A Usage-Centered Evaluation Methodology for Unmanned Ground Vehicles

Jurriaan van Diggelen, Rosemarijn Looije, Tina Mioch, Mark A. Neerincx, Nanja J. J. M. Smets
TNO
Kampweg 5, 3769 DE Soesterberg, The Netherlands
{*jurriaan.vandiggelen,rosemarijn.looije, tina.mioch,mark.neerincx,nanja.smets*}@*tno.nl*

***Abstract*—This paper presents a usage-centered evaluation method to assess the capabilities of a particular Unmanned Ground Vehicle (UGV) for establishing the operational goals. The method includes a test battery consisting of basic tasks (e.g., slalom, funnel driving, object detection). Tests can be of different levels of abstraction, and be performed in a virtual or real environment. In this way, several candidate UGV's in a procurement program can be assessed, and thus compared. Also, it can give directions to research on improving human-robot interfaces. A first case study of this methodology conveyed capability differences of two alternative user interfaces for a specific UGV with their operational impact.**

***Keywords*-Human-robot cooperation; Performance evaluation**

I. INTRODUCTION

Usage-centered Unmanned Ground Vehicle (UGV) evaluation is important for a number of purposes. It can be used in a procurement program, allowing several off-the-shelf candidate UGV's to be assessed against their operational needs, and thus compared. Furthermore, it can give directions to UGV development by clearly identifying shortcomings in operator-robot interaction. It can also help to select an adequate set of components (e.g., sensors, controllers, user interfaces) to be combined in the robot.

Nevertheless, performing a structured, well-founded UGV evaluation in practice poses a number of difficulties. Firstly, the usage environment may not be easily accessible. This can be because the UGV is intended to be used at a remote or dangerous location, or because the operator's final work environment is difficult to simulate in a laboratory setting. Secondly, the UGV may not be entirely available at the time of evaluation. For example, the UGV platform may be available and ready for testing before a decision is made about which sensors will be mounted on the UGV. Another example occurs when a UGV is still in its specification phase, and the manufacturers wish to perform an early evaluation of the requirements baseline before actually buying the physical hardware.

Because current UGV benchmarks are scarce and fail to adequately address these problems, we have developed a usage-centered evaluation methodology. The methodology is based on a test-battery and makes no prior assumptions on the *location* of the evaluation, or the UGV's *phase of development*. A test can be regarded as an *exam exercise*, constituting a simple atomic task, such as doing slalom, funnel driving, or object detection. If the robot, controlled by a qualified operator, passes the exam, it can be concluded that it has the basic capabilities that were identified as critical for the operational task. Because tests can be defined in an abstract way, and can be performed in both the real world as well as in the virtual world, they do not impose any constraints regarding the location of the UGV evaluation. Because tests are designed to be as elementary as possible, they can be used to partially evaluate a robot which has not been fully assembled yet. As missions are changing substantially, robot technology is progressing quickly, and relevant human factors knowledge is increasing continuously [1], the test battery should not be seen as a static entity. Rather, we regard it as a standard toolkit for UGV operators which should be updated regularly and tailored to the specific situations encountered. Furthermore, an additional summative evaluation with adequate fidelity will most often be needed at some point in the development cycle to assess all the dependencies between context, work organization, personnel, UGV and operational outcomes.

The purpose of this paper is threefold. Firstly, we present a structured and well-founded methodology for usage-centered UGV evaluation and design. Secondly, we give guidelines for UGV test development, considering issues such as abstractness and realism of tests and how this affects their validity, taking into account operational demands, human factors issues, and technological constraints and opportunities. Thirdly, we present a standard test battery and report on our experiences on using the methodology for robot evaluation. For a more in-depth investigation of the soundness of our approach, the reader is referred to [2].

We have started developing the standard test battery by designing around forty basic abstract tests together with the corresponding evaluation criteria, and their relation to operational demands. Because the tests are abstract, we can perform these tests at the location of our customers (either procurement officers or UGV manufacturers). The current test battery contains tests which must be performed in the real world, i.e. we have not yet included any virtual environment tests. Nevertheless, we have set up the methodology in such a way that we can straightforwardly extend the test battery with tests in a virtual environment, e.g., USARSim [3]. This allows us to design tests with virtual robot configurations in simulated environments, which are

ISBN: 978-1-61208-177-9

hardly available for (cost-effective) robot evaluations and/or might involve a danger of damaging the robot.

