Predicting Performance and Situation Awareness of Robot Operators in Complex Situations by Unit Task Tests

Tina Mioch, Nanja J. J. M. Smets, Mark A. Neerincx
TNO
Kampweg 5, 3769 DE Soesterberg, The Netherlands
{*tina.mioch, nanja.smets, mark.neerincx*}*@tno.nl*

***Abstract*—Human-in-the-loop field tests of human-robot operations in high-demand situations provide serious constraints with respect to costs and control. A set of relatively simple unit tasks might be used to do part of the testing and to establish a benchmark for human-robot performance and situation awareness. For an urban search and rescue (*tunnel accident*) scenario, we selected and refined the corresponding unit tasks from a first version of a test battery. First responders (fire-men) conducted these unit tasks with a state-of-the-art robot and, subsequently, had to perform the *tunnel accident* mission in a realistic field setting with the same robot. The *Detect objects* unit task proved to partially predict operator's performance and the operator's collision awareness in the scenario. Individual differences, particularly age, had a major effect on performance and collision awareness in both the unit tasks and scenario.**

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

I. INTRODUCTION

Unmanned Ground Vehicles (UGVs) are intended to be deployed in diverse, high-demanding environments. Human-robot team performance is often critical (e.g., time pressure, high error costs), and dependent on team's skills to cope with the dynamic situational conditions, for example in the urban search and rescue domain. Evaluation of the robots before actual deployment is of utmost importance, but the opportunities to conduct realistic field experiments are constrained due to the limited availability of end-users and test sites. Furthermore, objective evaluation poses fundamental difficulties due to the 'situatedness' of robots' effectiveness and efficiency so that outcomes may be hard to generalize.

In this study, we investigate the applicability and validity of a usage-centered evaluation methodology for unmanned ground vehicles. This evaluation methodology provides standard task assignments and metrics on human-robot collaboration. The idea is that a set of relatively simple and abstract unit tasks can be used to assess basic aspects of this collaboration and to establish a benchmark for human-robot performance and situation awareness. Such tests can decrease the need for evaluating the human-robot performance in the environment in which it will actually be deployed. The assumption is that these tests predict the performance in a realistic scenario for an important part. The application of the proposed test battery with 'unit tasks' should help

- to generalize,
- to standardize (compare results of different tests), and
- to interpret outcomes in terms of the robot's functional components.

For a more detailed motivation and positioning in a usage-centered UGV evaluation and design methodology, see [1]. It should be noted that the emphasis of this research lies on a first evaluation of the applicability and validity of the methodology. Our approach is to instantiate the methodology for one particular research question, namely human-robot collaboration in an urban search and rescue scenario ("tunnel accident"). For each task in the test battery, the interaction of the whole system, meaning one robot together with its operator, is evaluated. The test battery tasks are not intended to do isolated tests of specific robot technologies or performance tests of either the individual robot or individual operator (i.e., the focus is on joint human-robot operation).

The following research questions can be identified:

- Is the performance and situation awareness of the participants in the test battery a good prediction of the performance and situation awareness in the scenario?
- Can the unit tasks help to explain operator performance in complex scenarios?

Individual differences can have a major effect on operational outcomes. To get first insight in such effects, we will analyze whether individual factors such as spatial ability and experience in computer games influence the performance and situation awareness of the operator, and whether these effects are similar for the test battery and scenario setting.

The paper is structured as follows: first, we will describe how this research can be placed in the context of performance evaluation for human-robot cooperation, followed by a description of the method to answer the research question. Subsequently, the results of the experiment are given and discussed.

II. BACKGROUND

In this section, it is described how this research can be placed in the context of performance evaluation for human-robot cooperation.

 ISBN: 978-1-61208-177-9

A. Situated Cognitive Engineering Methodology

To establish the set of functional requirements with the corresponding metrics for evaluation, the situated Cognitive Engineering (sCE) Methodology [2] is applied. Following the sCE methodology, the operational demands, human factors knowledge and technological constraints were analyzed and used to specify design scenarios and a requirements baseline. An example of a requirement is given in Figure 1. The requirements baseline consists of claims that justify the requirements, and use cases that contextualize and organize these requirements.

