Experimental realism in user-tests related to medical systems

Faber, Jurjen

Award date:
2022

Experimental realism in user-tests related to medical systems

Jurjen Faber
Technical Univerisity of Eindhoven

Risk control measures are an important aspect in the development of new medical systems. However, current ways of testing have a hard time simulating a realistic environment during their user test. In this research we explore the basis of realism in user-tests of medical systems. First, interviews were conducted with team member of Philips, after which an experiment was constructed that could get better insights in the realism of user-tests with regards to medical systems. These experiments consisted of two methods of performing a user-test (task-wise vs. procedure wise). We found that perceived realism is higher in procedure wise user-tests. Furthermore, interviews after the experiment showed ways of improving the realism in user-tests. One aspect participants often mentioned is the amount of equipment and sound in the operation room as well as stress levels during the user-test. We conclude with suggestions of how the realism can be improved.

Keywords: realism, risk control measures, medical systems, user-tests

Introduction

Introduction to this topic

Medical systems are becoming more and more advanced as technology continues to develop new solutions. However, even with all the technological advances, it is important to keep potential risks in mind when developing medical systems, as well as make sure the systems are tested thoroughly. Through these tests, risk control measures that are implemented by a manufacturer can be tested to see if they prevent what they should be preventing and have the impact that they should have. Adverse events that one would like to avoid for example are a patient becoming stuck between moving parts of the machine or a doctor using the wrong settings resulting in a radiation dose that is too high for either patient or doctor. According to standard ISO14971 of the International Organization for Standardization (ISO), "the manufacturer shall determine risk control measures that are appropriate for reducing the risks to an acceptable level." (ISO, 2019) Furthermore, several options of achieving this have been listed in the standard. This standard is a good guideline with regards to risk management, as most regulatory bodies also use this document to determine their standards. Currently, user-tests face the problem that they can feel unrealistic for a participant. Currently, attempts are being made to make user-tests more realistic, but right now it is not clear what factors contribute to this unrealistic feeling and what can be done about it. It for example could be the amount of attributes in the room, noise levels, stress levels or even the used equipment during a user-test.

In order to release a medical system to the market, it has to be approved by the respective regulatory bodies. To be approved, it must be shown that the risks are at an acceptable level. Because regulatory bodies like the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) recognize the ISO 14971 standard, the document can be an important rule-set to adhere to. This can for example be shown through the results of tests done, or by handing over documentation about the fabrication and design process. However, the testing done by manufacturers can provide some difficulty, as the testing of risk control measures in itself is a paradoxical task. Normally, situations where risk control measures are triggered are situations one wants to avoid. Therefore, the testing of these situations might feel unrealistic for the participants of these tests. As a consequence of these unrealistic tests, results of these tests can be hard to generalize to real-world situations and therefore not showing the true impact of the risk control measures. As realism of the user tests is improved, results are more able to be extrapolated to real-world scenarios.

Within the tests we are talking about, several procedures and actions done by the user are tested. These procedures and actions are done by the doctor or technical assistant who is using the medical system. However, a problem found within the testing of these risk control measures, is the fact that testers were not sure how effective singular risk control measures that should be preventing some hazardous situations were. This means that sometimes it can be hard to determine the effectiveness of one risk control measures, as they are often stacked together to mitigate a risk. Part of the problem is that these singular risk control measures are tested in a task-wise approach, where the realism in the experiments are not that high. Therefore, during this research we will look into what defines realism of user-tests for medical systems. Furthermore, we will also look into how one could improve

the realism of a user-test.

Related work

Risk control measures

To adhere to the ISO (2019) standard, and therefore to the respective regulatory bodies as well, the manufacturer of a medical device shall implement, document and maintain an a constant monitoring of their risk management. Within this risk management, several processes exist. These processes are for example the identification of hazards and hazardous situations, estimating and evaluating associated risks, controlling these risks and monitoring the effectiveness of the risk control measures.

Through the use of risk management, new risk control measures can be implemented. Risk management is used to identify risks, analyze and prioritize risks, and handle and monitor all the risks (Boehm, 1991). Examples of risk management are reactive risk management and proactive risk management. Through reactive risk management, a risk can be seen as an anomaly that surfaces in form of a failure (Corciovă et al., 2013). After this failure is analyzed and the operational procedures are optimized, the problem will disappear and form no risk anymore. This happens within manufacturers as well, as they get feedback from the market when a problem occurs and then change their system accordingly. However, it is also necessary that risks are assessed and dealt with before anomalies can occur in the real world.

Within proactive risk management, a risk can be defined and estimated by using the answer to following three questions: (1) What can go wrong? (2) How likely is it to go wrong? (3) What are the consequences in case something goes wrong (Johansen & Rausand, 2014)? The process of answering these three questions can be illustrated by a bow-tie diagram and is called risk analysis.

In this bow-tie model, risk control measures are in place to prevent undesired events, and in case the undesired events happen, risk control measures are also in place to prevent or reduce the harm that follows the respective consequences. Furthermore, the expression of a risk can be done through the use of a risk metric, which can be defined as "a mathematical function of the probability of an event and the consequences of that event". (Jonkman et al., 2003)

The answers to the previously mentioned questions can be approached according to a top-down approach (Apostolakis, 2004), where first a set of undesirable end states is defined. Then for each end state, actions that lead to the end state are identified, after which the probabilities of these actions and scenarios are evaluated. This is done through the use of all available evidence, past experience and expert judgment. Lastly, the scenarios are ranked according to their expected frequency of occurrence.

Some of these scenarios have been present in previous iterations of medical devices as well, and therefore do not need to be tested to see if their respective risk control measures are working. However, when a new product is released or new features are added, new risk control measures are implemented as well. To verify and validate these new risk control measures, user-tests can be performed. Within these user-tests, specific scenarios are simulated in order to try to trigger the risk control measures. It is through this procedure that risk control measures are verified and validated. However, as mentioned previously, realism in these tests is something that can form a problem. This is something we will look deeper into in this research.

Realism in experiments

The realism of a test can be divided in two parts, experimental realism and mundane realism (Difonzo et al., 1998). Experimental realism can be defined as the degree of involvement and affectiveness of the experiment to the subject. Mundane realism is the likelihood that experiences encountered in the study will occur in the field as well. Morales et al. (2017) states that experimental realism can be placed on a continuum that ranges from very artificial to very realistic. Furthermore, they propose that when a setting involves more realism and when a setting is more naturalistic, the easier the generalization is of that experiment.

Another point made about realism in experiments is made by Herziger and Hoelzl (2017), where they stated that participants in hypothetical tasks have an intern bias. When making hypothetical choices or taking hypothetical actions, this leads to a situation where participants are likely to underestimate the cues that normally trigger a response. They therefore suggest that in an experiment you should give participants more real choice and rely less on their hypothetical situation thinking.