The paper is organized as follows. In the next section, we describe our abstract evaluation framework, and how testing can be performed at different phases of development. In Section III, we describe how tests can best be designed. Section IV describes an initial version of our basic test battery, and reports a case study, followed by a conclusion in Section V.

II. UGV Evaluation Method

The methodology follows a human-centered approach, i.e., we focus on a human operator who interacts with a UGV within a certain context. This is depicted in Figure II.

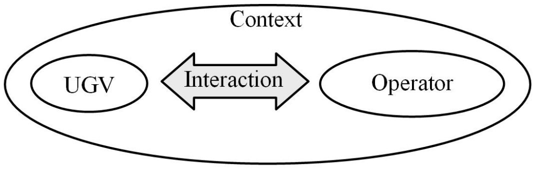


Figure 1. UGV-Operator System in Context.

Each of the four components in Figure 1 has an influence on the overall performance of the system. Below, we will outline some relevant factors for each of these components.

- UGV factors: These factors are typically described in the UGV's specification document. Examples are the energy consumption properties of a UGV, the amount of horsepower, the availability of different sensors, the availability of robotic arms and grippers, the maximum speed, size, weight, etc.
- Interaction factors: These are factors concerning the interaction between the UGV and operator. This may concern information which must flow from UGV to operator, such as sensor images and information on the slope of the terrain. It may also concern information which flows from operator to UGV, such as directions to the UGV that it must adjust its camera angle, or any other type of control action performed by the operator.
- Operator factors: Examples are: knowledge, skills, abilities and training level of the operator, fatigue, motivation, etc.
- Context factors: Example of relevant context factors are: properties of the soil, weather conditions, light conditions, etc.

Obviously, these four factors are interrelated. For example, darkness (a context factor) may obstruct proper UGV operation, unless the UGV has a night vision camera (UGV factor), and the interface allows to properly view these camera images (an interaction requirement). In this paper, we have chosen not to evaluate these different aspects separately, but to take all of these aspects into account at once. Hence, the evaluation measures the total operator-UGV performance.

A. Situated Cognitive Engineering

Our UGV evaluation methodology is a special component of situated cognitive engineering [4] that views system development as an iterative process in which system's functions are specified and assessed in a systematic way to establish a sound foundation of the specified and/or selected functions. An overview of this evaluation methodology is depicted in Figure 2.

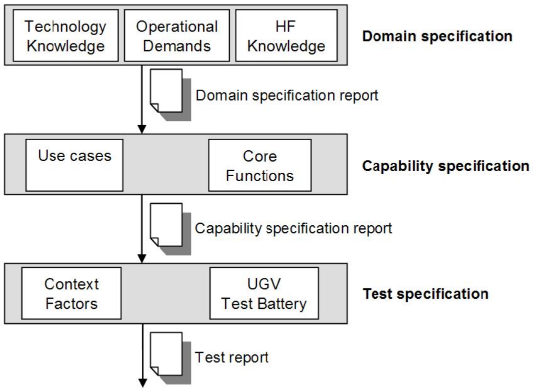


Figure 2. UGV Evaluation Process.

In the domain specification layer, the technology knowledge, operational demands and human factors knowledge are investigated. Usually this is done in a series of workshops with HRI experts and envisioned end users. During these workshops, the HRI experts are provided with information on what kinds of operations the users intend to use the UGV for, and what kind of UGV candidates they were considering. This information results in the domain specification report which specifies the following aspects:

- Envisioned operational use (which kind of environments, on which terrains, indoor/outdoor, weather types, envisioned operators, etc.)
- Initial technological investigation (potential UGV candidates, using which sensors and which interfaces, potential technological pitfalls)
- A list of human factors issues which are relevant for the envisioned technology and operational demands (such as operator sickness or risk of information overload).

The domain specification report forms the basis upon which the use cases and core functions are derived, as depicted in the capability specification layer. These two components are described in the capability specification report, which makes the required functionality of the robot operator system more concrete. This report describes a list of use cases which demonstrate the nominal and extreme use of the UGV, and a list of core functions.

In the final phase (the test specification layer), this knowledge is further refined into a concrete list of relevant context factors, and a concrete UGV test battery to which the robot is subjected. These tests comprise assessments for capabilities that are tailored to the (envisioned) UGV-supported operations, technological demands and human factors. The final judgment whether the robot is appropriate, is reported in the test report which describes the following:

- A selection of tests from the test battery, including information on why the test was selected, how the test was instantiated (which context factors were taken into account, practical constraints, setup of the environment).
- The UGV-operator system performance on the different tests, describing the performance on objective measures and subjective measures (feedback from the user about their performance).
- A final judgment which summarizes the test results.