Requirement 1.1	**The robot should be able to be manually controllable by an operator**		
Claim 1	If the robot needs navigational help (e.g. getting stuck or precise maneuvering), the operator can take manual control over the robot		
	+	The efficiency of the robot and operator cooperation will increase if the operator is capable to take over control for situations in which the operator thinks that manual control is needed	Measuring the time it takes for the operator and the robot to finish a task in which the robot gets stuck
	–	Manually controlling the robot will take up attention of the operator, increasing the experience of the workload	Question(s) about the amount of workload the operator is/was experiencing when teleoperating
Use cases	UC_0004(Main success scenario: step 1, 2, 5)		
Task battery tasks	T1: Stop before collision		Tests manual control of operator over the robot and a good awareness about speed and distance to the object

Figure 1. An example of a requirement, with a claim and the corresponding unit task.

Subsequently, we identified unit tasks in the test battery set, which addressed these requirements. This means that to execute the unit task successfully, the requirement must be met, just as this should be the case for the scenarios implementing the use cases. For each requirement, at least one unit task that manifests this requirement in the scenario was selected.

B. Performance evaluation

Several categories of human-robot cooperation metrics can be distinguished: general metrics, collaboration, and user interfaces. In this paper, we concentrate on the general performance metrics. These include for example efficiency, effectiveness, task load and emotions, and situation awareness. In the following, the predictability for general performance and situation awareness is analyzed.

In general, experiment setups for evaluations can differ in the dimensions *fidelity* and *realism* [3]. *Fidelity* expresses how close the collaborative operations resemble the actual ”rules” of operations and their internal and external dependencies (i.e., the social and environmental dependencies). *Realism* specifies whether the evaluation environment is represented realistically (“Does it look, feel and smell like a disaster?”), for example from low realism in a virtual environment, to a high realism in an earthquake site.

Different experimentation environments have different advantages and disadvantages. Evaluating a robot in a ”real disaster site” for example has high realism and high fidelity, but it is costly. Furthermore, there is lack of controllability and you cannot test all kind of settings without the risk for damage or injuries. Therefore, specific test arenas are being set up, such as NIST, which have different levels of realism [4]. However, fidelity may remain somewhat lower, because the rescue team cannot operate conform their complete set of coordination and collaboration policies.

As a complementary approach, we propose to identify unit tasks that resemble basic functionality of human-robot collaboration in envisioned scenarios. The higher the resemblance, the higher the fidelity. Here, we will focus on the collaboration between two actors, the robot and operator, however, this approach can be extended to more actors. Subsequently, these tasks are applied to test the collaboration in a controlled setting (preferably with the same environmental constraints as the real setting). In this paper, we evaluate whether the human-robot performance in a test battery can predict actual performance in a field test. For a more extensive motivation, overview and placement of the test battery in comparison to other evaluation environments, see [1]. The field test performed in this study has a high realism.

III. METHOD

In this section, the method is described in detail.

A. Task

As described in Section II-A, the unit tasks were selected by requirements matching. The experiment consisted of two parts, namely the test with a selection of tasks from the test battery, and the test with the scenario. The following unit tasks were selected:

- *Detect objects in the environment*. The robot is placed at the entrance of a room. In the room, several warning signs printed on A4 paper can be found. The participant has to find the signs, and situate them on a map, with a time limit of two minutes.
- *Slalom*. The participants have to drive slalom around pylons as fast as possible without touching the pylons.
- *Move through narrow hallway*. The participants have to drive through a narrow hallway as fast as possible without touching the walls.
- *Stop before collision*. At the end of the hallway, participants have to maneuver the robot as close as possible to the wall, without touching the wall.

The second part of the experiment was the execution of the scenario. The scenario was a car accident in a tunnel. The situation in the tunnel was not clear, and more information was needed. There was smoke development in the tunnel.

 ISBN: 978-1-61208-177-9

A robot, controlled by the participants, was deployed to gather information. The participants were asked to answer the following questions:

- Are there cars in the tunnel? If so, where are these?
- How is the layout of the situation?
- Are there victims? And if there are, how many were there, and where?
- Look for fire and dangerous substances, depicted by pictures of warning signs.