A way to make experiments more realistic in health services interventions has been proposed by Hayes et al. (2020). In their framework they identified four categories of factors that lets hypothetical decisions predict real-world behaviours. They advise that in order to test what factors are important in your specific setting, their four categories can be helpful. These categories include decision maker factors, cognitive factors, task factors and matching hypothetical and real-world tasks.

Decision maker factors are traits or capacities that relate directly to the decision maker. Cognitive factors are characteristics related to the decision-making process, with one important factor being "salience of or concern about the task". This factor implies that an increased salience of the decision or of the task can increase consistency between a hypothetical and a real-world situation. Task factors are factors that describe the aspects of the hypothetical decisions being made. Lastly, matching hypothetical and real-world tasks

include factors that increase the consistency of hypothetical and real-world situations by matching in ways not fitted in the other 3 categories. An example of this factor is that the study procedure should match the real-world decision context. The latter two factors will be examined in this research.

This research

Because the aforementioned problem about realism of testing the verification of effectiveness of risk control measures, we will look into the realism of user-tests in medical systems. This is brought forward in our research question "How can use-related risk control measures be tested in such a way that the experiment feels realistic?". We will answer this using sub questions: "How can realism be measured when testing a medical system?", "How does realism have effect on the results of the testing of a medical device?" and "How can the effectiveness of risk control measures be tested?"

With these questions we aim to get a better view of what constitutes realism in user-tests with regards to medical systems. Furthermore, we hope to find what realism in user-tests can say about the effectiveness of risk control measures and the relationship between these two aspects. During this research we will specifically focus on the Azurion line of Philips Medical Systems. Currently, this system adheres to the latest standards and regulations that apply. The Azurion line is an image guide therapy intervention tool and these systems are placed in operation rooms and interventional suites. Through the use of these systems, minimally invasive procedures can be performed using medical imaging guiding. Interviews will be done with employees of Philips in order to determine what aspects they think are important when testing risk control measures, as well as trying to figure out what they think entails realism in a user-test with regards to a medical system. With this information an experiment will be setup, where these aspects can be controlled and tested.

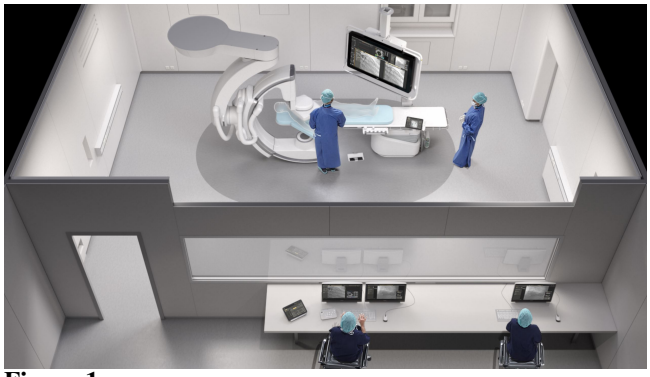


Figure 1

Azurion FlexArm test room. In the upper part the operation room with an Azurion FlexArm system and moveable table. In the lower part the control room.

Method

Pre expirement interviews

To help with designing the experiment, interviews were conducted to get more knowledge of the setup of the risk control measures, and the way of performing user-tests within Philips. These interviews were done with members of several teams within Philips Medical Systems. Eight participants with relevant experience on this topic were interviewed. Interviews were done in a semi-structured way, where several questions were prepared. For each team different questions were prepared, but the same basis and structure was used (see Appendix A). Follow up questions that were not prepared were asked when interesting answers were given, or when more insights were needed in a certain topic. These insights included information on risk control measures. For example how they are formed, how they are implemented and how they are tested. Furthermore, each interviewee was asked about their view of realism within the process. As this was more of an exploratory part, the answers given by the interviewees were used to form the experiments in the later part of this research. These answers were analyzed by summarizing the interviews and putting the themes on a digital post it board using Miro.com. These themes were then inspected to see which ones would be most interesting to explore further through our experiments. The selected themes were realism and the effect of procedure wise testing during user-tests. This information then was used to design the main experiment of this research.

Participants

Five participants between the age of 27 and 42 took part in the user-tests and following interviews. All participants were employees of Philips Medical Systems, with different backgrounds, expertise and team membership. The participants were contacted via e-mail, and chosen because of their experience of working with the Azurion system or previous experience in hospitals. The list of team-members that were contacted was given by one of the interviewees of the previous part. It was important that people had experience with Azurion because the tests would be done on an Azurion system. Furthermore, participants should have some prior experience in hospitals, as they would have to compare test situations with real-world situations. Participants were not compensated, as they were employees of Philips.

Materials

All tests were performed in a room located on the Philips campus in Best (see figure 1 for a similar room). These suites were built as testrooms for Philips, and resemble the setup of

a real-world hospital environment. This means that the operation room and control room were adjacent to each other and you can control the machine from within the control room as well as from within the operating room. In the operating room an Azurion FlexArm was installed, together with peripheral accessories often found in operating rooms equipped with the same system, such as a move-able table and a monitor that works together with the system. The rooms themselves are close resembles of real-world rooms, including radiation safety, meaning that rooms have to be secured and checked before radiation can be applied.

As part of the questionnaire an adaptation of the questionnaire used by Hill et al. (2012) was used. This questionnaire was developed to assess the realism of all key aspects of a colonoscopy simulator. Because the questionnaire from Hill et al. (2012) was about colonoscopy simulators we had to adapt the questionnaire. We used the same framing of the questions, i.e. "How realistic was ...", and then asked it about several characteristics of the experiment. Furthermore, to get insights in how likely it would be to see a similar setting in the real-world more questions were added. These questions were framed as "How likely is it to encounter a similar ...". The questionnaire consisted of 12 Likert scale questions with questions on a 5-point scale (see Appendix C). Answers ranged from "not realistic at all" to "very realistic" for questions about realism, and from "very unlikely" to "very likely" on questions about likeliness to a real-world scenario. Cronbach's alpha was high for the scale relating to realism (12 items; α= .92), while it was acceptable for the scale relating to likeliness (6 items; α= .77). Furthermore, a self-assessment manikin was used to measure the pleasure, arousal, and dominance of each participant after a test (Bradley & Lang, 1994).

Procedure

All participants were told that the experiment was to get insights about realism in user-tests. Therefore they would be performing two types of user-tests. Then they were randomly assigned an order of the tests, either the task-wise task first, or the procedure wise task first.