The seven components represented by the white boxes in Figure 2 can be regarded as an evolving toolbox, which grows over time when more UGV's are subjected to the evaluation method and more operational demands and human factors are taken into account. The advantage for the UGV-evaluator is that he does not start from scratch each time a UGV is evaluated; the advantage for the method is that it affords continuous updates of the task battery. Thus, the evaluator can reuse use cases, tests, technology knowledge, and so forth, which have been developed in previous UGV evaluations. In this way, a set of generic, reusable core components are iteratively developed.

For a specific evaluation, an instance of the core test battery should be formulated, consisting of the selection and prioritization of the tasks and criteria for evaluation. Because the foundations of the core test battery evolve continuously, the compilation and definition of the constituting tasks should be updated regularly. As depicted in Figure 2, the three evaluation reports form the milestones in this process.

B. Evaluation dimensions

In general, evaluation experiments can differ in fidelity and realism [5]. We can apply the same categorization to the UGV evaluation tests. In this domain, fidelity indicates how close the test environment resembles the environment in which the UGV is planned to be used. For example, we can perform low fidelity tests in a laboratory, or in the real environment (high fidelity). Realism varies from one extreme-the real environment-to the other, a virtual environment. For example, instead of going to a real earthquake site to test a prototype, the prototype can be tested in a virtual environment. The test space is depicted in Figure 3.

Typically, UGV's are evaluated in a series of tests, starting with easily performable tests with low fidelity and realism to experiments in the real world. This is indicated by the arrows in Figure 3.

An example of a low fidelity and low realism test is a cognitive walk through. This was for example done in the Mission Execution Crew Assistant (MECA) project [4], which uses Situated Cognitive Engineering to validate the requirements baseline of an astronaut's cognitive support system. During the cognitive walkthrough evaluation, participants were talked through a use case and answered questions during and after the walkthrough. Storyboarding was used to illustrate the environment and operations.

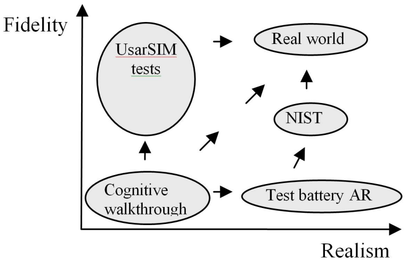


Figure 3. UGV Test Space.

After performing a cognitive walkthrough, you'd typically want to perform more extensive tests by either adding more realism, or by adding more fidelity. More realistic tests are described in Section IV of this paper, were we present the current version of our core test battery, i.e. test battery AR (Abstract&Real-world). The tests in test battery AR must be performed in the real world and have low fidelity, meaning that they abstract from real world details.

Realistic tests can also be performed with higher fidelity, for example the NIST test arena [6]. NIST evaluates robots in test arenas that are as realistic as possible. Because the tests we propose in test battery AR are more abstract (i.e. have less fidelity) than the test in the NIST arena, they are a valuable contribution to NIST evaluation. Although it is very important to evaluate the robots also in a realistic environment, we state that for a good assessment, the environment and its resemblance are not always essential. To test the functions or requirements in an early phase design specification, it is only necessary to let the robot and the operator interact and operate on tasks that need similar capabilities as the real operational task [7]. This makes testing more easy (relatively cheap and under controlled conditions) to perform.

Another path to follow when testing UGV's in an iterative way is to perform tests in virtual environments, such as USARSim [3]. USARSim is a high fidelity virtual environment, used for research of human robot interaction in urban search and rescue. USARSim uses the Unreal Engine and accurately represents interface elements (such as camera) and robot behaviour, furthermore the users of USARSim can create their own robots and environments [8]. Eventually, robots should be evaluated in the real end-user environment, which is represented in the upper right corner of the graph.

III. Test Design

This section describes how tests can be properly designed, regardless whether they are performed in the virtual world, or in the real world. We will first explain the general structure of the test battery. Then we will describe some general guidelines for test design, after which we discuss some issues regarding interpretation of UGV test performance.