While navigating through the scenario area, participants had to indicate on a whiteboard what they saw, by using magnetic icons and whiteboard marker. The magnetic icons were: pallet, truck, warning sign for fire, warning sign for dangerous substance, car, barrel, victim and a cardboard box.

B. Design

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

C. Materials

The following materials were used in the experiment:

- An unmanned ground vehicle, the *Generaal* (see Figure 2), has been custom-made at TNO in Soesterberg and has been used in other studies as well. For a detailed description, see [5]. The vehicle has been specifically designed for telepresence control, with a pan-tilt-roll unit with a camera system mounted on top of it. The telepresence control station consists of a head-tracking head-mounted display (HMD) (see Figure 3), a steering wheel and an accelerator. The head-tracker directs the pan-tilt-roll unit, and the HMD displays the sensor images. This gives the operator the experience of naturally looking around at the remote location. Vehicle control is facilitated by two 'antennas' at the side of the robot. These indicate the width of the vehicle as well as the front of the vehicle.
- Hall with separate area for test battery tasks and scenario.
- For setting up the scenario we used the following items: three cars, one motor, five dummy victims, three barrels and three 'danger' signs.

1) Participants: Nine male participants took part in the experiment as volunteers. All participants were firemen from the fire department of the city Dortmund with an average age of 34. The mean number of years the participants had a driver's license was 18.

D. Measures

The following measures were taken during the execution of the test battery tasks and the scenario:

1) Performance data
   - Time to finish task
   - Number of collisions
2) Situation Awareness
   - Number of correctly identified objects
3) Performance Perception
   - Perceived collisions
4) Personal characteristics

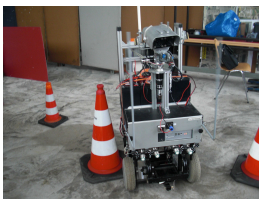

Figure 2. *Generaal* robot of TNO

Figure 3. Head-mounted display interface of Generaal robot

In the following, we will analyze the performance data, the situation awareness, and the operator's perception on the performance to determine, whether for these metrics, the test battery is a predictor for the field test measures. The information gained about the personal characteristics is also analyzed.

E. Procedure

At the beginning of the experiment participants were given a general, written instruction about the experiment. Then a spatial ability test was conducted. Then participants had to fill in a general questionnaire about their background, computer and game experience. An extensive training was conducted with afterwards a learnability questionnaire. Then the participant performed the test battery tasks, with after each test battery task a questionnaire. Then the scenario was performed with a workload questionnaire and map drawing during the scenario, followed by several scenario related questionnaires. The experiment ended with an end questionnaire.

IV. Results

As depicted in Figure 4, we performed several analyses. First, we performed correlation analysis and multiple regression analysis for performance, situation awareness, and performance perception measures for the scenario, with the performance, situation awareness, and performance perception of the unit tasks as predictor variables (arrow A in Figure 4). In addition, multiple linear regression analyses were performed for the unit tasks and the scenario based on the following predictor variables: age, the amount of kilometers the participant drives per year, and the experience with computer gaming (see arrow B and C in Figure 4). We

 ISBN: 978-1-61208-177-9

decided to use age as a predictor variable and not the number of years the participants had their driver's license, because some participants did not fill in the question correctly.

Performance: For both the unit tests and the scenario, as performance measure, we analyzed the number of collisions. The time it took to finish a task was measured for some of the unit tasks, but not for the scenario, as the operators were given 15 minutes to finish the scenario.

Situation awareness: As mentioned above, the operator drew a map of the environment of the scenarios and the test battery tasks. As situation awareness measure, the number of correctly identified objects was analyzed.

Performance perception: To measure the performance perception, we selected the measure of collision awareness, as this measure was most practical in defining and applicable for all test battery tasks. For both the unit tests and the scenario, the awareness of the operator of having collided with an object was measured as the difference between the actual number of collisions and the number of collisions reported by the participant.

