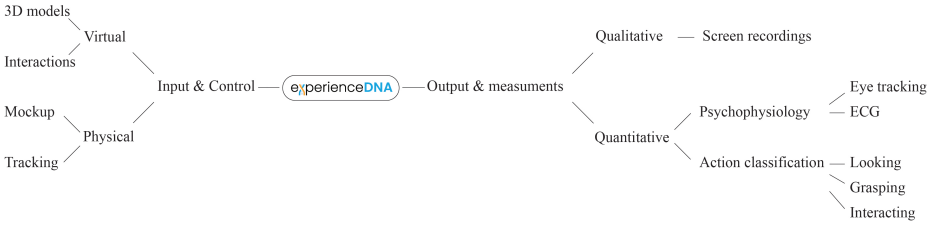
A third requirement was the ability to measure all possible interactions in an objective way, along with cognitive-affective and behavioral markers of the user during these interactions. Traditionally, product design and QoE research heavily rely on self-report methods: through focus groups, think-aloud protocols and post-hoc interviews, users can communicate their thoughts and experiences in detail with the designer. In addition, standardized questionnaires can be used before and after the test, such as the System Usability Scale [1], AttrakDiff [7] questionnaire or the Unified theory of acceptance and use of technology scale (UTAUT) [30]. However, although subjective reporting will always be very important during product design and user experience research, user responses might be clouded by various biases [18]. For example, the social desirability bias – i.e., the human tendency to give socially desirable responses instead of responses that are reflective of true feelings [23] – can be a significant problem.

The approach described in this paper combines VR technology, WoZ test protocols and an objective measurement strategy in a single tool in order to benefit to both the product development and user research domains, especially in the early stages of the (mostly IoT) product development process. The technical details of our tool are described in the next section.

2 Framework overview

ExperienceDNA represents a research framework that makes it possible to create immersive experiences for product testing, control the interactions between users and their surroundings, and collect a wide range of fine-grained sensor data, allowing researchers to map the entire user experience. In the following sections, this system is described in greater detail. This section is structured in line with the requirements mentioned earlier. Figure 1 presents the reader with an overview of the ExperienceDNA framework's key components.

Fig. 1: Overview of the ExperienceDNA framework.



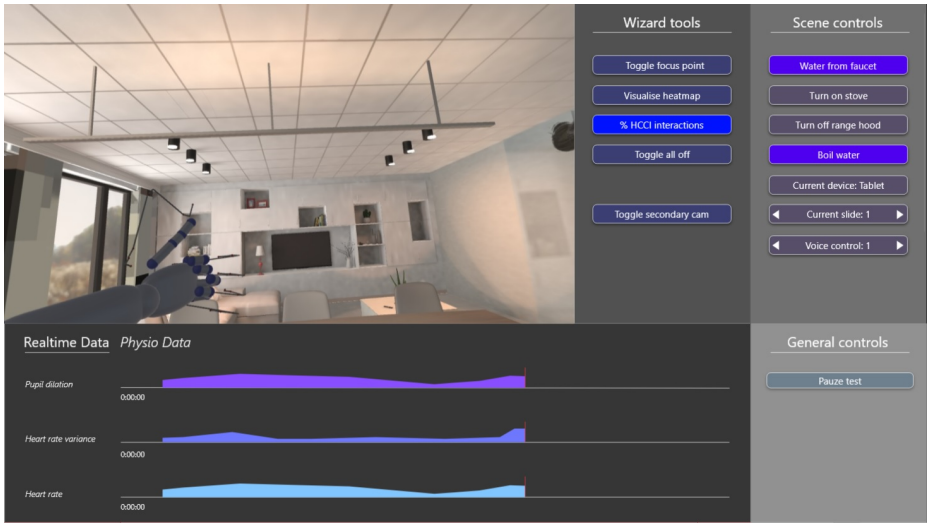
2.1 Creating a highly immersive product testing experience

ExperienceDNA is developed in Unity (version 2019.1.14f1), allowing researchers to both customize the visual aspect of their test setup (room, objects, materials, lighting) and the procedural aspect (animated events or prerecorded actions).