In the task-wise task, participants were asked to perform certain steps during the task (see Appendix B). These tasks included operations that should be done by the participants. Operations were chosen in such a way that certain risk control measures would be triggered. The risk control measures that would be triggered were in collaboration with an employee of Philips, with an eye on that it should be able to check them in both the task-wise test as well as the procedure test. First it was explained what the current situation was and the final goal of the task. The experimenter instructed each participant which task to perform. These tasks were all in the same order for every participant.

For the procedure wise task, a short prompt with information was given, after which it was expected that the participant would complete this procedure. The prompt was chosen such that it was expected that all the participants chose roughly the same approach to the procedure.

Analysis

For analysis of the quantitative data, it was chosen to perform exploratory research as the final sample size was smaller than expected. This was done through a t-test. In this t-test, the difference between the two user-tests will be analyzed. This was done for both the realism part of the questionnaire as well as the likeliness part of the questionnaire. All the respective individual questions, except the ones on lighting, were used to make an average scoring of each concept. Questions about lighting were omitted because it was not able to manipulate them as we wanted. Furthermore, a t-test will also be performed on all the 3 attributes of the self-assessment manikin. The groups again will be task-wise and procedure wise. The qualitative data of the interviews was first transcribed, after which an analysis of occurring themes was performed. The forming and grouping of themes was done by using a Miro.com board with an electronic post-it setup.

Results

Questionnaire

All five participants completed the whole experiment. The realism of the task-wise user-test (M = 3.7, SD = .32) was lower than the realism of the procedure user-test (M = 4.33, SD = .36). This was tested through a t-test and results were statistically significant, $t(4)$ = -3.41, $p < 0.03$.

The likeliness of encountering a scenario like the experiment in the real world was not different between the procedures, $t(4)$ = 0.0, p = 1. With the task-wise (M = 4.13, SD = .23) and procedure test (M = 4.13, SD = .2) rated the same.

No significant difference was found between the groups with regards to the attributes of the self-assessment manikin (mood: $t(4)$ = 1.63, p = 0.18. arousal: $t(4)$ = 0.0, p = 1. valence: $t(4)$ = .59, p = .59).

Interviews

Room characteristics

Participants pay close attention to what their environment looks like. They see this as a key characteristic of how realistic they experience the user-test. When the experiment is setup in a room that is closely resembling the real-world rooms, they experience higher realism during their experiment. This already starts with the layout of the room and the general feel:

[Participant 2]: Sometimes usability tests might be in a office

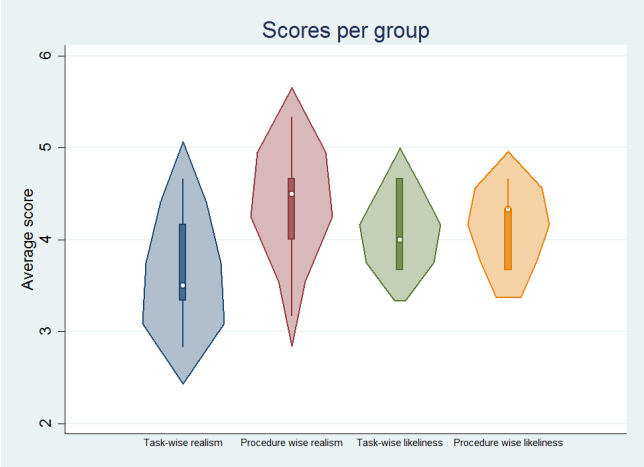


Figure 2

Violin plot of the data. White dot in the middle is the median. Realism is the average of all the realism questions related to the specific task method. Likeliness is the average of all the likeness questions related to the specific task method.

room where you put the medical equipment, but this very much has the feel of an actual hospital.

Furthermore, a big factor in how realistic the setting in the room was, depends on the sound cues in the room. There is a wide variety of sound cues that can improve realism, such as alarms of medical equipment, chatting team members and general operating noises from medical equipment:

[Participant 1]: You could also have an alarm going from the hemodynamics, yelling doctor etcetera etcetera, to include other stimuli that would normally be in the operating room or the control room.

[Participant 2]: It is very quiet right now, with no other people equipment and noises coming from other equipment I guess. And from the other people, noises coming from the other people.

According to the participants it was not only about sounds in the room, but also other things in general. Even though the rooms looked like real operating rooms, participants stated that the rooms were very empty and 'clean':

[Participant 2]: So especially us testing the system, part of me thinks maybe the realism, a bunch of other equipment could help, in terms of thinking where can you move the c-arm without making collisions

[Participant 5]: I think one thing that would be definitely different from a real case scenario is the amount of stuff around, and on the table and around the actual table and the system. Obviously here we are lacking all kinds of drapes, catheters, other machines and whatnot.

So we can see that realism during a user test can be improved by improving the conditions of the room the participants are working in. Right now, the rooms that are used are empty and void of sounds. According to the participants, additions to these elements during an experiment could improve the perceived mundane realism of the experiment itself. The first step is using a real operating and control room, instead of just an office room where you pull in the equipment. Furthermore, sounds add to a realistic experience as well. Lastly, an operating room that is filled with more equipment than just the system you are performing a user-test on also adds realism to the test. This is because in normal operating rooms there is also more equipment and stuff related to operations present in the room.

Secondly, all participants stated that having a phantom on the table on which you can perform scans drastically improves realism:

[Participant 5]: I also think it helps a lot to actually have the patient or the mannequin on the table, because it is very hard to visualize that without it, or you know be imaging something empty and be getting picture of grain.

This was also commented on by another participant:

[Interviewer]: What do you think makes a user-test realistic?

[Participant 1]: Most of all the dummy, and of course be able to control the system like we normally do, that's what made it realistic

Therefore, a dummy or phantom is really important in performing a user-test. Because using a real participant is not possible through ethical concerns, a dummy comes closest to what you can expect in the real-world. Moreover, participants likes being able to interact with a 'patient' and seeing a result of their scans. As some participants also mentioned, this looks to be coming from the fact that the participant does not have to imagine things, and can be more focused on performing the user-test instead.

Task procedure

Another theme that emerged with regards to realism is the method of the tasks. As explained, two types of tasks were performed during this research, the task-wise test and the procedure wise test. During the task-wise test, participants followed precise tasks instructed by the experimenter. During the procedure wise task participants were instructed to perform a particular procedure that was suited for the proposed problem. Participants stated that the latter task was

more realistic, but the uses of the tests were different. Participants were also in agreement that both types of testing methods were useful as each type has its own benefits.

Regulations require the check of the presence and the working of a risk control measure. A combination of both task wise and procedure wise approach are found to support fulfilling this requirement.

So with regards to the method of the task, there is no real consensus. Most participants state that it is harder to validate risk control measures that are not part of the normal way of working during a procedure.

