Towards Virtual Reality Crisis Simulation as a Tool for Usability Testing of Crisis Related Interactive Systems

Kristian Rother, Hamm-Lippstadt University of Applied Sciences, Lippstadt, Germany

Inga Karl, Hamm-Lippstadt University of Applied Sciences, Lippstadt, Germany

Simon Nestler, Hamm-Lippstadt University of Applied Sciences, Lippstadt, Germany

ABSTRACT

Usability testing is expensive in some domains due to the resource requirements that go hand in hand with taking a complex context of use into account. Crisis-related research is one such domain, typically requiring the reenactment of an extensive crisis scenario. To lessen the resource requirements and provide a more flexible setup geared towards testing, crisis scenarios can be reconstructed as virtual reality simulations. This paper outlines the development of an initial prototype of such a simulation following the design science method. The prototype is used to test if injecting an item that will be tested into the simulation affects the realism of the virtual reality crisis simulation. The realism was measured in a within-subject experiment and equivalence tests showed that injecting a representation of a simple app had no significant influence on the realism of the simulation.

Keywords: Crisis Management, Design Science, Interactive Systems, Usability Testing, Virtual Reality

1. INTRODUCTION AND METHOD

This paper outlines the development of a virtual reality crisis simulation (VRCS) prototype to enable a novel form of usability testing for crisis-related interactive systems. The nature of the research, namely the development of an artifact (instantiation) that solves a previously unsolved problem, suggests a design science (DS) approach (Hevner, March, Park, & Ram, 2004). While DS has mostly been discussed in general information systems research during the last ten years (Hevner et al., 2004; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007; Offermann, Levina, Schönherr, & Bub, 2009; Österle et al., 2011) it is equally applicable for human-computer in-

DOI: 10.4018/IJISCRAM.2015070103

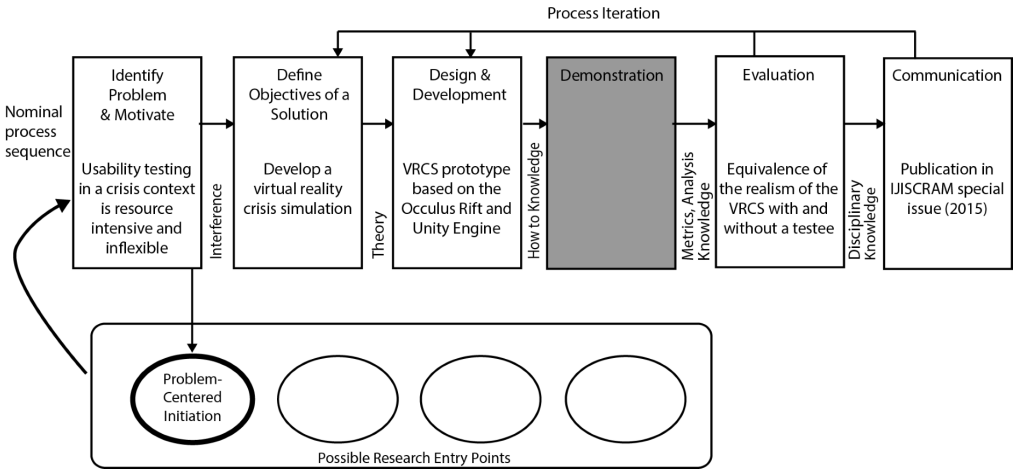
teraction research projects (Hevner & Zhang, 2011; Tarantino & Spagnoletti, 2013) and in crisis contexts (Schryen & Wex, 2015). The design science process model (DSPM) suggested by Peffers et al. was chosen because it synthesizes different previous models and "defines a mental model for presenting and evaluating DS research" (2007). The remainder of this paper is structured according to the DSPM. Figure 1 shows the adapted process.

The research can be classified as a *problem-centered initiation* because the construction of the artifact was motivated by an identified problem, namely that usability testing based on real crisis simulations is resource intensive and inflexible. The activities *problem identification and motivation*, *definition of objectives of a solution* and *design and development* were conducted in sequential order. While a general problem and general objectives for a solution were identified the research question of this paper relates to the identified sub-problem that injecting an item into the VRCS for testing purposes can influence the realism of the VRCS. Accordingly, the *evaluation* activity is not aimed at the identified general problem but rather at advancing the prototype to a stage where that problem can eventually be tackled. To reach that stage the evaluation conducted in this paper is aimed at the identified sub-problem. The *demonstration* activity was not conducted because the prototype is still at an early stage and cannot be used to solve concrete problems yet. The publication of this paper and the anticipated discussions serve as the beginning of the *communication* activity.