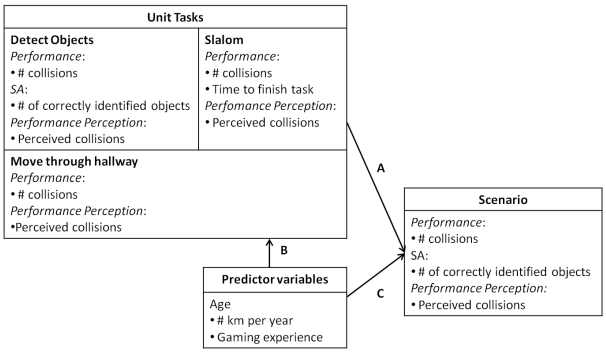

Figure 4. Overview of the analysis

A. Analysis of the predictive power of the unit task performance for the scenario performance

One of the questions we want to answer is in how far the unit task performance can predict the performance in the scenario, see arrow A in Figure 4. We are analyzing this for the performance measures (the number of collisions), the SA measure (the number of correctly identified objects), and the operator's collision awareness.

Performance: We conducted a correlation analysis on the performance measure. There was a positive correlation (trend) for the number of collisions, i.e., when a participant collided more in the test battery task *Detect Objects*, the participant also collided more in the scenario, with r = 0.44, p = 0.063 (for the scatterplot, see Figure 5).

Situation awareness: The correlation for the number of found objects in the *Detect Objects* test battery task and the scenario was not significant. Of the task battery tests, the number of objects found in the *Detect objects* task explains 24 % of the variance in the scenario (see Table I).

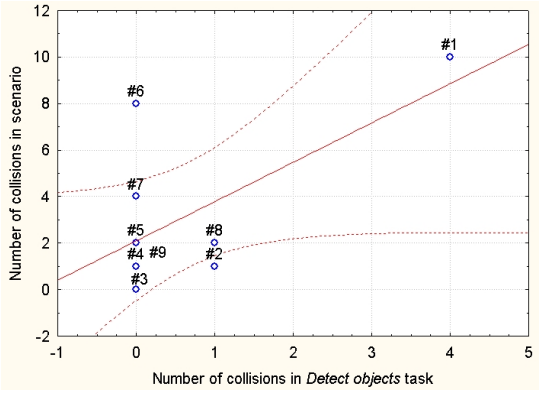

Figure 5. Scatter plot of the performance measure, number of collisions per participant in the *Detect objects* task and the scenario.

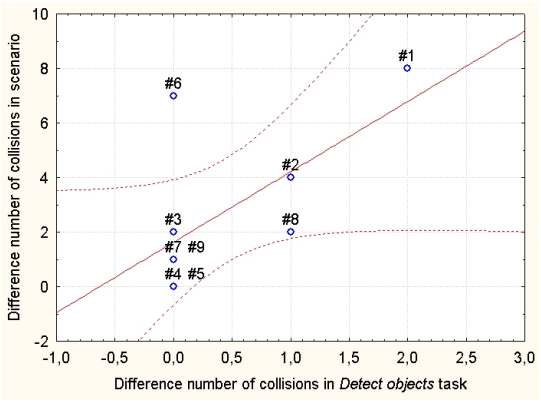

Figure 6. Scatter plot of the performance perception measure, difference between the actual number of collisions and the number of collisions reported per participant in the *Detect objects* task and the scenario.

Performance perception: Correlation analysis showed a positive trend between the operator's collision awareness in the unit task *Detect objects* and the collision awareness in the scenario, r = 0.64, p = 0.066. When there was a larger difference in the actual number of collisions and the number of collisions reported in the *Detect objects* task, this was also the case in the scenario, see Figure 6. When performing a multiple linear regression analysis with the task battery tests, the difference in the number of collisions in the *Detect objects* task explains 40 % of the variance in the scenario (see Table I).

Table I
PERCENTAGE OF EXPLAINED VARIANCE THE UNIT TASK *Detect objects* ADDS FOR THE SCENARIO.

Criterion	**Explained variance R2 (%) by the three predictor variables**
Number of objects found in *scenario*	Number of objects found *Detect objects* = 24%
Difference in number of collisions *scenario*	Difference in number of collisions *Detect objects* = 40%

Table II
PERCENTAGE OF EXPLAINED VARIANCE FOR THE PERFORMANCE AND SA MEASURES THAT THE DIFFERENT PREDICTOR VARIABLES ADD FOR THE UNIT TASKS AND THE SCENARIO.

