# GAME-BASED EVALUATION OF PERSONALIZED SUPPORT FOR ASTRONAUTS IN LONG DURATION MISSIONS

N.J.J.M. Smets
TNO /DUT
Soesterberg, The Netherlands
Nanja.smets@tno.nl

M.S. Abbing
TNO /University of Utrecht
Soesterberg, The Netherlands
msabbing@gmail.com

M.A. Neerincx
TNO /DUT
Soesterberg, The Netherlands
mark.neerincx@tno.nl

J. Lindenberg
TNO
Soesterberg, The Netherlands
Jasper.lindenberg@tno.nl

H. van Oostendorp
University of Utrecht
Utrecht, The Netherlands
herre@cs.uu.nl

## ABSTRACT

Long duration missions set high requirements for personalized astronaut support that takes into account the social, cognitive and affective state of the astronaut. To cope with the high demand situations astronauts benefit from personalized support. Such support should be tested in different stages of development and as thoroughly as possible before testing and deployment into space. The in-space influences of the astronaut's state factors are hard and expensive to simulate on earth. But testing personalized support for astronauts in the right context (e.g. space) is important. We investigated if evaluation of prototype systems in a game-based environment (where the space environment can be simulated) contributes to producing a more elaborate, in-depth and realistic user experience. The outcomes of this research are important to everyone who wants to develop a support system for people in high demand situations (e.g. astronauts in space.)

## 1. INTRODUCTION

Future manned missions to the Moon or Mars set high requirements on astronauts. The environment is harsh and communication with Earth may show long delays. Because of these delays there is a greater need for autonomy in nominal and off-nominal situations. The astronauts on long duration missions have to cope with high demand situations themselves and they benefit from personalized support that takes into account the social, cognitive and affective state of the astronaut. The Mission Execution Crew Assistant (MECA) project aims at developing the requirements for such a system that empowers the cognitive capacities of human-machine teams during planetary exploration missions in order to cope autonomously with unexpected, complex and potentially hazardous situations (Neerincx et. al, 2006). We specified a theoretical and empirical founded Requirements Baseline (RB) for MECA, and its rationale consisting of scenarios and use cases, user experience claims, and core support functions. In order to refine and test the requirements baseline it has to be further tested using human-in-the-loop evaluations.

To conduct a good evaluation of personalized support, the context of evaluation must closely match the eventual context of use (Bevan, 1995; Jokela et. al, 2003). It is especially important for mobile devices, where the context can change constantly (in contrast to the static context of desktop applications) and environments such as space that are impossible to test in. Streefkerk and colleagues clarify the relationship between use context and user experience for context-aware mobile interfaces comprehensively (Streefkerk et al., 2008). The authors state that context-aware mobile user interfaces are developed to improve the user experience by adapting the system behavior, based on a model of relevant use context factors. Traditionally, evaluation is limited to laboratory settings and lacks the use of methods such as survey research, case study research and evaluation in real use contexts that give validity to the research. Although recreating central aspects of the mobile use context in the lab is sufficient to identify usability problems, the added value of field evaluation lies primarily in a deeper insight into the user experience in a dynamically changing context. According to Streefkerk et al., user experience validity can only be

assessed in expensive field tests. During evaluations early in the development process however, the validity of evaluation results is assumed to be improved by simulation of contextual factors that are presupposed to have an influence on user experience and system use. We investigated if evaluation of a prototype system (such as MECA) in a game-based environment contributes to producing a more elaborate, in-depth and realistic user experience (cognitive task load, situation awareness, trust and emotion), better task involvement, valid performance and more realistic experience. We distinguish two research questions:

1. Does the evaluation of prototype systems in rich evaluation conditions contribute to producing a more elaborate, in-depth and realistic user experience (affective and cognitive), better task involvement and more realistic experience of events in the scenario?
2. Does the evaluation condition influence the evaluation results of the tested system, i.e. does the context of evaluation have an influence on how the system is used.

In this paper we will first give some background on game-based evaluation and the research approach. Then the experiment is described and results are presented. We end with a discussion of the results and draw conclusions.

## 2. BACKGROUND AND RESEARCH APPROACH

Streefkerk et al. identify important future trends in the evaluation of mobile, context-aware systems. One of these trends is the employment of more and more diverse game-based techniques; another trend is the use of mixed-reality settings. Using a game to build a simulated environment in which a scenario has to be played and a system has to be used in real life, is an example of both these identified trends. An important difference between evaluating systems in such a manner as opposed to traditional usability evaluation is the simulation of context of use and the influence of the played scenario. An important issue in this type of evaluation in a simulated environment is the effect of the scenario and the environment (i.e. contextual factors) on the user experience and use of the evaluated system. For adaptive, context-aware systems the simulation of these factors is presumed to be valuable in particular, because this type of systems is designed for providing support to the user that is tailored to the dynamic context of use.