2. PROBLEM IDENTIFICATION AND MOTIVATION

Professionals (emergency medical services, fire and rescue service, police) use crisis-related interactive systems during their work processes. Citizens can uses crisis-related interactive systems like apps or web applications to help them prepare for crises or in case a crisis breaks out. However, crisis situations are a complex domain. In complex domains, the context of use has to be taken into account for usability testing (DIS, 2009; Redish, 2007). Consequently, usability testing of these systems in the lab is necessary but not sufficient. Human-computer interaction methods that focus on the context of use such as contextual inquiry (Holtzblatt & Jones, 1993) and field research methods (Kantner, Sova, & Rosenbaum, 2003; Rosenbaum & Kantner, 2007)

Figure 1. Adapted design science process model according to Peffers et al. (2007)



are typically conducted during common work processes or day to day activities. Unmodified, these methods by themselves are not suitable for crisis-related interactive systems because a crisis happens unexpectedly and is not part of the routine work or typical daily activities. Even if a crisis would occur while these methods are used they could negatively affect the outcome of the crisis, for example by disturbing domain experts during their tasks. Therefore, these activities should be stopped immediately if a crisis breaks out.

Due to the outlined problems, field exercises also known as crisis simulations (Boin, Kofman-Bos, & Overdijk, 2004; Kleiboer, 1997) are used for usability testing of crisis-related interactive systems (Nestler, 2014). These *real crisis simulations*[1] are resource intensive because they require actors, extras, vehicles, equipment and space (see Figure 2). Additionally, changing variables during real crisis simulations, which is often desired for usability testing purposes, is not easy.

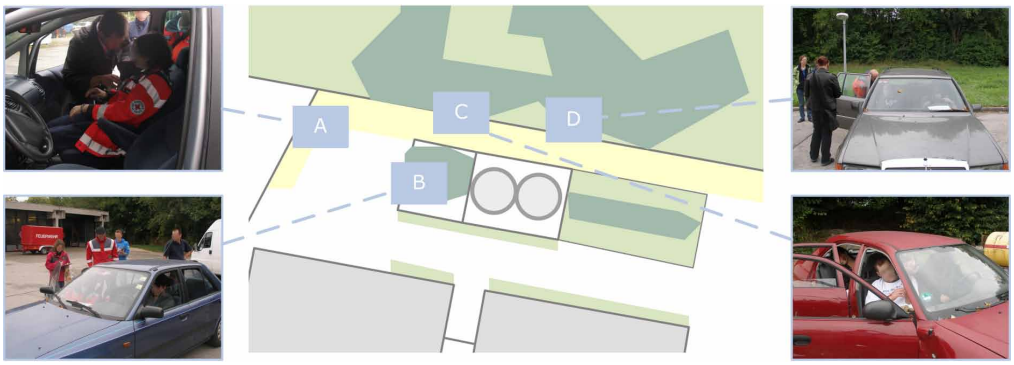
During the ongoing research project INTERKOM[2] the outlined problem of taking a complex context of use into account became evident while prototyping a mobile app for the communication between citizens and crisis-professionals. This motivated the outlined research.

3. DEFINITION OF THE OBJECTIVES FOR A SOLUTION

To counteract the resource requirements of real crisis simulations these simulations can be transferred into virtual worlds. The resulting simulations are *virtual reality crisis simulations (VRCS)*. The development and use of VRCS is associated with costs. To provide a benefit for usability testing these costs have to be lower than the resources saved by using the VRCS. It is currently assumed as a working hypothesis that this can be achieved. Under this assumption objectives can be identified based on the initial conditions of the usability testing setup (i.e., are real crisis simulations already used or are they not used at all so far):

1. Real crisis simulations are not used for usability testing yet:
   - **Problem:** Real crisis simulations are too resource intensive and as a result at most lab based usability tests are conducted. The crisis context is not taken into account;
   - **Objective 1:** If some additional resources are available but not sufficient to conduct an entire real crisis simulation they can be used to conduct VRCS and as a result the crisis context is taken into account;

Figure 2. Excerpt of the overview of a crisis simulation (Nestler, 2014) which shows some of the required vehicles and indicates the space requirements (faces blurred)



2. Real crisis simulations are already used for usability testing:
    - **Problem:** Due to the resource requirements of running an entire real crisis simulation both the number of design solutions that can be tested and the scenarios in which they can be tested are limited;
    - **Objective 2:** VRCS can serve as a pre-test to reduce the number of design solutions that have to be tested in the real crisis simulation. VRCS can also be used to pre-configure the real crisis simulation to fit the testing needs;
    - **Objective 3:** VRCS replace the real crisis simulations entirely. Due to the reduced resource requirements more scenarios can be tested or scenarios can be tested more in depth and varied easily.

Since real crisis simulations are common practice and accepted as useful as a general tool outside the realm of usability testing objective 3 is excluded as a candidate. The initial goal is to develop a prototype that can eventually fulfill objective 1 or objective 2.

The development of the VRCS and its integration into the usability testing process should be *possible* and *feasible* (Peffers et al., 2007). It is possible in principle because virtual reality has been used successfully for other purposes like training in different domains (Orr, Mallet, & Margolis, 2009; Seymour et al., 2002) therapy (Riva, 2005) and way finding (Tang, Wu, & Lin, 2009). Furthermore, virtual prototypes (Kuutti et al., 2001) and virtual worlds have been suggested as potential tools for usability testing (Chalil Madathil & Greenstein, 2011). However the representation of the interactive system for which the usability test will be conducted in the VRCS (henceforth referred to as *testee*) could influence the realism of the simulation. Therefore, a necessary first step is to conduct a suitable evaluation to make sure that the presence of the testee doesn't influence the realism of the VRCS. Thus, the initial prototype was built with this goal in mind and serves as tool to conduct the required equivalence tests. Consequently, the research question of this paper is: "Can a testee be injected into a VRCS without influencing the realism of that VRCS?"