When tackling a new use case, the researchers first design the corresponding VR environment. A VR scene can be designed from scratch (3d modeling), imported from asset stores (e.g. a surgery room, surveillance control room, classroom, etc.), or built using a combination of custom models and pre-made assets. In some cases, 3D scanning techniques can be used to create digital copies of existing environments or objects.

Next, the researchers determine which of these static assets need to be made interactable, so users are able to experience the affordances of an object (e.g. a tablet or Alexa speaker). This is achieved using custom 'behavioral' scripts, which are appended to the assets in Unity. Apart from simulating interactive functionalities, these scripts (the backend of our framework) also allow researchers to manipulate the flow of the user testing through our dashboard (the researcher-friendly frontend), which is described in the next section.

Fig. 2: The multi-functional wizard-of-Oz dashboard allows researchers to trigger actions and monitor the data in real-time.



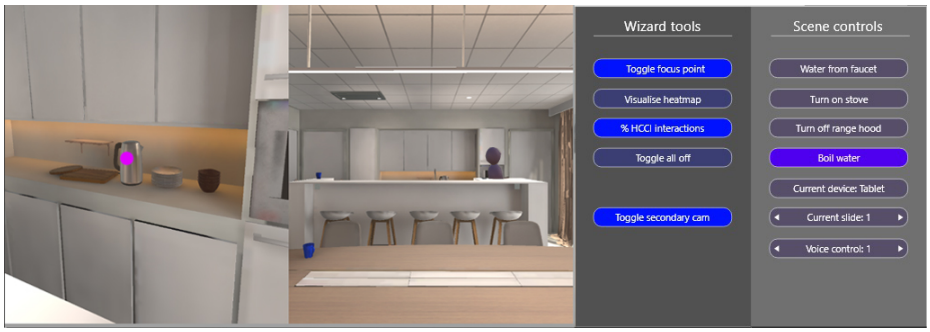
2.2 Controlling the product testing experience and initiation of interactions

Researchers interface with the ExperienceDNA framework through a visual dashboard (fig. 2), which is composed of three sections. First, the wizard – i.e. the researcher who is at the helm of the WoZ protocol – can manipulate the experiment in real-time by pressing scene control or general control buttons.

Second, the dashboard offers a real-time window on the data streams as the experiment proceeds. Presenting this live data view also helps the wizard to assure that all measurement channels are actively recording. In addition, and more importantly, it allows the wizard to monitor and compare the interaction data (looking at objects, grasping object and interacting with objects) with the real-time physiological data of the participant (see section 2.3). This gives the wizard a first, momentaneous indication of the participant's response to certain triggers.

Third and finally, the wizard can monitor what the participant sees through the headset, as the HMD's video feed is forwarded to the dashboard (fig. 3). Several visual overlays are added to this 'wizard view', such as a dot representing the participants gaze (as measured by the built-in eye tracker), a heatmap function highlighting the objects that received the most eye contact, and the possibility to shift into the position of one of several static scene cameras rather than monitoring the user's point of view.

Fig. 3: image of the wizard view with gaze tracking toggled on (left) and a third person view showing the avatar (right)



2.3 Objective measurement of interactions, cognitive-affective states, and user behavior

The ExperienceDNA framework facilitates both qualitative and quantitative QoE evaluations. To accommodate qualitative research efforts, researchers and

designers can rely on screen recordings – capturing both the subject's PoV and static perspectives from cameras placed in the scene. Quantitative measures consist of two main categories: psychophysiological measures and behavioral measures (actions).

Our use of a fully immersive virtual environment facilitates the capture of user actions to a great extent. ExperienceDNA is able to register various types of interaction, such as 'looking at' (using the built-in eye tracker), 'grasping' or 'interacting' (using the handheld controllers and positional trackers). Every interaction is timestamped, allowing researchers to explore reaction times or durations. These events can then be analyzed in combination with physiological data, providing researchers with insights into the affective characteristics of the user's experience.