Computer games which serve a scientific purpose are mostly used in the field of training and education. Using games for the purpose of creating artificial environments in which (prototype) systems can be evaluated is relatively new. In the cognitive engineering method for the development of a decision support system for crisis situations, described by Te Brake et al. (2006), development iterations in a synthetic environment are followed by iterations in a real-world setting. The system in development is first evaluated and improved in an artificial setting until the system is ready for a real-world evaluation. Using games to create the artificial evaluation environments is done because of good visualization and user interaction possibilities at a relatively low cost (Te Brake et al., 2006). The Unreal 2 game engine (Epic Games Inc.), for instance used in the PC-game Unreal Tournament 2004 (UT2004), is one of the popular engines used for scientific purposes in general and the purpose of creating evaluation environments in particular. Modifications to this engine have been created, to allow external software to communicate with the game. UT2004 has been used by Te Brake and Smets (2007) for an experiment in which participants had to rescue victims in a simulated crisis environment. In order to execute this task, participants had to use a map (generated by an external application) on which victims were shown. Several advantages of game-based simulated environments in combination with an external application are mentioned, i.e. the high level of control over events taking place in the environment during experiments, the ability to automatically create a log file (e.g. for keeping track of when events occurred and storing task times), and the fast development and re-usability of created artificial environments (Te Brake & Smets, 2007).

The use of a virtual environment in the game-based evaluation is assumed to have both a momentary effect (depending on the specific moment in the evaluation) and an overall effect (which is built up during the whole evaluation). Additionally, the richer evaluation conditions in the VE are expected to induce a greater perceived feeling of presence. In this conceptual overview, presence is regarded as a direct effect of the VE, and as a premise for effects on the other variables. The measurement of presence assesses the VE's potential to affect the human-in-the-loop. The individual user differences are expected to influence the feeling of presence that is evoked by the VE, and influence the momentary and overall effects that the VE evokes. Fantasy proneness, an individual user difference, is expected to have an influence on the presence results. Participants with a vivid fantasy are assumed to be able to compensate for the lower

richness in a non-VE evaluation. The momentary effect consists of: emotion, situation awareness (Endsley, 2000), mental effort and performance. The momentary effects have influence on the overall effects. The overall effect consists of the following factors: satisfaction, acquired knowledge by using the system, trust in the system and the provided user feedback

With regard to the momentary effects, the richer evaluation conditions are expected to cause more intense and extreme emotions and a better feeling of situational awareness. Furthermore, the VE is expected to stimulate natural responses to arousing events in the scenario, e.g. if the colleague of the user faints in the scenario, this will have a more realistic emotional response in the VE. This is expected to be resembled in system users' mental effort being higher in predefined critical parts, because of an increase in motivation to take action, in order to respond to the situation. Performance is expected to be positively related to mental effort, and is consequently expected to increase more if participants witness an arousing event in the VE. In other words, system users in a rich VE are expected to have a higher motivation for task execution, and correspondingly have a higher mental effort and a better performance compared to their counterparts in a less rich environment.

The momentary effects altogether affect the overall effects. The overall effects consist of the following factors: satisfaction of use, acquired knowledge by using the system, trust in the system and the provided user feedback. The choice for measuring acquired knowledge was made because it resembles how thorough the system was used by the participants. If the information provided to the user via the system was processed in a deep and meaningful way, more of the information will be remembered at a later stage. It is assumed that the VE stimulates participants to use the system more thoroughly, and that the acquired knowledge will therefore be higher when a VE simulation is used. Trust is measured because it is closely related to the concept of user experience, and it is especially important for context-aware mobile interfaces for the professional domain. The mean values for trust are expected to differ between conditions. A higher variance in trust (more extreme scores) is expected to be manifested when a VE is used, because system users will be able to better describe how good they trust the system because of the immersive experience. How content the users are with the system altogether can be measured trough satisfaction. A difference in how satisfied users are between rich and poor evaluation can be expected because of the different experience (the same reason for the expected higher variance in trust). Finally, participants are presumed to provide a larger quantity and higher quality of user feedback (e.g. missing system functionality) on the tested prototype in rich evaluation conditions, because the immersive experience enables them to be more detailed and profound in their comments. The experience is expected to enhance the level of feedback because users are able to imagine (or even experience) what problems arise during the use of the system in the real use context.

# 3. METHOD

## 3.1 Design

A between subjects design with two conditions was adopted; the Cave Automatic Virtual Environment (CAVE) condition (game-based evaluation) and the static condition (scenario-based evaluation). Both conditions were controlled by a Wizard of Oz (WoZ) to simulate not yet implemented functionalities of the prototype. A WoZ application is traditionally intended for the facilitation of human intervention during experiments, as Streefkerk et al. describe in their framework for selection of evaluation methods. During a WoZ evaluation, the participant interacts with a seemingly fully functional system that is actually (partially) operated by the experiment leader. The advantages of using a WoZ evaluation method are the low-cost and early in the development process. The WoZ application can be extended to add game communication functionalities used for the simulation. By doing this, an application is created that both controls the simulated game environment (e.g. UT2004) and the missing functionalities of the system in development that is subject to evaluation.