To ensure that the development of the VRCS is feasible the scope was limited by concentrating on a single crisis scenario and by creating this scenario ad hoc without the direct consultation of domain experts. The selected scenario was a *prolonged power outage* because it is described in literature (Petermann, Bradke, Lüllmann, Poetzsch, & Riehm, 2014) and the scenario is used in the INTERKOM research project. This ensures access to domain experts for future iterations of the VRCS. Limiting the testee to a simple virtual recreation of a mobile app inside the VRCS further reduced the scope.

4. DESIGN AND DEVELOPMENT

The developed prototype served as a basis to test the influence of the testee on the realism of the simulation and to get a general feeling for the feasibility of creating a VRCS. The two major design decisions were the transformation of the crisis scenario into a VRCS and the representation of the testee. The simulation was limited to a small city that was constructed from scratch by using preexisting city components such as buildings and streets. The city is in a state of power outage during the entire simulation with limited sound effects where appropriate and additional small visual indications of the power outage such as garbage that wasn't picked up. Obstacles that strategically limit the route to a predetermined one were used which means that one essentially walks from start to finish within the city while still retaining a feeling of free movement. This feeling of free movement was not measured as part of the experiment and only

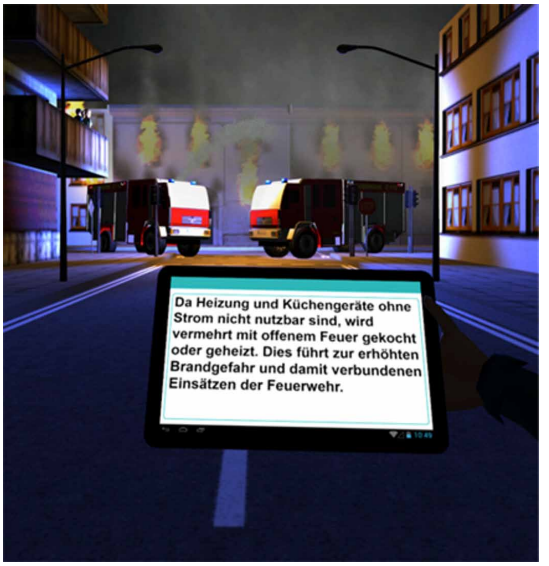
confirmed informally in talks when the VRCS was shown to another group of students later as a technology demo (i.e. not in an experimental setting).

With one exception, buildings cannot be entered. A playback component for a before-after effect of ten identified events, like increased accidents due to the lack of working traffic lights, was also added. These effects are triggered upon entering certain zones. A scene from the VRCS that shows that people will resort to open fires due to the lack of electric heating is depicted in Figure 3. The testee was prototyped as a simple virtual recreation of a mobile app inside the VRCS. A tablet and a virtual hand that holds it fade in at the bottom of the simulation when a trigger points is passed. There is no interaction with the virtual tablet. The device simply shows text related to the ongoing crisis. For example, when the trigger point for a burning apartment is passed, the tablet fades in and shows a note that explains that open fires are used for cooking and heating due to the power outage and as a result more fires break out (this is depicted in Figure 4). The German text translates to "Because electric heating and kitchen appliances cannot be used without electricity the use of open fires for cooking and heating increases. This leads to an increased risk of fires and corresponding increases in fire fighter deployments."

The two technology choices for the development of the VRCS were (a) picking the virtual reality hardware and (b) picking a 3D-engine. While there are many possible virtual reality hardware combinations an approach based on the Oculus Rift Development Kit 2 and an Xbox 360 controller was selected because this hardware was already available and integrated into the teaching process[3]. Furthermore, this setup can be used with a laptop, which makes the solution portable. Likewise there are many different 3-D engines. The Unity Engine was selected because it is free, wide spread, supports direct rendering of Oculus VR views and is already used in other projects at the Hamm-Lippstadt University of Applied Sciences.

Figure 3. A scene from the VRCS depicting a power outage

Figure 4. A scene with the virtual tablet (testee) from the VRCS depicting a power outage



5. EVALUATION

To test if the injection of a testee influences the realism of the VRCS, the realism was measured for the VRCS with and without the testee. This chapter outlines the experiment and the equivalence tests that were conducted to answer the research question.

5.1. Participants

32 participants were recruited from undergraduate students of the Hamm-Lippstadt University of Applied Sciences by asking for participation during their lectures. 26 participants were male and 6 participants were female. The age of the participants was between 18 and 29 (M = 23.28, SD = 2.67). The convenience sampling of the participants lead to a participant pool that has a high pre-education in the field of virtual reality as it is part of their studies (47% of the participants had previous knowledge about VR).

The participants were randomly assigned to two groups of 16 members each without any pretests or matching. No power-analysis to determine the required sample size was conducted prior to the experiment since the sample-size was limited to the available students regardless.

5.2. Variables

The nominal-scale variable *simulation type* (with the two levels "without testee" and "with testee") is the treatment variable. The four variables *scene realism*, *audience behavior*, *sound realism* and *realism of the VR-application* are the dependent variables. They are interval scaled from 1 (very low realism) to 5 (very high realism). The dependent variables are grand means of corresponding sub-items from the questionnaire that was used.