Stress

Lastly, a theme that emerged was the fact of stress levels during a test. Right now, participants stated that they were very relaxed during the test because there was no threat of doing something wrong. This made it feel less realistic for them.

[Interviewer]: And do you think stress levels have an effect, like stress levels of the user has en effect on the test?

[Participant 5]: I think it can, I think it can definitely you know, relating back to me there was actually a patient lying there, I would have been way more stressed with what I am pressing ... I think that would be adding to the stress of the user in the sense that yeah I could actually potentially cause harm and that a simulated environment makes you much calmer.

So a realistic situation has a user more stressed because there are more risks to an operation.

Another thing that is mentioned is that because of low levels of stress, doctors know that the test they are doing is not real, diminishing the realism of the experiment.

[Participant 3]: And the question of course is how realistic you can really make it, because there is not really someone on the table who gets a cardiac arrest. That is, they know that it is not real.

This also highlights something that was mentioned by all participants, that it is impossible to test certain scenarios, like for example the patient getting a cardiac arrest.

Discussion

We have found that realism can be measured with our adapted questionnaires. Furthermore, there are clear ways to improve realism further within the user-tests of medical systems. Participants stated that the characteristics in an operation room can improve the realism. We also gained insights in what kind of test method is best to use in what situation. Another point made about current user-tests is that the stress levels are not as high as in real-world situations.

In this research we wanted to look into how realism can be measured when testing a medical system. According to our data, the questionnaire that was used gives a good result of how realistic participants perceive a user test. The high internal validity of the questionnaire indicated that in order to get insights in the realism of a user-test of medical systems, it is important to ask the participants about the realism of the procedure, tasks, control room, materials and the images. Of course this questionnaire can be expanded with more questions, one example is how realistic participants find the overall 'cleanliness' of the operating room. That is, how realistic do the participants think all the other equipment is in the operating room. As our participants said during the interview, they thought the operating room was very clean, something that they normally do not encounter in real-world situations.

Although the questions with regards to realism had a good Cronbach's alpha, the questions with regards to likeliness did not have such a good internal consistency, but instead acceptable levels (above .75). Even though these scores are decent, it is important to see how we could improve our questionnaire.

As mentioned previously, several items could be added to get feedback on the realism of more elements in the user-test. However, a high Cronbach's alpha might show redundancy in the questions (Tavakol & Dennick, 2011). Tavakol and Dennick (2011) even states that a maximum of .90 should be used. That could mean that instead of adding more elements of the user-test, the questionnaire should be shortened. However, with the Cronbach's alpha values in our research, we have shown that our questionnaire can be used to test the realism of a user-test with regards to medical systems.

Next to the measuring of realism, we also looked in how this realism has an effect on the results of a user-test. Participants stated that it is important to come as close to the real world as possible when you want to prove certain risk control measures are present and working. However, participants also stated that it should still be possible to validate the risk control measures if you do user-tests in a procedure wise approach. Therefore, it seems that currently it is still better to use a setup in your user-test where you use both methods. This results in the ability to still check certain risk control measures and if they are working as intended in your system.

Furthermore, when experiments are more realistic, Morales et al. (2017) states that it is easier to generalize the experiment. This means that the results of a more realistic user-test are easier to generalize. If we then take a look at our examples, this means that results of procedure wise user-tests are better to generalize in comparison to task-wise user-tests. Our results are also in line with that statements of Hayes et al. (2020), where they state that in order to make experiments more realistic, one should match the study procedure with the real-world decision making context.

Something else that is stated in previous research is the

fact that having an experimental design utilizing realistic independent variables ensures that researchers are manipulating what they intend to manipulate (Lieberman et al., 2019). This then also enhances the external validity of your research. Our participants have also stated this, as they said that it was important to come as close to real world situations as possible.

During our research we also looked into how the testing of the verification of effectiveness of risk control measures can be improved. Our participants stated that having a procedure wise task scenario might result in lots of different actions by the participant. With this they mean that one prompt as procedure might result in participants performing the procedure on their own way. Therefore it might be hard to validate the risk control measures you want to test in a realistic setting, as you do not know beforehand exactly which risk control measures the participants will trigger. This results then in having to incorporate the risk control measures in a task-wise user test, in which you have less realism, but more control over what specific measures you want to test. Because you have a less realistic setting, it is harder to assess the effectiveness of a risk control measure. This could be explained by the fact that participants are not in the right mindset, and would not react to a risk control measure in the same way as they would in a real world scenario. Overall, participants stated that test scenarios should be created in which realism is kept high, but it is still possible to verify your risk control measures.

Another aspect that was brought up by some participants is that right now risk control measures could be validated at the end of a user-test. When they do this, they first run through a task-wise script, and then at the end validate risk control measures that could not be fitted in the task-wise script. This is mostly done for measures that can not be incorporated in the clinical relevance of the task-wise user-test. However, participants would prefer to test as much as possible within the clinical workflow.

Room characteristics

Something that became pretty clear from the interviews, is that participants felt that there were several characteristics during a user-test that are important to improve realism. First of all, an often mentioned part of the user-test was the room characteristics of the operation and control room. In order to feel like a realistic scenario, a room should also look and sound like a real operation or control room. The characteristic that was most named was sound from other sources.

Sound is known to bring about a negative mood, increased stress and difficulty in concentrating (Frankenhaeuser & Lundberg, 1977). This could mean that participants feel that currently they are not being influenced by sounds that would normally have an impact on them in real-world situations. Furthermore, previous research has shown that for example an increase in loudness results in greater arousal (Loewen & Suedfeld, 1992). This then in turn improves simple task performance, but could hinder more complex task performance.

In our experiment, participants thought the room was lacking beeps from other medical equipment or chatting from their coworkers. This is something that currently is not simulated in user-tests yet, but could be easy additions to a test setup, while adding to realism. Chatting team members could for example be reached by inviting a team instead of just one doctor or technician, something that is being done more and more already. Other sounds like the sounds from medical equipment could be simulated using speakers in the room which are playing recorded sounds from a real world operating room. Even better could be having the real equipment in the room, as participants also stated that the current testing environment was very 'clean'.

These statements from participants could also be seen as a combination with the previously mentioned rise of arousal through sound. Instead of just missing the sounds themselves, maybe participants thought they were missing the consequences of these sounds. Participants also stated that stress levels were different during the user-test and during real-world situations.