## 3.2 Task

In both conditions, the participants had to conduct tasks and monitor objects and persons relevant for a long duration mission to Mars. The instructions for the tasks were the same in the two conditions, the mediation differed however. All tasks were assigned in the form of conversational messages, to make it seem like a (virtual) character in the scenario was giving the task. An example of a message is:

> "Hey Brenda, this is Benny. I'm to your right. Let's start working. We have to collect some research samples. The procedure is in the knowledge base. Could you look this procedure up in the knowledge base of your MECA-device and read it to me, please?".

In the static condition, the conversational message was displayed on screen. In the CAVE condition, the conversational message was provided by playing back a prerecorded message that contained the exact same information. In both conditions, the task time was recorded from the moment the participant started executing the task. Task instructions were only provided once, and could not be reviewed later; in the static condition, the screen went blank as soon as the participant started executing it and the conversational voice messages in the CAVE condition were not repeated. If a participant was not able to complete a task within 180 seconds, they were told that an alternate solution was found so the scenario could continue.

The scenario used for the experiment was adopted from one of the use case scenarios made during the development of MECA. This scenario contains a setting of two astronauts: Benny and Brenda. The latter was played by the participant. The astronauts are on an Extra Vehicular Mission (EVA) on the surface of the moon. The scenario starts with the astronauts performing nominal procedures (collecting research samples of rocks). After a few minutes, the temperature of Benny's spacesuit starts to increase, followed by the increase of Benny's temperature. MECA discovers this and warns the astronauts. A self-check reveals the cause of the discovered abnormality: the suit cooling device is failing. A rescue operation is set up, and for this purpose the participant instructs a Rover to pick up the team, and transport it to the habitat (lunar base). Meanwhile, the participant checks if anything can be done for Benny at this stage and monitors his vital signs. Because of the still increasing temperature, Benny faints before the Rover arrives. The participant instructs the Rover to pick up her fainted teammate and climbs on the Rover. The Rover heads for the habitat, where the medical facility has been prepared.

## 3.3 Variables

The dependent variables that were measured in the experiment are listed below:

- *Situation awareness.* SA was measured by a subjective questionnaire about the order of events in the scenario and relative location of objects or persons.
- *Emotion.* The participant had to fill in an adapted version of the Self Assessment Manakin (SAM) developed by Bradley and Lang (1994). SAM consists of three scales: arousal, valence and dominance. Because the dominance scale proved to be the most ambiguous and to explain the least variance, the choice was made to not measure dominance in the experiment (Neerincx & Streefkerk, 2003).
- *Mental effort.* Mental effort was measured using a scale based on the Rating Scale for Mental Effort (RSME) developed by Zijlstra (Zijlstra, 1993).
- *Performance.* The performance consisted of how fast the participants completed the tasks.
- *Satisfaction.* Subjective data on satisfaction was gathered in a questionnaire.
- *Trust.* Trust was measured by means of a questionnaire.
- *Presence.* Presence was measured by filling in the Igroup Presence Questionnaire (IPQ) (Schubert et al., 2001).
- *User feedback.* Participants had to fill in which functionalities they missed in MECA, and were asked to rank these missing functionalities from least to most important.

## 3.4 Material

The device used in the experiment was an Ultra Mobile Personal Computer (UMPC), of the type Sony Vaio UX running Microsoft Windows Vista (Business edition). The device has a 1.33 GHz processor and 1 GB of DDR2 RAM. The screen has a resolution of 1024x600 pixels. A development version of MECA was used on this device, with limited functionality to simplify the program for the inexperienced participants. This version included a knowledge base (taxonomy of procedures, objects, astronauts and other concepts in the environment) and monitoring functionality (an overview of properties for humans and vehicles in the surroundings of the user). See Figure 1 for a picture of the UMPC device running the MECA prototype.


**Figure 1: UMPC device on which the MECA prototype was running.**

The participant had to interact with simulated elements in the scenario by means of voice communication, e.g. a Rover had to be ordered to come to a certain location by giving voice commands. To facilitate this kind of functionality, a WoZ application was developed in the C# programming language using Microsoft's Visual Studio .NET 2003. The WoZ application had functionalities for playing conversational pre-recorded messages at the right moment in a conversation with the participant, measuring of task times and controlling the vehicle used in the VE.

The virtual world was made in UT2004, and contained a lunar setting (unadorned, with various sized rocks and a black sky with stars). Furthermore, the world featured moving objects, namely a moving astronaut a Rover. Urban Search and Rescue Modification (USARSim), a modification to the UT2004 game engine, was used to facilitate the use of vehicles in the VE. USARSim was originally developed for experimenting with human robot interaction and robotic behavior in simulated urban search and rescue environments (Wang et al., 2005).

Both test conditions took place in the same room, but not simultaneously. The two conditions were separated by screens to block vision and to diminish distractions. In the Cave condition, the participant had to use the prototype in front of a large screen, on which a virtual world was shown.