In the present iteration of ExperienceDNA, both heart rate (HR) monitoring and eye-tracking measures have been implemented[3]. The former allows researchers to analyze heart rate (HR) and heart rate variability (HRV). HRV has been associated with emotional valence (i.e., affective quality: positive or negative), whereas HR has been shown to reflect arousal (i.e., physical intensity of responses to emotional stimuli) [17,9,12]. The current setup uses the affordable Polar H7 heart rate monitor, though other, more high end sensors can easily be integrated.

Eye tracking, apart from its aforementioned use in determining what users look *at*, also yields a measure for pupil dilation (associated with emotional arousal and cognitive effort [2,21]) and eye openness (a marker for drowsiness [16]). In addition, these raw data streams can be used to determine blink rate, which has also been identified as a marker of cognitive load [15,29]).

Once a session is concluded, ExperienceDNA saves all the recorded data (i.e. behavioral and psychophysiological data) in a .csv and .json format for post-hoc processing. In addition, aggregated output is generated, such as the durations and counts of *looking*, *grasping* and *interacting* events for each category of the human-computer-context interactions (user-to-object, user-to-user, user-to-platform, user-to-content and user-to-context). Psychophysiological data is stored in separate files, albeit with synchronized timestamps to facilitate post-processing.

3 Applied use case

In order to evaluate the core principles of the ExperienceDNA framework, an inaugural 'test flight' was conducted. In a series of pilot tests, the aim was to qualitatively validate the effectiveness of the framework, as well as the ease of implementation. To this effect, a cooking experience was created in which users need to cook a dish (i.e. bacon and eggs) following a recipe that was presented on a tablet next to the stove. This scenario was derived from a previous project on 'smart kitchen appliances', involving a tablet-based cooking assistant. The recipe was chosen to amount for a distinct , though limited, set of interactions to be implemented and evaluated in the pilot test. In the scene, the wizard could

[3] In further iterations of this framework, we foresee the integration of brain signals using an electroencephalography (EEG) headset.

initiate certain events, such as letting a phone ring to distract the user from the main task.

First, a Unity scene was designed in which users were able to perform the necessary actions (e.g. pour water from a faucet, heat it on a stove and boil eggs). Simultaneously, automated instructions were implemented on the virtual tablet, which presented users with the steps they needed to complete during the cooking task. In a second trial, this smart tablet was interchanged with a smart speaker – instructions were now presented verbally, and users were able to use voice commands to interact with the speaker. The intent of both trials was to evaluate the user experience of both the auditive and visual assisted cooking process.

The virtual environment was modelled after a real-life kitchen setting, which had already been used in the context of physical user testing (the "Homelab"). A true-to-life 3d model of the space was created using the 3D modelling software 'Rhino'. This model was then imported in Unity, where materials (e.g., textures) and assets (e.g., furniture and cooking items) were added. The process of creating environments and assets can be further sped up using 3D scans or pre-existing models.

In a second phase, all potential interactions were implemented in the scene. Using the WoZ prefabs, custom scripts were linked to the interactive 3D models. For instance, sound was linked to the phone asset, the bacon's model was made to change appearance based on the cooking time (raw, cooked, and burnt), realistic physics were assigned to the cooking assets (i.e., gravity, interaction with the controllers for grasping actions, collisions). Additionally, the scripts for logging physiological data were attached to the project. At the time of writing, this phase remains most time consuming, however there are many opportunities that will be pursued in the near future to further streamline the process (e.g. optimization of the code, cultivating a library a standardized objects types and behaviors, etc.).

In a last phase, and in order to deliver a high degree of realism to the participant, the visual quality of the scene was increased. This process involved placing extra lights and reflection probes to mimic real world lighting conditions. Furthermore, a more realistic shadow was achieved by creating baked light-maps that capture light bouncing of walls and objects in the scene. This phase is recommended if visual (photo)realism is deemed important for the test.