Stress

Stress also seemed to be an important characteristic of realism in a user-test. Participants mentioned that some situations could not be simulated during user-tests. One often mentioned situation was the event of a cardiac arrest. In this situation, doctors experience higher levels of stress (Hunziker et al., 2013). Hunziker et al. (2013) also state that stress levels during a simulation of cardiac arrest are probably lower than cardiac arrest situations in a real-world scenario. In another research, it was found that doctors perceived stress levels during a cardiac arrest situation as a score of 9 on a scale from 0 to 10 (0 = "no stress at all felt", 5 = "some stress felt", 10 = "very high stress felt")(Hunziker et al., 2009).

As a participant mentioned, they feel much calmer during a simulated environment where stress levels are not as high, thus suggesting that higher stress levels improve realism in user-tests. Furthermore, it is shown that stress can alter the underlying mechanisms of decision making, like 'adjustment from automated response', 'feedback processing' and 'strategy use' (Starcke & Brand, 2012). This then shows that decision making is different in a more realistic setting in which stress is more present.

Because some situations would be very hard, or even impossible, to simulate, instead what one can do is raise stress levels through other means to still simulate a realistic environment. However, one thing that was unclear is whether participants were really thinking about stress, or just feeling stressed by for example higher arousal levels. Both of these situations can be indicated by increased heart rate (McEwen,

1998). Currently, it is very easy to measure heart rate, for example through the use smart watches. These can be used in future research in order to see what levels of stress are measured during a user-test. To simulate scenarios where participants still feel like they could be in a stressful situations, several actions can be taken to increase heart rate, arousal or stress levels.

Several methods are known to increase arousal levels within a participant. There are a multitude of mental and physical tasks to use, for example having a participant perform in a quiz for 3 minutes or counting backwards from a certain number in steps of 7 (Faulstich et al., 1986). Longer methods to induce stress can also be used, such as the Trier Social Stress Test (20 minutes) (Kirschbaum et al., 1993), or the Sing-a-Song Stress Test (Brouwer & Hogervorst, 2014). However, it is important to notice that stress and arousal levels varies between each individual (Arena et al., 1989). This is caused by two factors during an experiment. The first is the average level of anxiety (Arena et al., 1989), while the second factor is the intensity of the stressor (Neiss, 1988). Therefore it is important to keep in mind individual trait anxiety, when designing your experiment.

Limitations

Although results from the interviews were clear and contained good insights, there are some limitations to this study. First of all, the results of the questionnaires should be considered carefully, as the sample size only was five. Because this low sample size, it should be good for future research to see if the differences are still present in greater groups.

Next to the differences between the two test methods, the alpha value could also be more confident if the sample size would be bigger. With a bigger sample size, comparisons of Cronbach's alphas becomes more reliable (Bujang et al., 2018).

Moreover, part of our sample size consisted of team members of the usability validation team. Members of this team are very proficient in using user-tests to acquire data and running experiments with doctors. This means they have good knowledge of the user-tests themselves as well as have a good view of what doctors think. However, it should be noted that results gathered from doctors themselves would always be better, as they have more direct experience with real-world scenarios.

Future research

In future research, one big step that can be taken is using participants that are currently employed in hospitals. By doing so, one will get better insights with regards to the realism of your user-test.

Besides that sample of the experiments, future research should also look into refining the questionnaire used in this research. As shown, the Cronbach's alpha values were sufficient, but an optimization of which items are most important when it comes to realism could be very helpful. Furthermore, the questionnaire on likeliness could be improved by adding more items. A study where these questionnaires will be further validated could be good.

Lastly, using stress in the setup of an experiment can be a good approach. Currently there was no manipulation, nor was there any measurement of stress levels during the experiment. As it was mentioned by almost all participants, stress seems to have an influence on the realism of the user-tests. To start it could be good to measure stress levels in real-life, such that when you start manipulating it in an experiment, you know what the levels should be have high realism. These measurements could be done in a non-invasive way by using smartwatches to measure heart rate.

Conclusion

Our research has shown that there is still a lot to be explored with regards to realism in user-tests with regards to medical systems. Our participants stated that realism in user-tests can still be improved by a lot of factors, for example by removing the cleanliness of the operating room and incorporating stress in your user-tests. When trying to improve the realism in your user-test, you can start by adding sound cues found in hospitals, as well as applying a procedure wise test instead of task-wise tests.

References

Apostolakis, G. E. (2004). How useful is quantitative risk assessment? *Risk Analysis: An International Journal*, *24*(3), 515–520.

Arena, J. G., Goldberg, S. J., Saul, D. L., & Hobbs, S. H. (1989). Temporal stability of psychophysiological response profiles: Analysis of individual response stereotypy and stimulus response specificity. *Behavior Therapy*, *20*(4), 609–618.

Boehm, B. W. (1991). Software risk management: Principles and practices. *IEEE software*, *8*(1), 32–41.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49–59.

Brouwer, A.-M., & Hogervorst, M. A. (2014). A new paradigm to induce mental stress: The sing-a-song stress test (ssst). *Frontiers in neuroscience*, *8*, 224.

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for cronbach's alpha test: A simple guide for researchers. *The Malaysian journal of medical sciences: MJMS*, *25*(6), 85.

Corciovă, C., Andritoi, D., & Ciorap, R. (2013). Elements of risk assessment in medical equipment. *2013 8th International Symposium On Advanced Topics In Electrical Engineering (ATEE)*, 1–4.

Difonzo, N., Hantula, D. A., & Bordia, P. (1998). Microworlds for experimental research: Having your (control and collection) cake, and realism too. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 278–286.

Faulstich, M. E., Williamson, D. A., McKenzie, S. J., Duchmann, E. G., Hutchinson, K. M., & Blouin, D. C. (1986). Temporal stability of psychophysiological responding: A comparative analysis of mental and physical stressors. *International Journal of Neuroscience*, *30*(1-2), 65–72.

Frankenhaeuser, M., & Lundberg, U. (1977). The influence of cognitive set on performance and arousal under different noise loads. *Motivation and Emotion*, *1*(2), 139–149.

Hayes, T., Hudek, N., Graham, I. D., Coyle, D., & Brehaut, J. C. (2020). When piloting health services interventions, what predicts real world behaviours? a systematic concept mapping review. *BMC Medical Research Methodology*, *20*(1), 1–20.

Herziger, A., & Hoelzl, E. (2017). Underestimated habits: Hypothetical choice design in consumer research. *Journal of the Association for Consumer Research*, *2*(3), 359–370.

Hill, A., Horswill, M. S., Plooy, A. M., Watson, M. O., Karamatic, R., Basit, T. A., Wallis, G. M., Riek, S., Burgess-Limerick, R., & Hewett, D. G. (2012). Assessing the realism of colonoscopy simulation: The development of an instrument and systematic comparison of 4 simulators. *Gastrointestinal endoscopy*, *75*(3), 631–640.