The CAVE condition featured a beamer, used to project the virtual surroundings of the test subject in the scenario on two screens in a 90-degree angle. CAVEs originally have projections on three or more sides, but with modest means (e.g. blocking peripheral vision) an attempt was made to create a resembling immersive experience. To block the peripheral vision and consequently force attention to the projected image on the screen, participants wore a helmet. See Figure 2 for a schematic overview of the CAVE condition and Figure 3 for a picture of a participant in the CAVE condition. While conducting the CAVE condition, participants kneeled on cushions. This choice was made to trigger an effect of actively taking part in the scenario instead of statically viewing (which was expected to occur in a normal sitting position, e.g. on a chair). Participants received sound (heartbeat and breathing sounds) and audio messages through ear buds.

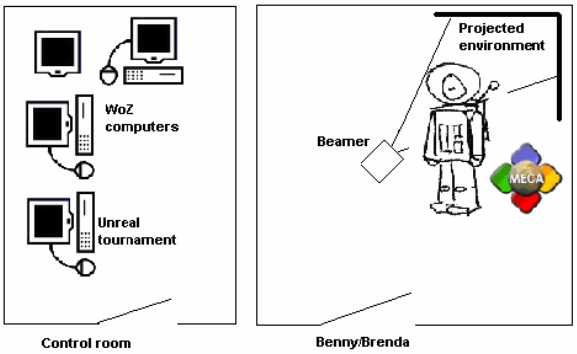



**Figure 2: schematic overview of the Cave condition.**

MECA was used by holding the UMPC with both hands. Because the participants were instructed not to physically move during the experiment, the wired network connection was not inconvenient. Participants could also not move in the VE, all movement was controlled by the experiment leader, via the WoZ application and UT2004 running on the control computers. Although interaction with the VE is one of the main contributors to presence, this choice was made to limit the participants' freedom to make sure important scenario events were not missed. Environmental interaction did occur in the form of commanding a vehicle and communicating with a virtual colleague. Conversations with colleagues were simulated by pre-recording several sentences, which could be played back at the right moment in the scenario by either the WoZ application automatically or by the experiment leader through this application. Messages from MECA to the participant were provided in the same manner. The use of pre-recorded sentences was possible because of the strictly structured scenario, in which ambiguity of 'what to say' in order to execute a task was slim to non-existent.

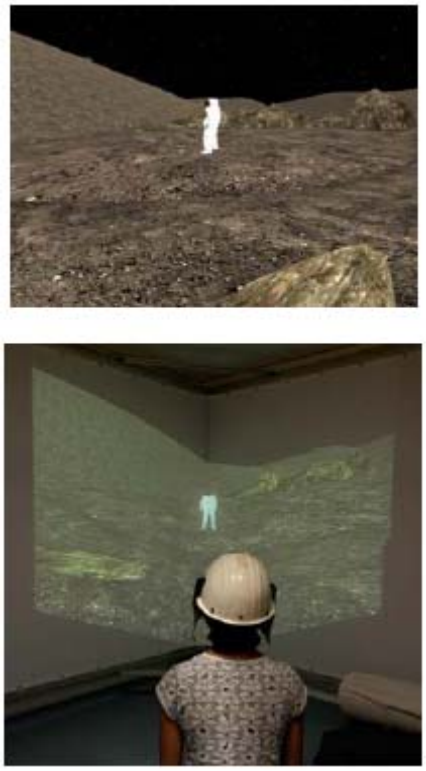
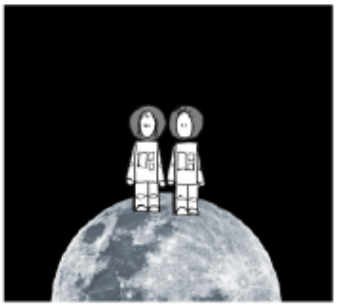


**Figure 3: screenshot of the projected virtual environment and the CAVE.**

The static condition was situated in an office-setting, with the participants sitting behind a desk with a standard TFT monitor on it; input devices were not provided. The scenario was presented by showing the storyboard drawings on the computer screen. Conversational task messages were shown below the scenario picture. A screenshot of the screen image and the setting of this condition are shown in Figure 4. MECA was used by placing the UMPC in a holder for using the device on a desk, provided by the manufacturer. The difference between having to hold the device up in the air and being able to control it from the standard is not expected to have influenced results, because hand placement on the device was similar.



Brenda Ross (you) and Benny Parker are on the surface of the moon, on a EVA (Extra Vehicular Activity). Your goal is to collect rocks for research purposes.

**Figure 4: screenshot and setting of the scenario-based evaluation.**

## 3.5 Participants

Twenty-five paid volunteers participated in the experiment, thirteen in the CAVE condition (seven male and six female) and twelve in the static condition (six male and six female). The average age was 23, with extremes at 18 and 31. All participants used a PC on a regular basis, one subject reported using a PDA before and none of the participants used a UMPC before. Few of the participants had experience with VEs: two of the twenty-five participants had been in a CAVE before, and three participants had worn virtual reality 3D-glasses prior to the experiment.