3.1 Setup

In order to have an objective evaluation of our framework, a professional design researcher (female, 26) was recruited to test the 'Wizard' functionalities of the system. In addition, an experienced user researcher (male, 28) to participated as a test user. Their feedback was gathered in order to optimise the virtual reality aspect of the experience as well as to enhance the usability of the ExperienceDNA dashboard. Both the design researcher and participant were accustomed to the use of VR.

The test was conducted in the Ghent University Art and Science Interaction Lab [26]. This lab is a state-of-the-art research facility able to effectively bring,

analyze and test experiences and interactions in virtual or augmented contexts. The test (fig. 4) was conducted using a high-end rendering machine equipped with a VR-ready graphics card (NVIDIA RTX 2080Ti), which was connected to an untethered HTC Vive Pro Eye headset. This setup delivered optimal free roaming capabilities, allowing the user to walk freely in a $\approx$ 10x10m area. The user's position was tracked using six HTC Vive 2.0 base stations. In addition, the user was fitted with a Polar H7 in order to monitor HR and HRV.

As specified earlier, the ExperienceDNA framework was designed in Unity. A custom (java)script was used to log heart rate data from the polar H7 monitor, and steam it in real time towards the Unity framework.

Since the focus of this first pilot test was to evaluate the overall ExperienceDNA framework (including the dashboard), the interactions are performed with the HTC Vive controllers. However, future iterations of ExperienceDNA will accommodate the use of (virtually tracked) real-life objects and haptic gloves. This makes haptic interactions, such as tapping on a tablet for example, more realistic.

3.2 Procedure

Our evaluation followed a think-aloud protocol [27] to detect usability problems during the VR user test. After the VR portion of the test, a semi-structured interview was conducted with the wizard, containing general and qualitative questions inspired by usability frameworks such as the System Usability Scale (SUS) [1], Unified Theory of Acceptance and Use of Technology (UTAUT) [30].Twelve questions from the SUS and 17 questions from the UTAUT framework were used to assess usability, confidence, performance expectancy, effort expectancy and behavioural intention towards the tested dashboard. Although these questions could be used in every experiment performed with the ExperienceDNA system, their goal was mainly to get deeper insight in usability problems when using the dashboard during this pilot test. Typical questions of the UTAUT framework are: "Does the use of this dashboard allow you to conduct experiments faster?", "Are you confident using this dashboard?", "Would people be willing to learn how to use this dashboard?", "Would this system be used in the future and by whom?". Important questions related to usability are:"Do you think the system is easy to use?", "Do you think you will need technical support when using this system?", "Do you think that the functionalities of this system are well integrated?". In order to collect more specific feedback, the wizard (i.e. the researcher) was asked how they evaluated their interactions with the three main features of the dashboard. These specific questions were oriented towards using the 'Wizard View', 'Scene controls' and data visualisation. In another interview, a QoE (Quality of experience) [31] assessment was done with the (VR) participant and semi structured interview followed to evaluate his experience and his tasks in VR. This pilot study was videotaped and comments were recorded and annotated for a more complete and unbiased overview of the responses from wizard and participant.

Fig. 4: The setup during pilot test with the wizard (right) and the participant (left).

3.3 Results

This section sheds light on the usability of the ExperienceDNA framework from two vantage points: that of a participant (taking part in the immersive experience), and that of a researcher (the wizard at the helm of the experience flow). The results of two semi-structured interviews are restructured in paragraphs highlighting both the framework's merits, as well as its current points of pain.

Subjective experience of the wizard The following subsection elucidates the wizard his general impression and the feedback on three specific functionalities of the system: the wizard view, scene controls and experiment data. In a final paragraph we included some future improvements suggested by the wizard.

General impression The wizard complimented the system's ease of use and the integration of different functionalities. She also commented that chances to conduct a good test increase because this system is easy for a single person to operate.

"This system increases my productivity, because it is much more realistic than building a quick 'wizard of test' yourself. You don't have to rebuild everything from scratch."