Hunziker, S., Tschan, F., Semmer, N., & Marsch, S. (2013). Importance of leadership in cardiac arrest situations: From simulation to real life and back. *Swiss medical weekly*, *143*(1516).

Hunziker, S., Tschan, F., Semmer, N. K., Zobrist, R., Spychiger, M., Breuer, M., Hunziker, P. R., & Marsch, S. C. (2009). Hands-on time during cardiopulmonary resuscitation is affected by the process of teambuilding: A prospective randomised simulator-based trial. *BMC emergency medicine*, *9*(1), 1–10.

ISO. (2019). *Medical devices — application of risk management to medical devices* (ISO 14971:2019(E)). International Organization for Standardization. https://www.iso.org/standard/72704.html

Johansen, I. L., & Rausand, M. (2014). Foundations and choice of risk metrics. *Safety science*, *62*, 386–399.

Jonkman, S., Van Gelder, P., & Vrijling, J. (2003). An overview of quantitative risk measures for loss of life and economic damage. *Journal of hazardous materials*, *99*(1), 1–30.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76–81.

Lieberman, A., Morales, A. C., & Amir, O. (2019). Beyond the lab: Using data from the field to increase research validity. *Handbook of research methods in consumer psychology* (pp. 41–60). Routledge.

Loewen, L. J., & Suedfeld, P. (1992). Cognitive and arousal effects of masking office noise. *Environment and Behavior*, *24*(3), 381–395.

McEwen, B. S. (1998). Protective and damaging effects of stress mediators. *New England journal of medicine*, *338*(3), 171–179.

Morales, A. C., Amir, O., & Lee, L. (2017). Keeping it real in experimental research—understanding when, where, and how to enhance realism and measure consumer behavior. *Journal of Consumer Research*, *44*(2), 465–476.

Neiss, R. (1988). Reconceptualizing arousal: Psychobiological states in motor performance. *Psychological bulletin*, *103*(3), 345.

Starcke, K., & Brand, M. (2012). Decision making under stress: A selective review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1228–1248.

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *International journal of medical education*, *2*, 53.

Appendix A
Pre-experiment interview questions

Usability Tester:

[Uitleg over waarom ik bezig ben met dit onderzoek en wat het doel van het interview is: In het onderzoek wil ik het realisme van user tests verbeteren en de effectiveness van Risk control measures beter in kaart brengen. Focus van de risk control measures is user-related. Als onderdeel hiervan zullen we interviews houden om meer informatie te krijgen over de huidige manier van werken, vooral op betrekking tot risk control measures. We willen gaan kijken naar categorisatie van risk control measures zodat we kunnen kijken naar de effectiveness van de measures.]

Wie ben je en kan je jezelf introduceren, bijvoorbeeld over je rol binnen Philips?

Wat zijn jouw taken met betrekking tot Azurion?

Als usability tester, hoe komen risk control measures terug in jouw rol?

Kan je omschrijven hoe een test voor een nieuwe feature er bij jou uit ziet en hoe zich dit verhoudt met risk control measures?

[Als stap om RCM wat duidelijker te krijgen heb ik categorien gemaakt waar measures staan die in mijn ogen bij elkaar horen. In de categorisatie heb ik gekeken naar hoe een risk-control measure getriggerd wordt.]

Herken je dat verschillende measures op deze manier bij elkaar te groeperen zijn?

[Een categorie is een system failure, die getriggerd wordt als er iets fout gaat in het systeem. Een voorbeeld hiervan is: “Imaging unit shall provide status information to the user in case there is a movement limitation due to failures.”]

Hoe testen jullie system failures bij een eindgebruiker?

Hoe wordt de effectiviteit van zo’n maatregel getest?

[Een andere overkoepelende categorie die ik heb geclassificeerd is user actions, hier vallen measures onder die worden getriggerd als de gebruiker een bepaalde actie onderneemt. Een voorbeeld hiervan is “The user shall be warned about the possibility of misaligned overlay images.”]

Hoe testen jullie dit soort measures bij de eindgebruiker? Wordt dit scenario door jullie er in gezet of forceren jullie de gebruiker tot dit soort triggers?

[Een volgende categorie die ik heb gevonden is “system information”, wat een categorie is waar risk control measures instaan die statisch informatie weergeven. Een voorbeeld hiervan is: “The system shall display and report the skin dose information and warnings.”]

Hoe testen jullie of zo’n maatregel met statische informatie het gewenste effect heeft?

[Soms kan het best onrealistisch aanvoelen voor een eindgebruiker tijdens een test.]

Hoe proberen jullie het realisme van een test zo hoog mogelijk te houden?

Wat is de algemene ervaring van een eindgebruiker tijdens zo’n test?

Een veel voorkomende maatregel, en dus ook categorie, is dat er iets in de Instructions for Use moet staan, hoe wordt dit door jullie geverifeerd? En wordt de IfU ook getest op de eindgebruiker?

Hoe wordt over het algemeen de effectiviteit van maatregelen getest?

[Verder spreek ik nog met 3 andere teams, safety designers, clinical marketing en verification testing.]

Hoe staat jouw team in verband met deze andere teams?

Heb je nog verdere aanvullingen op de categorien?

Hoe zou je het proces om RCMs te formuleren en te testen verbeteren?

Safety Designer:

[Uitleg over waarom ik bezig ben met dit onderzoek en wat het doel van het interview is: In het onderzoek wil ik het realisme van user tests verbeteren en de effectiveness van Risk control measures beter in kaart brengen. Focus van de risk control measures is user-related. Als onderdeel hiervan zullen we interviews houden om meer informatie te krijgen over de huidige manier van werken, vooral op betrekking tot risk control measures. We willen gaan kijken naar categorisatie van risk control measures zodat we kunnen kijken naar de effectiveness van de measures.]

Wat zijn jouw taken met betrekking tot Azurion?

Ben je bekend met risk control measures?

Als Safety Designer, hoe ben je betrokken bij risk control measures?

Kan je omschrijven wat het proces is voor jou wat je doorloopt bij de ontwikkeling van een nieuwe feature?

Hoe komt een risk control measure tot stand?

[Risk control measures zijn te ranken op severity of hoe erg ze nodig zijn.]

Hoe wordt deze assessment gedaan?

Hoe weten/testen jullie of een risk control measure effectief is?

Hoe belangrijk is de effectiviteit van een measure voor jullie?

[Als stap om RCM wat duidelijker te krijgen heb ik categorien gemaakt waar measures staan die in mijn ogen bij elkaar horen. Dit zijn onder andere dynamische en statische measures]

Herken je dat verschillende measures bij elkaar te groeperen zijn?