To control for external factors the experiment was conducted in the same conditions with each participant. All participants used the same equipment and computer and were tested in the

same room under the same lighting and sound conditions. The individual characteristics of the participants are treated as random variables.

5.3. Instruments and Materials

The experiment was conducted in a dedicated laboratory with constant lighting and no disturbing sounds. The same computer and hardware (Oculus Dev Kit 2, X-Box-Controller, 3D-Headphones) was used for all participants. A simple self-made questionnaire to collect demographic information was used.

The realism questionnaire that was used for the evaluation[4] is based on the Simulation Realism Scale (Poeschl & Doering, 2013) which is in turn based on the presence questionnaire by Witmer and Singer (1998). The Simulation Realism Scale was adapted to meet the needs of this evaluation in the following manner:

- The items relating to Audience Appearance were removed since Audience Appearance showed insufficient reliability in the original study by Poeschl and Doering and is not deemed essential for this research;
- Additional sound items were added as suggested by Poeschl and Doering. For the extension the sound items from the revised presence questionnaire (UQO Cyberpsychology Lab, 2004) were used;
- The realness items from Schubert (2003) were added;
- All items were adapted to five levels if necessary.

The resulting questionnaire contains a total of 17 items in 4 categories. Each category represents one of the dependent variables. The questionnaire is a summated rating scale. All items except the original sound item from Poeschl and Doering have five levels ordered from least agreement (left) to most agreement (right). Consequently, that sound item was collected but not used for the calculation of the grand mean for *sound realism*[5].

5.4. Experiment Design and Procedure

Due to the limited number of participants, the experiment was conducted as a within-subject design. To avoid an ordering effect, a counterbalanced design in which participants were randomly assigned to one of the two groups was chosen. Each group consisted of 16 participants. Following Campbell, Stanley, and Gage (1963) the design is depicted in Figure 5. All sessions were conducted individually. Participants who were assigned to group A started with the *simulation type* set to "without testee" (X_1) followed by *simulation type* set to "with testee" (X_2). The reverse setup was used for group B. O_D depicts when the socio-demographic questionnaire was conducted, O_R depicts when the realism questionnaire was conducted and CD depicts when the cool down period happened. The length of the lines indicates the duration of the exposure to the VRCS and the cool down time respectively. The cool down time was set at five minutes. Unfortunately the exact exposure time wasn't collected. On average it was 10 minutes for X_1 and

Figure 5. Design of the experiment following (Campbell et al., 1963)

Group A: O_D X_1 _____ O_R CD _____ X_2 _____ O_R
Group B: O_D X_2 _____ O_R CD _____ X_1 _____ O_R

X_2 for a total of 20 minutes of exposure with the cool down break in between (this is indicated by the length of the lines in relation to each other).

The same experimental procedure was followed for each participant. The participants were seated during the entire experiment. The experimenter greeted each participant and explained the health risks of the VRCS[6]. Afterwards a socio-demographic questionnaire was handed out and upon completion the VRCS was explained to the participant. After making sure that the HMD was set up properly and that the participant understood how to navigate the VRCS the first run through the VRCS begun. Upon completion of the first run, the participant filled out the realism questionnaire, which was collected upon completion. After a cool down period of five minutes the *simulation type* was switched and the participant conducted a second run through the VRCS. Upon completion of the second run a second realism questionnaire was filled out and the experimenter collected the questionnaire and thanked the participant.

5.5. Measures

Table 1 lists the descriptive statistics of the dependent variables for the conducted experiment as taken from the realism questionnaires on a scale from 1 (very low realism) to 5 (very high realism).

Since the objective of the evaluation was to see if adding a testee to the VRCS leaves the realism equivalent to the VRCS without the testee an equivalence test (Kirkwood & Westlake, 1981; D. L. Schuirmann, 1981) was conducted. The formulated research hypothesis (equivalence assumption) is "after injecting the testee into the VRCS, the realism of the VRCS is equivalent to the realism of the VRCS without the testee" which can be formulated for the scene realism, audience behavior, sound realism and the realism of the VR-application. The corresponding null hypotheses were tested with the Two One-Sided Tests procedure (TOST) (D. J. Schuirmann, 1987) for the four dependent variables *scene realism*, *audience behavior*, *sound realism* and *realism of the VR-application.* Following Wellek (Wellek, 2010, p. 11), the null hypothesis of nonequivalence has the general form:

Table 1. Descriptive statistics of the realism questionnaire

Variable	N	M	SD	Min	Max
Scene Realism (without testee)	32	3.92	0.54	2.60	5.00
Scene Realism (with testee)	32	3.84	0.54	2.40	5.00
Audience Behavior (without testee)	32	3.42	0.85	1.50	4.75
Audience Behavior (with testee)	32	3.47	0.80	1.25	4.75
Sound Realism (without testee)	32	4.38	0.63	2.33	5.00
Sound Realism (with testee)	32	4.40	0.60	2.67	5.00
Realism of the VR-Application (without testee)	32	2.67	0.72	1.25	4.50
Realism of the VR-Application (with testee)	32	2.61	0.69	1.50	4.25

$$\mathrm{H}: \theta \leq \theta_o - \varepsilon_1 \text{ or } \theta \geq \theta_o + \varepsilon_2$$

while the equivalence assumption has the general form:

$$\mathrm{K}: \theta_o - \varepsilon_1 < \theta < \theta_o + \varepsilon_2$$

The chosen epsilon value was 0.25 as this represents the suggested strict value for paired t-tests setups (Wellek, 2010, p. 16). The degree of dissimilarity θ was chosen to be the difference between the dependent variable before and after the independent variable was changed[7]. The reference value θ was set to zero as this represents the case when the distributions under investigation are equal. Table 2 summarizes the results of the equivalence tests for all dependent variables.