The VR setting of the framework was considered to be a versatile choice, mainly since it can be used to conduct the same experiment repeatedly even with slightly different interaction and context settings. The VR dashboard not only indicates whether people will use the tested product as envisioned but also delivers insight whether they engage in a positive experience regarding the tested product. "This system would be very useful for assessing whether people like to use something. It takes less time to find out that people don't like something because you don't have to build the thing first."

The used ExperienceDNA framework was perceived as especially interesting for evaluating new envisioned concepts as well as for testing bigger projects (such as escape games, public spaces, smart device interactions). Tests in these domains are generally difficult to recreate in a conventional wizard of Oz test.

Even though the system was perceived as easy to use, three functionality problems were detected during the pilot test. First, some functionalities remained unnoticed during the first-time use (e.g. a button for switching from the 'tablet interaction mode' to the 'voice assistant interaction mode'). Another functional problem that remained was anticipating behavior of the participant.In that case, the wizard could respond erroneous resulting in unnatural interactions during a user test (e.g. when the wizard reacts too late or not at all). In a final remark, the cost of the VR setup and limited knowledge of programming appeared to limit the chance of adoption of this kind of system by test designers.

Specific functionalities Next, the wizard commented on the 'wizard view'. Several features made it interesting for real-time evaluations. Even though it remains impossible to read body language just as in real life, it allowed the wizard to follow the participant and his gaze in the virtual space. The secondary camera

facilitated the overview and showed objects that are out of the field of view of the participant.

The wizard liked the integration of the scene control and general control buttons. This section functioned well, although user friendliness could still be increased. Buttons could be easier to read if button text were accompanied by icons .The wizard commented that she was eager to learn how to create her own buttons to trigger actions in future experiments. If possible, coding should be avoided and drag-and-drop functionality or a library of prefabricated components should help wizards to implement control buttons in future projects.

The 'live' interaction data can be useful for probing and real-time interpretations. Behavior can be triggered to draw live conclusions. The data generated in the experiment could be useful for post analysis of large samples. First, you can see in the data when something in the experiment went wrong. For example, the wizard can identify if text was hard to read for certain participants (mistakes or exceptionally long gaze times can be found in the data). Second, you can test if one scenario outperforms another. You could even see if data collection went wrong in an experiment by doing certain queries comparing different types of data streams. "You can analyze this data both horizontally as vertically, you can check if the user's eyes dilate when something explodes or you can check the sequences that people make (are they looking at the tablet after looking at the cooking pot). This is interesting to analyze what participants their next step will be."

Future improvements Two impactful improvements to the system were proposed. First, the system could be improved if the wizard is able to experience the same auditory stimuli as the participant at all times. Another useful adaptation would be a shadow mode where the wizard can prepare interactions before activating them for the participant. In the current version of the framework, the wizard and participant see the same scene at all times. Asynchronous interactions could alleviate the load of the wizard during the user test.

Subjective experience of the participant The semi-structured interview used questions from the QoE framework combined with in-depth questions to probe for a general impression of the participant. The answers of this semi-structured interview are restructured in paragraphs highlighting the positive and negative aspects of the experience. This subsection is finalised with a paragraph discussing future improvements suggested by the participant.

General feedback During the interview afterwards, the participant indicated that the tasks in VR were sufficiently developed to do a comparative assessment between the functionalities of the tablet interface and the voice assistant. Regarding the quality of experience, the participant commented that he felt immersed in the VR world. The participant also mentioned that the adaptation to the virtual environment occurred naturally. The participant was willing to wear all peripherals needed for the test (wireless HMD, battery, heart rate sensor). The use of

wireless peripherals in this test allowed for optimal freedom of movement for the participant. The participant did not notice that there was a 'wizard' controlling his actions. Afterwards it became clear to him that most interactions he had performed (e.g. filling a pot with water, baking bacon on the stove, controlling the tablet) were triggered by the 'wizard'.