A. Structure of the Test Battery

The tests in the test battery are categorized according to three levels in which robots are operated (from low-level operation to high-level operation):

- Executional: At the execution level, the operator performs elementary actions, e.g., accelerating the robot, observing an object, etc.
- Tactical: At the tactical level, the operator executes a plan of actions. Most of the time, this is done during the mission. For example, the operator follows a route, defined using a number of waypoints.
- Strategical: At the strategic level, the operator forms a plan. For example, for the basic function transit this could be deciding on the waypoints for the route based on meteorological data.

Proper operator-robot interaction at these different levels requires different interface properties. For example, accelerating a robot (a task at the executional level) requires the operator to be able to properly adjust the robot's speed. Deciding on the waypoints for a route (a strategic task), requires the operator to have spatial situation awareness of the area where the robot is to be operated. Therefore, the test battery should contain tests at each of the three levels.

Typically, higher level tests are dependent on the lower level tests. For example, the test ”do a slalom” (tactical level) is dependent on the test “make a turn” (executional level). If test A is dependent on test B, we mean that test A can only be passed if the agent also passes for test B. For each test in the battery, we the dependencies of the test must be made explicit.

B. Guidelines for Test Design

The experiment was within subject, and each participant first performed the test battery tasks, followed by the scenario.

C. Materials

A test can be viewed as a package containing a short name of the test, a unique identifier, a description of how the test is performed, the performance measures, the test's dependencies and a lab setup. Furthermore, we specify (in another table) the relation with operational demands. For the different elements of a test, related literature should be consulted on UGV metrics, testing environments, and operational literature.

Based on our experiences on designing tests for UGV's, we have identified the following aspects as deserving special attention:

How is the test evaluated?
In general, there are the following options:

- *Result-based*: Whether the operator is capable to accomplish a certain result. For example: can the operator do a slalom within 30 seconds? Within result-based evaluation we have two ways of doing that:
  - *Subjective*: an examiner determines whether the test has been passed
  - *Objective*: the test is passed by some objectively measurable criterium, such as time it takes to finish a trajectory.
- *Questionnaire-based*: useful for evaluating the operator's situation awareness. For example: ask the operator to draw a map of the environment after having moved around the environment for a while.

From our experiences with robot evaluation during the development of the test battery, we found that often a combination of result-based and questionnaire-based evaluation works well. For example, after the operator has finished a result-based test, we ask the operator what his or her experiences were and to estimate his or her performance on the test. It turned out that often the operator did not know that (s)he was disqualified for the test because (s)he was unaware of bumping into other objects during the task.

Can the test be decomposed? We aim for the tests to be as elementary as possible, following a reductionistic approach. This means that if we believe that a test actually measures two distinct independent aspects, we decompose the test in two separate tests for each of those two aspects. Of course, being able to do two things separately does not always imply that these two things can be done simultaneously as well. In those cases, we also include the test for doing the two things simultaneously in the test battery, provided that the composed task is realistic in an operational setting. The benefit of having the elementary tests as well is that it improves diagnostic power to our evaluation methodology.

Is the test discriminatory? Tests should be discriminatory in the following respects:

- It should not be a trivial test that every UGV can pass
- It should add something to the existing tests in the battery. The test should test a capability (or several capabilities) of the robot-operator system that in this combination are tested by no other test in the test battery.

D. Optimal Performance

To interpret the performance measure of a test, it is useful to understand what would be the *optimal* performance on the test. In general, the robot-operator performance is determined by two factors. Firstly, it is determined by the robot properties. This is referred to as the *inherent capability*. For example, a robot can have an inherent capability of moving at a maximum speed of 20 kmh. Secondly, the robot-operator performance is determined by the operator controlling the robot. This is referred to as the *piloted capability*. The piloted capability can never be greater than the inherent capability of the robot. For example, an operator cannot make a robot go faster than its maximum speed. Typically, the piloted capability is lower than the inherent capability. For example, when doing a slalom, it is unlikely that the operator can move the robot at its maximum speed. In case the piloted capability equals the inherent capability of the robot, we can say that the test has been passed with *optimal performance*.

When the inherent capability of the different UGV's is different, the test results of the different UGV's cannot be straightforwardly compared. For example, steering an unmanned tank during a slalom task is much more difficult than steering a medium-sized robot, because it is a much larger robot. If the evaluator's interest is at the HRI properties of the two systems, this would be an unfair comparison. The comparison could be made more ”fair” by comparing the actual performance of the test, with the optimal performance.