## 3.6 Procedure

At the beginning, participants were given a general, written instruction about the experiment. Then participants had to fill in a general questionnaire with demographic questions and questions related to experience with digital environments (e.g. computer games), immersive displays and handheld devices (e.g.

PDA's, portable game devices). After the general questionnaire the participants had to fill in the Creative Experiences Questionnaire (CEQ) by Merkelbach et. al (2001). The last pre-test questionnaire was the Big-Five Personality Questionnaire, used to assess personality traits (Goldberg, 1992). This questionnaire is one of the commonly used questionnaires to assess personality traits. It was presented in Dutch, because all participants were native Dutch speakers and the English terms for the personality descriptions were expected to be not commonly known among the participants. The Big-Five questionnaire and the CEQ were used to assess individual user differences.

The participants followed a training session to familiarize participants with the device in general, as well as with MECA's functionalities in particular. The session consisted of:

1. Learning how to answer questions by using MECA's knowledge base.
2. Learning how to carry out textual instructions in the form of step-by-step guidance of actions relevant for the experiment session.

Every time a participant finished a task, emotion and mental effort were measured by means of a short questionnaire. After the experiment was finished, the participants had to fill in a questionnaire to measure satisfaction of use and trust in the system, the IPQ, and a number of questions assessing situation awareness and acquired knowledge by using MECA

# 4. RESULTS

## 4.1 Presence

To assess differences in the perceived feeling of presence for the participants, scores for the three factorial components of presence (spatial presence, experienced realism and involvement) and a score for the overall presence were constructed per participant. A T-Test revealed a significantly higher mean for the CAVE condition ($t(11) = -2.824$, $p = 0.01$) for one of the three components (spatial presence). The mean values for the other two components (experienced realism and involvement) and the overall concept of presence were not higher in the CAVE condition.

## 4.2 Momentary effects

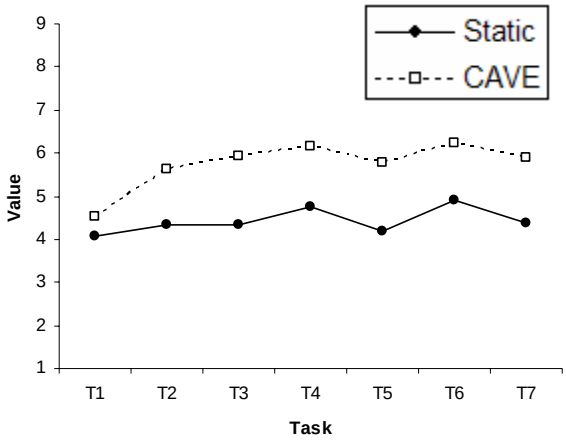
The expected momentary effects of the VE evaluation apply to the following variables: situation awareness, emotion, mental effort and performance.

Situation awareness

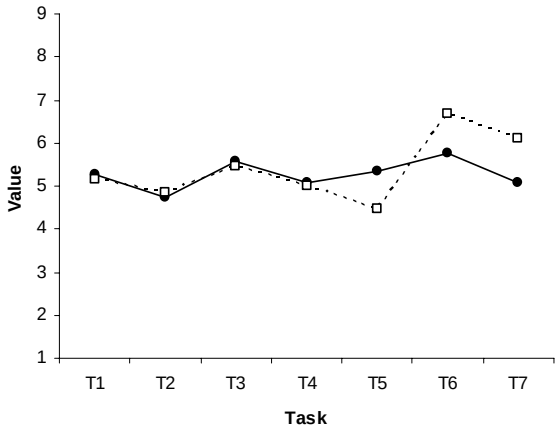
The mean value per participant over five situation awareness questions was significantly higher in the CAVE condition ($t(11) = -2.610$, $p = 0.017$), values range between 0 and 1.

Emotion

The participants' mean arousal over all measurements was higher in the CAVE condition (4.41 for the static condition and 5.73 for the CAVE condition, $t(11) = -3.318$, $p < 0.01$), the mean valence did not differ significantly.




a – Arousal




b – Valence

**Figure 5: Arousal and valence means per measurement**

Plots of the absolute scores of arousal and valence are displayed in Figure 5. The arousal scores measured after the nominal task 1 (Figure 5a) start at the same level for both conditions. After this task an unexpected event occurred to the participants: a warning message was received regarding an abnormal temperature value

for the suit of the virtual colleague. The first measurement of arousal after this event (T2) shows a non-significant difference between the values for the two conditions, which becomes significant ($p < 0.05$) from the third measurement onwards. Interesting to note is that the arousal plots are nearly parallel from T3 onwards, with a minor dip in both plots at T5. The valence plots (Figure 5b), show similar scores for the two conditions over the first four measurements. From the fifth measurement, the plots diverge; the mean valence score for the CAVE condition in measurement T5 is lower than the score for the static condition. Between measurements T4 and T5, the participants' virtual colleague fainted. Following measurement T5, a Rover that was instructed to pick up both astronauts and transport them back to the habitat (base camp) approached. Task 6 involved using the crane on the Rover to pick up the fainted colleague. Valence values that were measured directly after this positive scenario development and the final measurement values for valence were higher in the CAVE condition. Although the change in tendency of the valence plot for the CAVE condition is interesting, none of the differences in valence per measurement between conditions are significant.