The participant expressed his confusion regarding the interaction with some of the virtual objects. Specifically, actions that were not implemented in the kitchen were cutting vegetables and using the oven. This perceived lack of definition could be due to the open ended nature of Wizard of Oz tests. By displaying interactive and static objects in the same way, the participant assumed that all objects in the scene could be interacted with. Another discrepancy with real life was the timing of steps or tasks to be performed. For example, filling the pot with water goes instantly. The participant warns for a blind spot in the research due to not incorporating some realistic aspects, water flow, sunlight reflections. However, for the evaluation of the steps presented by the tablet this did not cause a problem, since the focus is on the smart interfaces. Apart from content related problems the VR test produced problems with visual focus, nausea and a small headache. This was caused mainly due to a bad calibration of distance between the lenses inside the headset. Being mentioned earlier in the paper, using hand-held controllers can contribute to a lack of realism or even cause interaction issues in a virtual environment. This indicates that user testing could benefit from more natural interaction modalities including haptic feedback. Consequently, confusion about controller input to interaction mapping could be minimised. The participant remarked that this discrepancy could be a bigger issue with persons who are less accustomed to testing new technologies. "It's in the details, interactions should be very detailed. Pressing buttons takes away from the naturalness of the interactions. Consequently, choices have to be made by the wizard which action he allows are performed well to move on with the experience."

Future improvements Making the interaction with controllers less ambiguous can improve the experience of the participant in VR. Furthermore, some usability issues could be resolved using a better onboarding strategy. For instance, a tutorial where the participant presses all controller buttons before the observations can start.

3.4 Comments on findings

Reflection on evaluation of the wizard When reflecting back on this first evaluation, it becomes clear that first time users are able to assess the functionality of the tool. Also, two usability issues were identified towards audio playback and option for an asynchronous workflow where the wizard can prepare future actions instead of working in real-time. When confronted with the logged output after the VR user test took place, both the participant and wizard reacted with the intention of using this data for comparing interactions and different scenarios. Further steps towards making it easier for the researchers to create interactions themselves using ready made modules will be important. Notwithstanding the prepared

interactions on the dashboard were used successfully, the wizard indicated being interested in adding interactions to the scene controls menu herself.

Reflection on evaluation of the participant Besides the evaluation of the wizard, the reaction of the participant towards this system was mostly positive. Some remarks were made regarding clarity about the distinction between interactive and static objects in the scene and a lack of realism during interactions. The comments indicate that onboarding is a crucial factor when introducing participants to VR experiments.

4 Discussion

In this paper, we presented the ExperienceDNA framework as an easy-to-use WoZ user testing tool in virtual reality with a strong emphasis on the capture of objective data. It aims to address three requirements: the framework facilitates the use of immersive environments (contexts), grants WoZ-style control to researchers and designers who use it, and aids in the capture of various behavioral and psychophysiological data streams. In doing so, we contend that the ExperienceDNA framework represents an overall useful methodology for testing products, services and systems, though it may be particularly suited to evaluate concepts and ideas that are difficult to test in real life. An example of such hard-to-prototype systems are so-called 'smart' systems, typically involving one or more IoT devices. Given the cost associated with developing a fully functional prototype, modelling these products and services in a VR simulation is budget- and cost-efficient, as virtual models can function in a 'black box' fashion: unlike the functionalities offered by a device, only the outcome needs to be modelled.

We believe that our framework has several other advantages over other traditional user testing methodologies. Where current user testing demands physical space, objects and people, VR user testing is possible in a virtual space – that can be modelled to represent any context – with virtual objects and people. This results in a faster workflow, a smaller development cost and an overall increase in versatility. The automatic logging of HCCI events, as well as psychophysiological markers allows researchers and designers to not only monitor the experience as it unfolds, but also helps them to analyze specific events post-hoc. Since interactions are automatically registered in the virtual world, these events can easily be synced with the behavioral (HCCI) and psychophysiological data streams. Finally, the virtual nature of our framework allows researchers to capture and replay experiences from the user's PoV, as well as from any number of (virtual) camera angles – a highly cumbersome and convoluted feat to be achieved in real-life user testing settings.

The modular structure of the ExperienceDNA framework allows us to gradually increase and improve functionalities. First, we will improve the current usability for the researcher designing a product testing experience.