Hoe bepaal je of een measure bij een bepaalde actie triggert (dynamisch) of altijd actief moet zijn (statisch)?

[Twee verschillende bewoordingen voor measures die ik vaak terug zag komen waren warnings en notifications.]

Zou je kunnen uitleggen wat het verschil zou kunnen zijn tussen een warning en een notification?

Hoe wordt bepaald waar een notification of warning weergegeven wordt?

[Een andere categorie die vaak terugkomt is het gebruik van de IfU.]

Hoe wordt bepaald welke maatregelen en warnings er in de IfU moeten komen?

[Verder zijn er in het proces nog 3 teams betrokken, usability testing, clinical marketing en verification testing.]

Hoe staat jouw team in verband met deze andere teams?

Hoe zou je het proces om RCMs te formuleren en te testen verbeteren?

Clinical Marketing Specialist:

[Uitleg over waarom ik bezig ben met dit onderzoek en wat het doel van het interview is: In het onderzoek wil ik het realisme van user tests verbeteren en de effectiveness van Risk control measures beter in kaart brengen. Focus van de risk control measures is user-related. Als onderdeel hiervan zullen we interviews houden om meer informatie te krijgen over de huidige manier van werken, vooral op betrekking tot risk control measures. We willen gaan kijken naar categorisatie van risk control measures zodat we kunnen kijken naar de effectiveness van de measures.]

Wat zijn jouw taken met betrekking tot Azurion?

Ben je bekend met risk control measures?

Als clinical marketing specialist, hoe ben je betrokken bij risk control measures?

Kan je omschrijven hoe een test voor een nieuwe feature er bij jou uit ziet?

[Als stap om RCM wat duidelijker te krijgen heb ik categorien gemaakt waar measures staan die in mijn ogen bij elkaar horen.]

Herken je dat verschillende measures bij elkaar te groeperen zijn?

Merk je als gebruiker een duidelijk verschil tussen een dynamische measure en een statische measure?

[Een veel voorkomende maatregel, en dus ook categorie, is dat er iets in de Instructions for Use moet staan.]

Hoe ervaar je dit document als potentiele eindgebruiker?

[Twee verschillende manieren van tonen die ik vaak terug zag komen waren warnings en notifications.]

Merk je als eindgebruiker dat er een verschil zit tussen een warning en een notification?

[Soms kan het best onrealistisch aanvoelen voor een eindgebruiker tijdens een test.]

Herken je dit als jij een systeem moet testen?

Wat zijn voor jou de grootste verschillen met een echte situatie en een test situatie?

Heb je ideen hoe deze situaties dichter bij elkaar kunnen komen?

[Verder zijn er in het proces nog 3 teams betrokken, safety designers, usability testing en verification testing.]

Hoe staat jouw team in verband met deze andere teams?

Heb je nog verdere aanvullingen op de categorien?

Hoe zou je het proces om RCMs te formuleren en te testen verbeteren?

Verification testers:

[Uitleg over waarom ik bezig ben met dit onderzoek en wat het doel van het interview is: In het onderzoek wil ik het realisme van user tests verbeteren en de effectiveness van Risk control measures beter in kaart brengen. Focus van de risk control measures is user-related. Als onderdeel hiervan zullen we interviews houden om meer informatie te krijgen over de huidige manier van werken, vooral op betrekking tot risk control measures. We willen gaan kijken naar categorisatie van risk control measures zodat we kunnen kijken naar de effectiveness van de measures.]

Wat zijn jouw taken met betrekking tot Azurion?

Ben je bekend met risk control measures?

Kan je omschrijven hoe een test voor een nieuwe feature er bij jou uit ziet?

[Als stap om RCM wat duidelijker te krijgen heb ik categorien gemaakt waar measures staan die in mijn ogen bij elkaar horen.]

Herken je dat verschillende measures bij elkaar te groeperen zijn?

[Verder zijn er in het proces nog 3 teams betrokken, safety designers, clinical marketing en usability testing.]

Hoe staat jouw team in verband met deze andere teams?

Heb je nog verdere aanvullingen op de categorien?

Hoe zou je het proces om RCMs te formuleren en te testen verbeteren?

Appendix B
Task-wise user test script

Label	RCM	Prompt to participant / scribe action	User actions to reach the success criteria	Success criteria	Score	Observational Data: use errors, user's comments, answers, actions
Checklist		- Position of patient is wrongly entered	-			
Introduction						
Introduction			-			
Azurion POF						
		Patient A is on the table, you would like to place a stent in the carotid artery.				
	56	Can you check the orientation of the patient?	Participant changes orientation of the patient	Patient orientation is set as * in system	☐1 ☐2 ☐3 ☐4 ☐N/A	
Task	27	Can you visualize the groin area by using an abdomen EPX protocol.	Participant selects the correct mapping	EPX mapping chosen	☐1 ☐2 ☐3 ☐4 ☐N/A	-
Task	80, 83	You have punctured the groin. Imagine you are moving the catheter up to the carotid under live fluoro. Let's simulate this from the control room.	Participant creates fluoro.	Fluoro is made.	☐1 ☐2 ☐3 ☐4 ☐N/A	-
Task		Can you visualize the carotid arteries (neck area).	Participant moves arm to neck area	Arm is located in neck area	☐1 ☐2 ☐3 ☐4 ☐N/A	
		To navigate the catheter you want to visualize the carotid arteries from the lateral view. Please do so now.	Participant moves the arms in lateral position.	Arm is situated in lateral position	☐1 ☐2 ☐3 ☐4 ☐N/A	
Task	96	Can you move the detector as close to the patient as possible?	Participant moves the detector closer to the table.	Trigger collision warning	☐1 ☐2 ☐3 ☐4 ☐N/A	
Task	39	Can you perform a simple measurement on the ROI with the TSM?	Participant makes measurement of patient.	Succesfull triggering of the warning.	☐1 ☐2 ☐3 ☐4 ☐N/A	
Task	65	You would like to know the dose information.	Participant reads actual dose information and explains their judgement.	Correct value is read out loud and clear judgement is explained	☐1 ☐2 ☐3 ☐4 ☐N/A	-
Facilitator		Follow up with 'Post-interview' questions.				

Appendix C
Questionnaire experiment

Participant Number

What is your gender?

- Male
- Female
- Prefer not to answer

What is your age?

Do you have any experience working in the hospital?

- Ye
- No

If yes, how many years of experience do you have?

How would you rate your knowledge of the Azurion system?