5.6. Discussion

Since all TOST-Confidence Intervals fall within the range of [-0.25; 0.25] the null hypothesis of nonequivalence can be rejected for all four cases. Thus, the VRCS with and without the testee is equivalent regarding all relevant realism measures. This indicates, that it is possible to add a simple virtual app to the developed VRCS without disturbing the realism. This result is a first step towards being able to test said app in the VRCS.

It was assumed that the constructed questionnaire measures the realism of the VRCS reliably and that realism is a good base measure for a VRCS. The idea behind the latter assumption is that if the VRCS is perceived as real, the results of usability tests conducted within the VRCS could be transferable to the real world. Existing research of the applicability of VR learning to the real world (Alexander, Brunyé, Sidman, & Weil, 2005) could be a good starting point to justify this assumption. However, the potential gap between the general realism of the VRCS and the realism of the crisis scenario was not addressed. Experiential Design applied to Virtual Environments (Chertoff, Schatz, McDaniel, Bowers, & others, 2008) in conjunction with domain experts would be a viable approach. Lastly it is necessary to construct and validate a more sophisticated questionnaire or other means of measuring the realism of a VRCS while taking the realism of the scenario into account.

Table 2. Results of the equivalence tests

Dependent Variable	Mean of the Difference	Standard Error of the Difference	Confidence Interval (1-α)	TOST-Confidence Interval (1-2α)	p-Value
Scene Realism	-0.081	0.063	[-0.188, 0.025]	[-0.170, 0.007]	.006
Audience Behavior	0.047	0.110	[-0.140, 0.233]	[-0.108, 0.202]	.037
Sound Realism	0.021	0.047	[-0.060, 0.101]	[-0.046, 0.088]	< .001
Realism of the VR-Application	-0.063	0.070	[-0.181, 0.056]	[-0.161, 0.036]	.006

Note: The alpha level was .05 and an epsilon of .25 was chosen. The order for the mean of the difference and standard error was kept constant. It was set to be the value of the case "with testee" minus the value for the case "without testee".

6. OUTLOOK

This paper outlined the general motivation for the development of a VRCS prototype as a means to solve the problem of taking the crisis context into account in a less resource intensive way than relying solely on real crisis simulations. It defined objectives for a solution and identified the sub-problem that injecting a testee into the VRCS could influence the realism of the VRCS. A high level view of the design and technology choices was sketched.

To answer the research question "Does the injection of a testee into a VRCS influence the realism of that VRCS?" equivalence tests with regards to the realism of the VRCS were conducted. The tests showed that the VRCS with and without the virtual app were equivalent with regards to *scene realism*, *audience behavior*, *sound realism* and *realism of the VR-application.*

While it proved to be feasible to build the prototype within a reasonable timeframe the construction of the first version of the VRCS and the conducted experiment have already revealed some defects and ideas for further improvements. Thus, the next step is to iteratively improve the VRCS before eventually moving on towards the goal of conducting usability tests inside the VRCS. Input from both domain experts in crisis management and human-computer interaction (HCI) specialists is needed and welcome to achieve this. To kick-start this communication activity (see Figure 1) we list some identified weaknesses and some of our own ideas for further discussion.

Cybersickness: A major drawback is that the well-known problem of cybersickness (Davis, Nesbitt, & Nalivaiko, 2014; McCauley & Sharkey, 1992) occurred during the early stages of prototyping the VRCS. Frame rate improvements offset the initial issues and none of the participants of the experiment reported any problems with cybersickness. However there is no guarantee that the issue is fully solved and further tests with a more heterogeneous group of participants are needed. The next iteration of the prototype will focus on following the best practices (Yao et al., 2014) that lead to a reduction in cybersickness even more in depth. Even if this problem can be reduced it may still have influence on the design choices of future experiments as cybersickness gets worse with prolonged exposure (Kennedy, Stanney, & Dunlap, 2000). For example, within-subject designs require a longer exposure to the VRCS than between-subject designs, which could lead to a higher number of subjects dropping out during the experiment due to the experienced sickness.

Relationship between the VRCS and the interactive system that will be tested: The exact specifications of the mobile app that will be tested within the VRCS are currently being developed in the INTERKOM research project. For now a simplified virtual tablet served as a placeholder and proved useful for equivalence testing. However the actual app that will be tested influences the design of the VRCS scenario and as such knowing the testee is a precondition for further advances. Additionally, the equivalence results cannot be generalized and a similar experiment needs to be conducted for every VRCS-testee pair that one designs.