From a more technical perspective, we will first make it possible for users to also physically touch objects while performing a VR user test. This new

implementation will make the use of controllers in VR obsolete since we will use wireless gloves with a kit to make mock-up objects (cardboard, foam or 3d printed) interactive. This kit (named 'reality blocks') of tangible buttons, connectors and wireless trackers will allow the wizard to build a physical WoZ test where the functionality and looks can be assessed in VR. The implementation of tangibles has two main advantages: the augmentation of realism enhances the experience for the user. It also results in deeper qualitative feedback.

Secondly, we will implement multiplayer interaction, which will be interesting to test multi-person experiences where, for example, an actor is involved to play along with the scenario. It can also be used to test a scenario with multiple test users at the same time.

Finally, with regard to the objective measurements, EEG will be added. EEG allows the wizard to have more accurate physiological data and better assessments can be done towards cognitive load and the emotional state of the participants.

In sum, we believe that the proposed system is a great step forward for interactive user testing of smart systems. This paper describes and validates the different aspects of a system for performing live VR user tests using the 'Wizard of Oz' method. This early validation was done through a pilot test of a smart kitchen use case.

References

1. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction **24**(6), 574–594 (2008)
2. Bradley, M.B., Miccoli, L.M., Escrig, M.a., Lang, P.J.: The pupil as a measure of emotional arousal and automatic activation. Psychophysiology **45**(4), 602 (2008). https://doi.org/10.1111/j.1469-8986.2008.00654
3. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of oz studies – why and how. Knowledge-Based Systems **6**(4), 258–266 (1993). https://doi.org/https://doi.org/10.1016/0950-7051(93)90017-N, `http://www.sciencedirect.com/science/article/pii/095070519390017N`, special Issue: Intelligent User Interfaces
4. Exner, K., Stark, R.: Validation of product-service systems in virtual reality. Procedia CIRP **30**, 96–101 (2015)
5. Geerts, D., De Moor, K., Ketyko, I., Jacobs, A., Van den Bergh, J., Joseph, W., Martens, L., De Marez, L.: Linking an integrated framework with appropriate methods for measuring qoe. In: 2010 Second international workshop on quality of multimedia experience (QoMEX). pp. 158–163. IEEE (2010)
6. Gengnagel, C., Nagy, E., Stark, R.: Rethink! Prototyping: Transdisciplinary Concepts of Prototyping. Springer (2015)
7. Hassenzahl, M., Burmester, M., Koller, F.: Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In: Mensch & computer 2003, pp. 187–196. Springer (2003)
8. Inc., A.: Mixamo - animate 3d characters for games, film, and more (2021), `https://www.mixamo.com/`
9. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H., Koo, B.H.: Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry investigation **15**(3), 235 (2018)