Mental effort

The required mental effort to perform tasks using the system was measured at seven times (directly after each task) during the experiment (Figure 6). The average mental effort over all measurements was significantly higher in the CAVE condition ($t(11) = -2.257$, $p = 0.038$). Only the individual measurement after task 3 differed significantly ($t(11) = -2.099$, $p = 0.048$), although the differences in measurement values after task 4 ($t(11) = -2.018$, $p = 0.055$) and task 5 ($t(11) = -1.927$, $p = 0.072$) approached significance. The static condition has lower mental effort values than the CAVE condition for all measures, except for the measure taken after the second task. Between task 1 and task 2, the operation changed from executing nominal procedures (planned task execution, conforming the expectable) to emergency procedure execution. An interesting analogy with the arousal plots (Figure 5a) is noticeable: the conditional plots for arousal are parallel from T3 onwards; this also applies to the plots for mental effort.

Performance

The task times of the four relevant tasks for performance were compared between conditions, and tested for significance using T-Tests. Plots of the mean task performances (less time is better) per condition revealed an interesting effect (Figure 5). The first measurement's (T1) mean performance comparison shows better performance for the participants in the static condition (not significant). This is followed by similar performances of both groups with regard to second performance measurement (T3). The faster performance of participants in the CAVE condition on the third measurement (T4) was significant ($t(11) = 2.584$, $p = 0.017$). The final mean performance comparison shows similar performances for both conditions. Task 4, the only task which showed significant performance differences, involved looking up a procedure to treat hyperthermia (of which the virtual colleague was suffering) at the astronauts' current location. The differences in variance of task performances were compared for the two conditions, but did not show significant differences.

To assess if individual user differences had an effect on the found momentary effects, the correlations between the individual user difference variables (personality trait scores and scores for fantasy proneness) and the discovered momentary effects were calculated. No significant correlations between individual user differences and the significant momentary effects results were found.

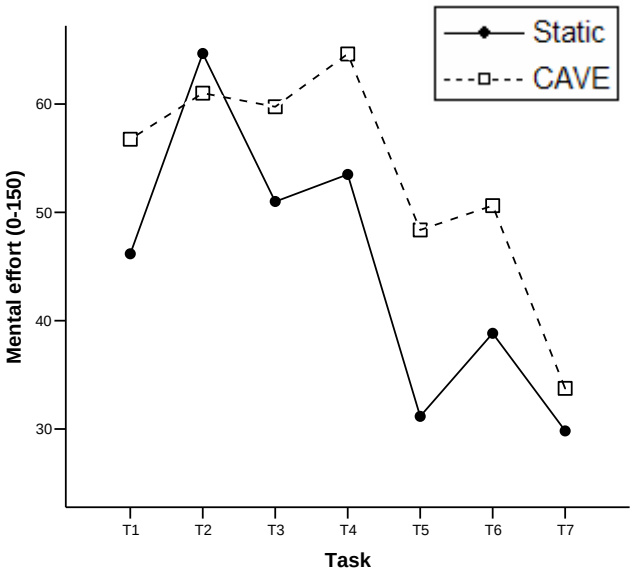



**Figure 6: Mental effort value means per measurement**

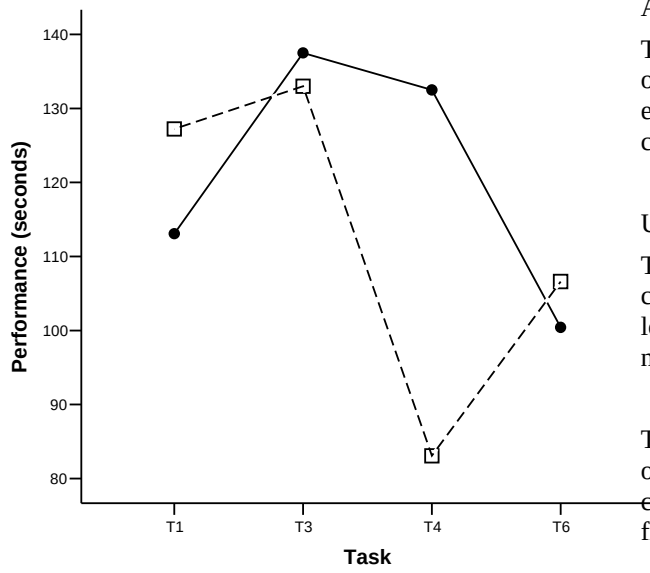



**Figure 7: Performance (task time means) for the four relevant tasks**

## 4.3 Overall effects

The expected overall effects of the use of a VE evaluation setting were value differences for the following variables: satisfaction of use, trust in the system, the acquired knowledge using the system and the provided user feedback (quality and quantity). To assess the quality of the feedback provided by the participants on missing system functionalities, the participants were given a quality rating by a usability evaluation expert. All overall effect variables were converted to a range between 0 and 1, for the user feedback variables this involved converting the values relative to the minimum (0) and maximum (1) values scored in the experiment. The results for these variables were compared between conditions, and tested for significance using T-Tests.