App representation: There are multiple possible ways of representing an app in a VRCS. Mirroring the screen of the actual device onto a virtual representation of the device or a recreation of the app within the VRCS (as was done in a simplified manner for this version of the VRCS) are two examples. One of the key problems is that it is hard to impossible to use the mobile device while wearing a head-mounted display (HMD). Even if that wasn't the case the interaction with the device provides interesting challenges. A transfer of ideas from the use of touch- or device-based gesture control (Turk, 2015) could prove fruitful. The problem of interacting with the mobile device can be mitigated by moving form a HMD to a CAVE (Cruz-Neira, Sandin, & DeFanti, 1993). Even if a CAVE is used, a HMD based VRCS can serve as a prototyping environment for the CAVE as long as the underlying technology (e.g. 3D-Engine) is compatible.

Crisis representation: The selected crisis scenario was built ad hoc without the input of domain experts based on the scenario description found in Petermann et al. (2014). Since most test subjects will not have experienced this crisis situation it is hard to measure how realistic the reconstruction actually is. The next iteration will involve domain experts and rely on their feedback about the realism of the VRCS. Three crisis scenarios (prolonged power outage, pandemic and bio terrorism) are currently being developed in cooperation with domain experts in the INTERKOM research project. Other alternatives are seeking subjects that have lived through the specific crisis and relying on their memory of the past experiences (which is limited to types of crises that have already happened) or developing a generic questionnaire to evaluate if the crisis scenario felt real. The questionnaire by Chertoff, Goldiez, and LaViola Jr (2010) could be a good starting point. Additionally, content development was an afterthought and mostly based on what was available for free and some intuition regarding the construction of the city. A more rigorous approach following established principles (Isdale, Fencott, Heim, & Daly, 2002) is planned for a future iteration. Another challenge is the potential use of real-time data in the VRCS. Wang, Bishop, and Stock (2009) provide an overview of a framework for the integration of real-time data in collaborative virtual environments.

Lack of interaction: Currently users can only walk through the city by using a gamepad or keyboard in combination with the direction they look in. The simulation ends when the final destination is reached which leads to a fade to black. While this is acceptable for a first prototype the next steps need to focus on actual actions that are taken during a crisis. These interactions depend on both the testee and the crisis scenario, which have to be developed. In the future, the crisis scenario could be split into smaller units that can be used to measure the performance during concrete tasks within the VRCS like using the app to navigate to collection points, getting warned by the app when near an exclusion zone or using the app to communicate with the rescue service via text message. The measurements of performance can be based on the work done by Lampton, Bliss, and Morris (2015). The addition of a walking device like the Virtuix Omni as a replacement for the need to walk via a gamepad could be considered.

Usability testing of the VRCS: From an HCI point of view we jumped straight into the development step of the EN ISO 9421-11 process (2009). While this is a good compromise when developing software for your own use or to quickly see how much time it takes to build a prototype, a usability test of the VRCS itself has to be conducted especially if it is to be used by other researchers. Bowman et al. (2002), Gabbard et al. (1999) and Tromp et al. (2003) provide some insight into how this could be done. Sutcliffe and Gault (2004) provide some useful heuristics.

Comparison of VRCS and other methods of generating a crisis context: To test the assumed working hypothesis of resource savings VRCS have to be compared to other methods of creating a crisis context like non-VR 3D simulations, storytelling, paper based descriptions, low-fidelity crisis-simulation and real crisis simulations.

ACKNOWLEDGMENT

This research is supported by a grant from the German Federal Ministry of Education and Research (BMBF) as part of the INTERKOM project (No. 13N1005, 01/2014 – 12/2016). The paper is an enhanced and improved version of a paper presented at the 2015 Workshop "KritischeHCI" (Rother, Karl, & Nestler, 2015). We would like to thank Alexander Giesbrecht, Sandra Jürgensmeier, Inge Kling, Sinan Mert and Dennis Ziebart for their contributions.

REFERENCES

Alexander, A. L., Brunyé, T., Sidman, J., & Weil, S. A. (2005). From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games. *DARWARS Training Impact Group*, *5*, 1–14.

Boin, A., Kofman-Bos, C., & Overdijk, W. (2004). Crisis simulations: Exploring tomorrow's vulnerabilities and threats. *Simulation & Gaming*, *35*(3), 378–393. doi:10.1177/1046878104266220

Bowman, D. A., Gabbard, J. L., & Hix, D. (2002). A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. *Presence (Cambridge, Mass.)*, *11*(4), 404–424. doi:10.1162/105474602760204309

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Boston.