Not good at all		Undecided		Very good
○	○	○	○	○

Participant Number

	Not realistic at all		Undecided		Very realistic
How realistic was the procedure?	○	○	○	○	○
How realistic were the tasks you had to perform?	○	○	○	○	○
How realistic was the control room?	○	○	○	○	○
How realistic was the lighting in the operating room?	○	○	○	○	○
How realistic was the operating room overall?	○	○	○	○	○
How realistic were the materials on the table?	○	○	○	○	○
How realistic were the images?	○	○	○	○	○

Participant Number

Very unlikely Undecided Very likely

How likely is it to encounter a similar procedure in the hospital?

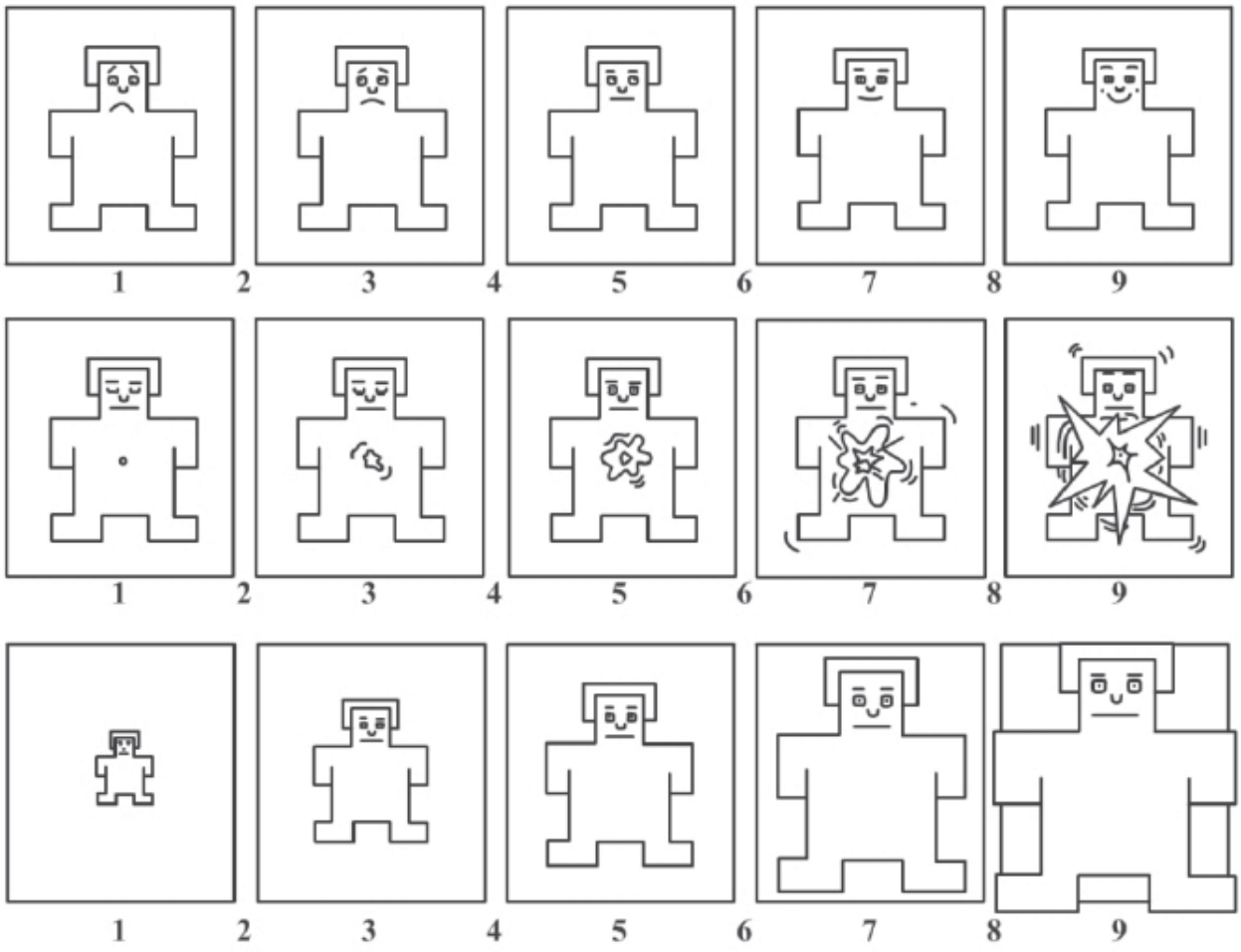
How likely is it to encounter a similar control room in the hospital?

How likely is it to encounter similar lighting in the operating room in the hospital?

How likely is it to encounter a similar operating room in the hospital?

Not realistic at all Undecided Very realistic

How realistic was this test overall?


1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9

Appendix D
Experiment interview questions

Interview

What do you think, makes a user-test realistic?

How would you compare the test-scenario with a real-world scenario in a hospital?

What is missing in the test-scenario compared to a real-world scenario in a hospital?

What aspects about a test are important if you want to improve realism?

Which test-scenario do you prefer and why?

What do you think are the effects of added realism to user-tests?

Can you think of events or errors that could happen in real life, that are impossible to test in the current setting?

Would you like to have a higher stress-level (for instance due to time-pressure) to make the procedure more realistic for risk-assessment?

Appendix E
Stata do-file

```
use "the data", clear

gen tw_light_real_avg = (tw_1+tw_2+tw_3+tw_4+tw_5+tw_6+tw_7)/7

gen tw_light_like_avg = (tw_8+tw_9+tw_10+tw_11)/4

gen tw_real_avg = (tw_1+tw_2+tw_3+tw_5+tw_6+tw_7)/6

gen tw_like_avg = (tw_8+tw_9+tw_11)/3

gen pr_light_real_avg = (pr_1+pr_2+pr_3+pr_4+pr_5+pr_6+pr_7)/7

gen pr_light_like_avg = (pr_8+pr_9+pr_10+pr_11)/4

gen pr_real_avg = (pr_1+pr_2+pr_3+pr_4+pr_5+pr_6+pr_7)/6

gen pr_like_avg = (pr_8+pr_9+pr_11)/3

label variable tw_real_avg "Task-wise realism"

label variable pr_real_avg "Procedure wise realism"

label variable tw_like_avg "Task-wise likeliness"

label variable pr_like_avg "Procedure wise likeliness"

vioplot tw_real pr_real tw_like pr_like, title("Scores per group") ytitle(Average score) xlabel(,
labsize(vsmall))

ttest tw_like_avg == pr_like

ttest tw_real_avg == pr_real_avg

ttest tw_mood = pr_mood

ttest tw_arousal = pr_arousal

ttest tw_valence = pr_valence

alpha tw_8 tw_9 tw_11 pr_8 pr_9 pr_11

alpha tw_1 tw_2 tw_3 tw_5 tw_6 tw_7 pr_1 pr_2 pr_3 pr_5 pr_6 pr_7
```