IV. Case Study

We have applied our method in the domain of military UGV's. During the domain specification phase, we have identified the following four operational demands: *Transit*: The UGV should be able to transit from one location to another; *Observe*: contains all tests that are focused on the collection of mission relevant information; *Manipulate*: The manipulation of (objects in) environment, both direct (disposal of IED) and indirect (grenade throwing) manipulations are possible; *Communicate*: communication between operator/UGV with the other stakeholders.

Use cases developed for the NIFTI project, contain observe tasks. For instance when a UGV is deployed to retrieve the exact location of a victim in a collapsed building, the UGV will be tele-operated through an entrance (this can be a narrow passageway) and the UGV and operator will observe the environment for the victim by means of camera and sound.

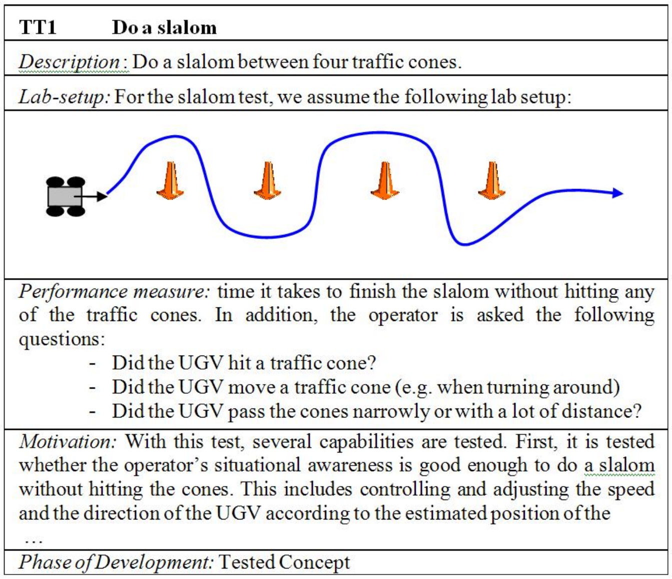
A. Test Battery AR

The investigation of this case-study has resulted in a first version of a standard core test battery, i.e. test battery AR. All tests are abstract real world tests. A fragment of the core test battery is shown in Table IV-A.

For each of these tests, more details are provided on how the test should be carried out and which aspects are deemed important. For the slalom task, this is described in Figure 4.

Nr	Name	Dependencies
Transit	**Executional**	
TE1	Accelerate	-
TE2	Slow down	-
TE5	Accelerate backwards	-
Transit	**Tactical**	
TT1	Do a Slalom	TE4, OE4, OT2
TT2	Stop before collision	TE1, TE2, OE1, OE4
TT9	Accelerate backwards following a straight line	TE5
Transit	**Strategic**	
TS1	Find a way through the maze	TT1
TS3	Return to starting point	TE4, OS1
TS6	Estimate whether the UGV can drive up a slope	OE2

Table I
Fragment of Test Battery AR



TT1 Do a slalom

Description: Do a slalom between four traffic cones.

Lab-setup: For the slalom test, we assume the following lab setup:

Performance measure: time it takes to finish the slalom without hitting any of the traffic cones. In addition, the operator is asked the following questions:

- Did the UGV hit a traffic cone?
- Did the UGV move a traffic cone (e.g. when turning around)
- Did the UGV pass the cones narrowly or with a lot of distance?

Motivation: With this test, several capabilities are tested. First, it is tested whether the operator's situational awareness is good enough to do a slalom without hitting the cones. This includes controlling and adjusting the speed and the direction of the UGV according to the estimated position of the …

Phase of Development: Tested Concept

Figure 4. Example of test description

B. Performing the tests

In our pilot experiment, we experimented with the Eye-robot (see Figure 6). The robot was operated from another room, using two possible user interfaces, i.e. a graphical user interface (GUI), and a telepresence interface (see Figure 5). The GUI interface consisted of a monitor with the robot's camera images and a joystick for steering. The telepresence interface is a more advanced interface, consisting of a head-mounted stereo vision device, together with a gas-pedal and a steering wheel for controlling the robot. Note that, in this paper, it is not our purpose to provide a complete investigation of these different interfaces. Rather, we wish to demonstrate our evaluation methodology by showing the feasibility of using early testing using the standardized test from test battery AR.