Chalil Madathil, K., & Greenstein, J. S. (2011). Synchronous remote usability testing: a new approach facilitated by virtual worlds. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2225–2234). ACM. doi:10.1145/1978942.1979267

Chertoff, D. B., Goldiez, B., & LaViola, J. J., Jr. (2010). Virtual Experience Test: A virtual environment evaluation questionnaire. *Proceedings of the Virtual Reality Conference (VR)* (pp. 103–110). IEEE. doi:10.1109/VR.2010.5444804

Chertoff, D. B., Schatz, S. L., McDaniel, R., & Bowers, C. et al. (2008). Improving presence theory through experiential design. *Presence (Cambridge, Mass.)*, *17*(4), 405–413. doi:10.1162/pres.17.4.405

Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. *Proceedings of the 20th annual conference on Computer graphics and interactive techniques* (pp. 135–142). ACM. doi:10.1145/166117.166134

Cyberpsychology Lab, U. Q. O. (2004). Presence Questionnaire Revised by the UQO Cyberpsychology Lab. Retrieved from http://w3.uqo.ca/cyberpsy/docs/qaires/pres/PQ_va.pdf

Davis, S., Nesbitt, K., & Nalivaiko, E. (2014). A Systematic Review of Cybersickness. *Proceedings of the 2014 Conference on Interactive Entertainment* (pp. 8:1–8:9). New York, NY, USA: ACM. http://doi.org/doi:10.1145/2677758.2677780

DIS. I. (2009). 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. International Standardization Organization (ISO). Switzerland.

Gabbard, J. L., Hix, D., & Swan, J. E. (1999). User-centered design and evaluation of virtual environments. *Computer Graphics and Applications, IEEE*, *19*(6), 51–59. doi:10.1109/38.799740

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly*, *28*(1), 75–105.

Hevner, A. R., & Zhang, P. (2011). Introduction to the AIS THCI Special Issue on Design Research in Human-Computer Interaction. *AIS Transactions on Human-Computer Interaction*, *3*(2), 56–61.

Holtzblatt, K., & Jones, S. (1993). Contextual inquiry: A participatory technique for system design. *Participatory Design: Principles and Practices*, 177–210.

Isdale, J., Fencott, C., Heim, M., & Daly, L. (2002). Content design for virtual environments. In K. Hale & K. Stanney (Eds.), *Handbook of virtual environments: Design, implementation, and applications* (2nd ed., pp. 519–532). Boca Raton, FL: CRC Press.

Kantner, L., Sova, D. H., & Rosenbaum, S. (2003). Alternative methods for field usability research. *Proceedings of the 21st annual international conference on Documentation* (pp. 68–72). ACM.

Kennedy, R. S., Stanney, K. M., & Dunlap, W. P. (2000). Duration and exposure to virtual environments: Sickness curves during and across sessions. *Presence (Cambridge, Mass.)*, *9*(5), 463–472. doi:10.1162/105474600566952

Kirkwood, T. B., & Westlake, W. J. (1981). Bioequivalence testing–a need to rethink. *Biometrics*, *37*(3), 589–594. doi:10.2307/2530573

Kleiboer, M. (1997). Simulation methodology for crisis management support. *Journal of Contingencies and Crisis Management*, *5*(4), 198–206. doi:10.1111/1468-5973.00057

Kuutti, K., Battarbee, K., Sade, S., Mattelmaki, T., Keinonen, T., Teirikko, T., & Tornberg, A.-M. (2001). Virtual prototypes in usability testing. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences* (p. 7). http://doi.org/ doi:10.1109/HICSS.2001.926545

Lampton, D. R., Bliss, J. P., & Morris, C. S. (2015). Human performance measurement in virtual environments. In K. Hale & K. Stanney (Eds.), *Handbook of virtual environments: Design, implementation, and applications* (2nd ed., pp. 749–780). Boca Raton, FL: CRC Press.

McCauley, M. E., & Sharkey, T. J. (1992). Cybersickness: Perception of self-motion in virtual environments. *Presence (Cambridge, Mass.)*, *1*(3), 311–318. doi:10.1162/pres.1992.1.3.311

Nestler, S. (2014). Evaluation der Mensch-Computer-Interaktion in Krisenszenarien / Evaluating human-computer-interaction in crisis scenarios. *I-Com*, *13*(1), 53–62. doi:10.1515/icom-2014-0008

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009). Outline of a design science research process. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (p. 7). ACM. doi:10.1145/1555619.1555629

Orr, T. J., Mallet, L. G., & Margolis, K. A. (2009). Enhanced fire escape training for mine workers using virtual reality simulation. *Mining Engineering*, *61*(11), 41.

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., & Krcmar, H. … Sinz, E. J. (2011). Memorandum on design-oriented information systems research. *European Journal of Information Systems, 20*(1), 7–10. http://doi.org/10.1057/ejis.2010.55

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. doi:10.2753/MIS0742-1222240302

Petermann, T., Bradke, H., Lüllmann, A., Poetzsch, M., & Riehm, U. (2014). *What Happens During a Blackout: Consequences of a Prolonged and Wide-ranging Power Outage*. BoD–Books on Demand.

Poeschl, S., & Doering, N. (2013). The German VR Simulation Realism Scale–psychometric construction for virtual reality applications with virtual humans. *Studies in Health Technology and Informatics*, *191*, 33–37. PMID:23792838

Redish, J. (2007). Expanding usability testing to evaluate complex systems. *Journal of Usability Studies*, *2*(3), 102–111.

Riva, G. (2005). Virtual reality in psychotherapy. *Cyberpsychology & Behavior*, *8*(3), 220–230. doi:10.1089/cpb.2005.8.220 PMID:15971972

Rosenbaum, S., & Kantner, L. (2007). Field usability testing: method, not compromise. *Proceedings of the Professional Communication Conference*.