Trust

The expected higher variance in trust values for the CAVE condition was not supported by the results.

Satisfaction

The non-significant difference between conditions with regard to the mean of the scored satisfaction values did not provide support for the assumed mean satisfaction difference.

Acquired knowledge

The results do not support the expected greater amount of acquired knowledge in the CAVE condition and the expected difference between mean trust values for both conditions.

User feedback

The user feedback quantity was higher in the CAVE condition; however the quality of user feedback was lower in this condition. Both mean differences were not significant.

The influence of individual user differences on the overall effects was checked by performing analyses of covariance. The results of these tests did not differ from the results without covariates described above.

# 5. DISCUSSION

Scenario and VE involvement

Results confirmed the expected more intense feeling of presence when using a VE: the IPQ component spatial presence was higher for participants in the CAVE-condition. Individual user differences such as character traits did not influence presence. Similar to the presence results, the values for situational awareness were also higher in the CAVE-condition. These results indicate that a richer scenario experience helps participants to get involved in not only the virtual scenario environment (sense of being there, etc.), but also in the scenario itself. The better scenario involvement can be induced from the higher situational awareness in the CAVE-condition, which was partly measured by having the participants recall the order of events that took place in the scenario. The higher value for situational awareness in the CAVE-condition indicates more thorough processing of the events that occurred in the scenario.

Emotion

In general, the richness of the scenario environment shows to have substantial momentary effects in prototype evaluation. Scenario storytelling in a more elaborate way is associated with higher participant arousal levels, indicating a possible positive causal relationship between richness of virtual scenario evaluation environment representation and arousal. The expected higher variance in emotional state for the CAVE-condition was not supported by experiment results, but this could be due to the manner of measuring (self-assessment); objective physiological

measurements might reveal a relationship. The used scenario featured events that were expected to evoke an emotional response, such as the transition from normal to abnormal procedure execution when the temperature problem occurred and the moment the participant's colleague fainted. Furthermore, the CAVE-condition was expected to evoke more extreme emotional responses than the static condition. The gathered arousal data do not support the first expectation: arousal values do not increase substantially after the critical events took place in the scenario. Arousal values for the CAVE-condition are generally at a higher level, but do not fluctuate more as was expected. The higher overall mean arousal value for the CAVE-condition might be an indication that participants enjoyed participating in the evaluation in that condition more; a high enjoyment level was also reported by some of the participants informally.

The valence data show an interesting difference between two conditions: after the moment the astronaut faints, the two plot lines diverge because of a decrease in mean valence for the CAVE-condition. The following event is the arrival of the rover that was instructed to come, and in the measurement following this event the participants in the CAVE-condition suddenly report more positive valence values than their counterparts in the static condition (in contrast to the previous measurement). This non-significant effect is an indication for more extreme emotional responses to critical events in the CAVE-condition. The first four measurements show similar, constant values for both conditions, a possible indication that the events occurring during that period might not be able to evoke emotional responses altogether.

### Personification

Some interesting conclusions can be drawn after analyzing mental effort and performance fluctuations before and after critical scenario events. The first critical event is between measurements T1 and T2: the discovery of the overheating. Mental effort for both conditions increases after this event, which might be an indication for stress. In task 4, participants had to use MECA to find out if they could do anything for Benny while waiting for the vehicle to arrive. The large performance difference between tasks for the measurement following this task (T4) is remarkable, because this effect only exists for this task. Task 4 might be characterized as the only task in which the required action of the user is pointed directly at providing help to the colleague in trouble. The better performance for CAVE-participants indicates that participants in this condition perceived Benny as being a simulated person, instead of a conceptual storyboard drawing for which it did not matter how they performed for the development of the scenario. The rising mental effort, which reaches the highest mean of all CAVE-measurements, supports this indication: participants' motivation to perform well at executing task 4 might be higher in the CAVE-condition. Possibly, the presumed boost in motivation is in consequence of cooperation and team feeling, as for example occurs in collaborative computer gaming, with artificial or human collaborators. A conditional difference that could be responsible for causing this effect is the visual richness. The participant's virtual colleague was struggling and clearly in desperate need of help in the CAVE-condition, whereas in the static condition a conceptual drawing of the suffering astronaut was shown.

### Value of use context simulation

The events in the scenario have shown to have a bigger impact on participants when a VE simulation is used. This has secondary effects, such as the presumed greater personification of simulated team members in such a setting. Indications for a better task involvement and a more realistic experience of events in the scenario have been found in the form of a presumed motivation boost for executing a task involving helping a virtual team member in trouble. This also lead to a general difference in system use, in the form of a better performance. The mean arousal was found to be higher in the CAVE-condition, perhaps an increase in the level of arousal towards a value occurring in a similar real environment evaluation.