10. Kirkevold, M., Bergland, A.: The quality of qualitative data: Issues to consider when interviewing participants who have difficulties providing detailed accounts of their experiences. International Journal of Qualitative Studies on Health and Well-being **2**(2), 68–75 (2007). https://doi.org/10.1080/17482620701259273, https://doi.org/10.1080/17482620701259273
11. Klemmer, S.R., Sinha, A.K., Chen, J., Landay, J.A., Aboobaker, N., Wang, A.: Suede: a wizard of oz prototyping tool for speech user interfaces. In: Proceedings of the 13th annual ACM symposium on User interface software and technology. pp. 1–10 (2000)
12. Lane, R.D., McRae, K., Reiman, E.M., Chen, K., Ahern, G.L., Thayer, J.F.: Neural correlates of heart rate variability during emotion. Neuroimage **44**(1), 213–222 (2009)
13. Lawson, G., Salanitri, D., Waterfield, B.: Future directions for the development of virtual reality within an automotive manufacturer. Applied ergonomics **53**, 323–330 (2016)
14. LLC, M.: Turbosquid - 3d models for professionals (2021), https://www.turbosquid.com/
15. Magliacano, A., Fiorenza, S., Estraneo, A., Trojano, L.: Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. Neuroscience Letters **736**, 135293 (2020). https://doi.org/https://doi.org/10.1016/j.neulet.2020.135293, http://www.sciencedirect.com/science/article/pii/S0304394020305632
16. Mandal, B., Li, L., Wang, G.S., Lin, J.: Towards detection of bus driver fatigue based on robust visual analysis of eye state. IEEE Transactions on Intelligent Transportation Systems **18**(3), 545–557 (2017). https://doi.org/10.1109/TITS.2016.2582900
17. Nardelli, M., Valenza, G., Greco, A., Lanata, A., Scilingo, E.P.: Recognizing emotions induced by affective sounds through heart rate variability. IEEE Transactions on Affective Computing **6**(4), 385–394 (2015)
18. Noble, H., Smith, J.: Issues of validity and reliability in qualitative research. Evidence-based nursing **18**(2), 34–35 (2015)
19. Ottosson, S.: Virtual reality in the product development process. Journal of Engineering Design **13**(2), 159–172 (2002)
20. Pardo, C., Ivens, B.S., Pagani, M.: Are products striking back? the rise of smart products in business markets. Industrial Marketing Management **90**, 205–220 (2020). https://doi.org/https://doi.org/10.1016/j.indmarman.2020.06.011, http://www.sciencedirect.com/science/article/pii/S001985011930330X
21. Piquado, T., Isaacowitz, D., Wingfield, A.: Pupillometry as a measure of cognitive effort in younger and older adults. Psychophysiology **47**(3), 560–569 (2010). https://doi.org/https://doi.org/10.1111/j.1469-8986.2009.00947.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.2009.00947.x
22. Rahman, L.F., Ozcelebi, T., Lukkien, J.: Understanding iot systems: A life cycle approach. Procedia Computer Science **130**, 1057–1062 (2018). https://doi.org/https://doi.org/10.1016/j.procs.2018.04.148, http://www.sciencedirect.com/science/article/pii/S1877050918305106, the 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops
23. Steenkamp, J.B.E., De Jong, M.G., Baumgartner, H.: Socially desirable response tendencies in survey research. Journal of Marketing Research **47**(2), 199–214 (2010)
24. Van Acker, B.B., Parmentier, D.D., Vlerick, P., Saldien, J.: Understanding mental workload: from a clarifying concept analysis toward an implementable framework. Cognition, technology & work **20**(3), 351–365 (2018)

25. Van Hove, S., De Letter, J., De Ruyck, O., Conradie, P., All, A., Saldien, J., De Marez, L.: Human-computer interaction to human-computer-context interaction: Towards a conceptual framework for conducting user studies for shifting interfaces. In: International Conference of Design, User Experience, and Usability. pp. 277–293. Springer (2018)
26. Van Kets, N., Moens, B., Bombeke, K., Durnez, W., Maes, P.J., Van Wallendael, G., De Marez, L., Leman, M., Lambert, P.: Art and science interaction lab – a highly flexible and modular interaction science research facility. arXiv open-access archive 2101.11691 (2021), https://arxiv.org/abs/2101.11691
27. Van Someren, M., Barnard, Y., Sandberg, J.: The think aloud method: a practical approach to modelling cognitive. London: AcademicPress (1994)
28. Wang, P., Sibi, S., Mok, B., Ju, W.: Marionette: Enabling on-road wizard-of-oz autonomous driving studies. In: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction. pp. 234–243 (2017)
29. van der Wel, P., van Steenbergen, H.: Pupil dilation as an index of effort in cognitive control tasks: A review. Psychonomic bulletin & review **25**(6), 2005–2015 (2018)
30. Williams, M.D., Rana, N.P., Dwivedi, Y.K.: The unified theory of acceptance and use of technology (utaut): a literature review. Journal of enterprise information management (2015)
31. Zheleva, A., Durnez, W., Bombeke, K., Van Wallendael, G., De Marez, L.: Seeing is believing: The effect of video quality on quality of experience in virtual reality. In: 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–4. IEEE (2020)