We have subjected both UI configurations to a selection of nine tests from the test battery. Some tests were discriminatory, among which TT1, TT2, and TS3. This means that they indicated a clear difference in performance between the GUI and the telepresence interface (mostly in favor of the

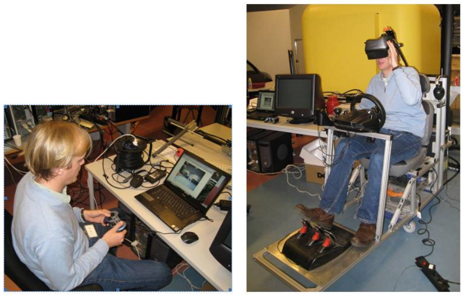


Figure 5. GUI interface (left) versus telepresence interface (right)

telepresence interface). Some of these differences appeared in the objective evaluation criteria specified with the test, such as the time it took to complete the slalom task. Other differences in performance were subjective, e.g., using the GUI interface the operator failed to notice that he bumped into one of the traffic cones. In conclusion, this case study of the methodology conveyed capability differences of two alternative user interfaces for a specific UGV with their operational impact.

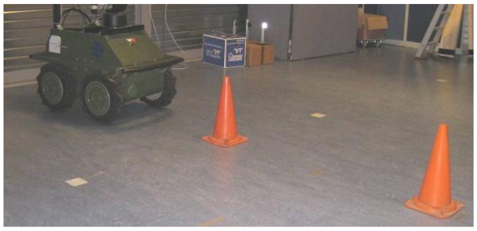


Figure 6. The eye robot during the slalom test

V. Conclusion and Future Work

Standardized test methods for human-UGV interaction are important for Human Robot Interaction research. In this paper we have presented a usage-centered UGV evaluation method which is centered around a battery of basic tasks. We have also proposed guidelines to design appropriate tests, taking into account ease of testing, phase of UGV development, and the situatedness of the robot.

From our experiences with using our UGV test methodology, we believe that the benefits of using the method are twofold. Firstly, it provides a well-founded and structured way to perform a single UGV evaluation. It allows us to compare different UGV configurations (such as alternative user interfaces), and to benchmark designs. Secondly, it provides a good way to organize and store knowledge in our research team which is gained during prior UGV evaluations. For example, if we would experience that some maneuver is difficult for many UGV's, we would add a test to the standard battery that targets exactly this aspect. In this way, we can guarantee that this aspect is properly addressed in next evaluations, and that it is recognized by all members of the UGV research team.

Addressing the interrelationships between the UGV, operator, interaction and context factors remains an important issue for further experimentation. In the future we plan therefore to specify and perform a summative experiment that is based on complex realistic use cases and utilizes the task battery knowledge-base. Also, we intend to extend the current test battery with tests that are to be performed in the virtual world. This would make our iterative testing method more complete, and would allow us to complement our existing abstract real world tests.

Acknowledgment

This research is supported by FP7-EU grant for the NIFTi project (#247870FP7).

References

[1] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[2] T. Mioch, N. Smets, and M. Neerincx, "Predicting performance and situation awareness of robot operators in complex situations by unit task tests," in *Proceedings of the 5th International Conference on Advances in Human Computer Interaction (ACHI)*, 2012.

[3] M. Lewis, "USARSim and HRI: from teleoperated cars to high fidelity teams," in *Proceedings of the IROS 2009 Workshop on Robots, Games, and Research: Success stories in USARSim*, 2009, pp. 18–22.

[4] M. Neerincx, J. Lindenberg, N. Smets, A. Bos, L. Breebaart, T. Grant, A. Olmedo-Soler, U. Brauer, and M. Wolff, "The mission execution crew assistant: Improving human-machine team resilience for long duration missions," in *Proceedings of the 59th International Astronautical Congress (IAC2008)*, 2008.

[5] N. Smets, J. Bradshaw, J. Diggelen van, C. Jonker, M. Neerincx, L. de Rijk, P. Senster, M. Sierhuis, and J. ten Thije, "Assessing human-agent teams for future space missions," *IEEE Intelligent systems*, vol. 25, no. 5, pp. 46–53, 2010.

[6] A. Jacoff, E. Messina, and J. Evans, "Experiences in Deploying Test Arenas for Autonomous Mobile Robots," in *Proceedings of the 2001 Performance Metrics for Intelligent Systems Workshop*, Mexico City, 2001.

[7] J. W. Streefkerk, M. P. van Esch-Bussemakers, M. A. Neerincx, and R. Looije, "Evaluating Context-Aware Mobile Interfaces for Professionals," in *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*. Information Science Reference, 2008, pp. 759–779.

[8] J. Wang, M. Lewis, S. Hughes, M. Koes, and S. Carpin, "Validating USARsim for use in HRI Research," in *In Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, 2005, pp. 457–461.