The presumed overall effects of evaluation richness (in the form of trust, satisfaction, acquired knowledge and user feedback) were not supported by empirical evidence. This leaves hypotheses on participants' specificity with regard to the system evaluation unsupported. For instance, the participants were not supported by the simulation in criticizing whether they trusted the system or not, and whether they were satisfied with the system or not. Perhaps the tasks in the experiment were too superficial and straightforward to measure these effects, trust differences might come to light when participants have to make more difficult decisions (and influencing the scenario) based on information provided by the system and execute more complex tasks using the system. Providing the participants with such complex tasks leading to different scenario paths was infeasible for this project. User feedback regarding system functionalities was not of higher quantity and quality in context simulated settings. More specific user feedback questions regarding the functioning of the system in a specific situation in the scenario (instead of

the general functionality feedback that was asked for in this experiment) might support the assumption that users can give more and better functionality feedback if scenario and use context involvement is higher.

Despite the unsupported hypothesized effects of use context simulation on system evaluation measures (trust, satisfaction and user feedback), the simulation of a system's use context for the purpose of evaluation is considered to be valuable and worth the effort. The added value of simulation lies in the facilitation of context and scenario involvement, which is expected to add to the validity of evaluation results.

## 6. CONCLUSION AND IMPLICATIONS

We investigated if evaluation of prototype systems in a game-based environment contributes to producing a more elaborate, in-depth and realistic user experience (cognitive task load, situation awareness, trust and emotion), better task involvement, valid performance and more realistic experience of events in the scenario. In the game-based evaluation, the participants showed higher arousal levels where expected, a more intense feeling of presence, better situation awareness, higher mental effort and faster performance when needed. A game-based evaluation seems to better address the social and affective aspects of the space mission. As a result of this research we have used game-based evaluation to refine the RB within the MECA project, see Neerincx et. al (2008).

Game-based evaluation makes it possible to evaluate mobile context-aware support systems in a simulated eventual context of use (e.g. space). The results showed that this has additional value to the outcomes of the evaluation of such a system. Game-based evaluation can be conducted in different stages of the development process. These results are important for everyone who wants to develop a support system for astronauts in space or any other high demand environment.

## 7. REFERENCES


[1] Neerincx, M. A., Lindenberg, J., Smets, N., Grant, T., Bos, A., Olmedo-Soler, A. et al. (2006). Cognitive Engineering for Long Duration Missions: Human-Machine Collaboration on the Moon and Mars. *Proceedings of the 2nd IEEE International Conference on Space Mission Challenges for Information Technology*, 40-46.

[2] Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal*, *4*(2), 115-130.

[3] Jokela, T., Iivari, N., Matero, J., & Karukka, M. (2003). The standard of user-centered design and the standard definition of usability: analyzing ISO 13407 against ISO 9241-11. *Proceedings of the Latin American conference on Human-computer interaction*, 53-60.

[4] Streefkerk, J. W., van Esch-Bussemakers, M. P., Neerincx, M. A., & Looije, R. Evaluating context-aware mobile user interfaces. *Handbook of research on user interface design and evaluation for mobile technology*, (2008).

[5] Te Brake, G., De Greef, T., Lindenberg, J., Rypkema, J., & Smets, N. (2006). Developing adaptive user interfaces using a game-based simulation environment. *Proceedings of the 3rd International ISCRAM Conference*.

[6] Te Brake, G. & Smets, N. (2007). Developing Adaptive Mobile Support for Crisis Response in Synthetic Task Environments. In *Usability and Internationalization. Global and Local User Interfaces.* (pp. 510-519). Berlin / Heidelberg: Springer.

[7] Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: a critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement* (pp. 3-32). Mahwah, NJ: Lawrence Erlbaum.

[8] 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.

[9] Neerincx, M. A. & Streefkerk, J. W. (2003). Interacting in Desktop and Mobile Context: Emotion, Trust and Task Performance. *Proceedings of the first European Symposium on Ambient Intelligence (EUSAI), Eindhoven, The Netherlands.Springer-Verlag*.

[10] Zijlstra, F. R. H. (1993). *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. Delft University Press.

[11] Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments, 10*(3), 266-281.

[12] Wang, J., Lewis, M., Hughes, S., Koes, M., & Carpin, S. (2005). Validating USARsim for use in HRI Research. *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*, 457-461.

[13] Merckelbach, H., Horselenberg, R., & Muris, P. (2001). The Creative Experiences Questionnaire (CEQ): A brief self-report measure of fantasy proneness. *Personality and Individual Differences, 31*(6), 987-995.

[14] Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26-42.

[15] Neerincx, M.A., Lindenberg, J., Smets, N.J.J.M., Bos, A , Breebaart, L., Grant, T., Olmedo-Soler, A., Brauer, U., Wolff, M. (2008). The Mission Execution Crew

Assistant: Improving Human-Machine Team Resilience for Long Duration Missions. *Proceedings of the 59th International Astronautical Congress (IAC2008)*, Glasgow.