Rother, K., Karl, I., & Nestler, S. (2015). Virtual Reality Crisis Simulation for Usability Testing of Mobile Apps. In A. Weisbecker, M. Burmester, & A. Schmidt (Eds.), *Mensch und Computer 2015–Workshopband* (pp. 69–76). Stuttgart: De Gruyter Oldenbourg. doi:10.1515/9783110443905-010

Schryen, G., & Wex, F. (2015). Risk Reduction in Natural Disaster Management through Information Systems: A Literature Review and an IS Design. *Transportation Systems and Engineering: Concepts, Methodologies, Tools, and Applications*, 79.

Schubert, T. W. (2003). The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, *15*(2), 69–71. doi:10.1026//1617-6383.15.2.69

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. doi:10.1007/BF01068419 PMID:3450848

Schuirmann, D. L. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, *37*, 617–617.

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual Reality Training Improves Operating Room Performance. *Annals of Surgery*, *236*(4), 458–464. doi:10.1097/00000658-200210000-00008 PMID:12368674

Sutcliffe, A., & Gault, B. (2004). Heuristic evaluation of virtual reality applications. *Interacting with Computers*, *16*(4), 831–849. doi:10.1016/j.intcom.2004.05.001

Tang, C.-H., Wu, W.-T., & Lin, C.-Y. (2009). Using virtual reality to determine how emergency signs facilitate way-finding. *Applied Ergonomics*, *40*(4), 722–730. doi:10.1016/j.apergo.2008.06.009 PMID:18708182

Tarantino, L., & Spagnoletti, P. (2013). Can Design Science Research Bridge Computer Human Interaction and Information Systems? In Organizational Change and Information Systems (pp. 409–418). Springer. doi:10.1007/978-3-642-37228-5_40

Tromp, J. G., Steed, A., & Wilson, J. R. (2003). Systematic usability evaluation and design issues for collaborative virtual environments. *Presence (Cambridge, Mass.)*, *12*(3), 241–267. doi:10.1162/105474603765879512

Turk, M. (2015). Gesture recognition. In K. Hale & K. Stanney (Eds.), *Handbook of virtual environments: Design, implementation, and applications* (2nd ed., pp. 211–231). Boca Raton, FL: CRC Press.

Wang, P., Bishop, I. D., & Stock, C. (2009). Real-time data visualization in Collaborative Virtual Environments for emergency response. *Proceedings of the Spatial Sciences Institute Biennial International Conference*.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press. doi:10.1201/EBK1439808184

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence (Cambridge, Mass.)*, *7*(3), 225–240. doi:10.1162/105474698565686

Yao, R., Heath, T., Davies, A., Forsyth, T., Mitchell, N., & Hoberman, P. (2014). Oculus VR Best Practices Guide. *Oculus VR*. Retrieved from http://treyte.ch/oculus/tools/0.4.2/documentation/OculusBestPractices.pdf

ENDNOTES

1. Throughout the paper, we use the term real crisis simulation to make the distinction between these simulations and the virtual reality crisis simulations more clear.
2. See acknowledgements for more details.
3. The fact that the technology is integrated into the teaching process corresponds to the Humboldtian model of higher education of integrating research and studies.
4. Since the experiment was conducted in Germany, the German version of all mentioned questionnaires was used if available. Otherwise the questions were translated.
5. The original sound item has the levels way too quiet, too quiet, right, too loud, way too loud. Alternatively this could be mapped as a 1,3,5,3,1 scoring.
6. An additional written explanation was given as well and a warning text was shown within the VRCS at the beginning of the simulation.
7. The order was kept constant. The degree of dissimilarity was always set to the value of the dependent variable for the case "with testee" minus the value for the case "without testee".

Kristian Rother studied Business Information Systems at the University of Duisburg-Essen, Germany (Dipl.-Wirt.Inf.). Before entering academia he worked as a programmer, project manager and VP of marketing at a software company. He was a researcher at the University of Duisburg-Essen in the field of Artificial Intelligence. Currently, he is a researcher at the Hamm-Lippstadt University of Applied Sciences with a focus on Human-Computer Interaction, Virtual Reality, Augmented Reality and Crisis-Related Interactive Systems.

Inga Karl studied Applied Cognitive and Media Science at the University of Duisburg-Essen, Germany (M.Sc.). Currently she is a researcher at the Hamm-Lippstadt University of Applied Science in the project INTERKOM with a focus on Crisis-Related Interactive Systems. Further research focuses on Human-Computer-Interaction, Usability and Social Media.

Simon Nestler studied Computer Science (Dipl. Inf.) at the Technische Universität München, Germany (TUM) and received a PhD for his work on Human-Computer Interaction in life threatening, time critical and instable situations (Dr. rer. nat.) from the TUM. Currently, he holds a professorship at the Hamm-Lippstadt University of Applied Sciences and leads the Human-Computer Interaction research group. His research interests include all topics in Human-Computer Interaction, Social Media, Mobile Computing, Virtual Reality and Augmented Reality, especially with a focus on Crisis-Related Interactive Systems. Simon Nestler is a member of the German UPA and the German Informatics Society.