Criterion	**Explained variance R2 (%) by the three predictor variables**
Number of objects found in *Detect Objects*	Age = 32%
Number of objects found in *scenario*	Age = 74% Add kilometers per year = 86% Add Gaming experience = 89%
Number of collisions in *Detect Objects*	Age = 13%
Number of collisions in *Narrow Hallway*	Kilometers per year = 57% Add age = 72%
Number of collisions in *Slalom*	Age = 59%
Number of collisions in *scenario*	Age = 38%

Table III
PERCENTAGE OF EXPLAINED VARIANCE FOR THE OPERATOR'S COLLISION AWARENESS THAT THE DIFFERENT PREDICTOR VARIABLES ADD FOR THE UNIT TASKS AND THE SCENARIO.

Criterion	**Explained variance R2 (%) by the three predictor variables**
Difference in number of collisions, *Detect Objects*	Age = 20%
Difference in number of collisions, *Slalom*	Kilometers per year = 45% Add gaming experience of collisions,= 60%
Difference in number *Move through narrow hallway*	Kilometers per year = 39% Add age = 52% ;
Difference in number of collisions *scenario*	Age = 37% Add kilometers per year = 60%

B. Effect of individual differences on the unit task performance and scenario

In this section, it is analyzed in how far individual differences effect the performance in the unit tasks and in the scenario (see arrow B and arrow C in Figure 4, respectively.) A multiple linear regression analysis was performed to predict the different measures based on the following predictor variables: age, the amount of kilometers driven per year, and the experience with computer gaming.

Performance and Situation awareness: Table II shows that age explains most of the variance for the test battery and the scenario. In the regression, it explains the largest part of the variance percentage-wise for all performance variables, of which two are significant (for the number of objects found in the scenario and number of collisions in the slalom). In the scenario, the number of kilometers driven per year and gaming experience is also of influence for the number of objects found. The number of collisions in the *narrow hallway* task is influenced by the amount of kilometers driven per year by the participant.

Performance perception: Table III shows that the age of the participants explains the variance percentage-wise for three out of four variables, in the scenario it is significant. Kilometers driven per year also explains the variance for three out of four variables, and is significant in the slalom task. In the slalom task, game experience is of influence as well.

V. DISCUSSION AND CONCLUSION

This study tested a recent method for the evaluation of human-robot collaboration with unit tasks [1]. The *Detect objects* unit task proved to partially predict operator's performance and the operator's collision awareness in the scenario. Individual differences, particularly age, had a major effect on performance and collision awareness in both the unit tasks and scenario.

It should be noted that the *Detect objects* task was the most comprehensive task; both the operational demand of transiting with the robot and observing the environment are included, whereas the other unit tasks are mostly transiting tasks. Hence, the *Detect object* task is the closest of all tasks to the scenario task, in which also transiting and observing the environment. Conversely, if the scenario would have had as main operational demand transiting around the environment, the other unit tasks possibly would have predicted the scenario outcomes better. Our study suggest that, when applying the methodology, the tasks that are used for predicting the performance in the scenario should address the concurrent operational demands.

In addition to the deficient mapping of operational demands on the two "other" unit tasks, effects may have been hidden due to some deficiencies in the amount and property of the data. As in most field studies with real end-users, the number of participants available was limited. In addition, the performance measures of the unit tasks proved not to match perfectly with the scenario measures. For example, the slalom task had two performance measures: the time it took to finish and the number of collisions with the cones. In the scenario, only the number of collisions was relevant, and the time, even though it was limited, was given as a constraint and not as a performance measure. Consequently, the measure *number of collisions* was different in the slalom task compared to the scenario, as the time the task execution took probably influenced the number of collisions. In general, the evaluation measures in the scenario proved to be quite difficult to establish and to incorporate in the unit task measures. Based on the experiences in this test, we will refine the measures in the next tests.

We can further conclude that the unit tasks can be used to explain some operators' performances. As they are specified with a particular challenge in mind, e.g., operational control of the robot, or gaining situation awareness, the reason for a bad or good performance is more easily inferred than when evaluating the scenario performance. For example, because of the *Stop before collision* task, we could determine that the perception of distance was not very good, and that this was the main reason for the number of collisions, instead of difficulty of maneuvring. In general, individual differences, particularly age, proved to have a major effect on performance and situation awareness in both the unit tasks and scenario. Unit tasks show the effects of these

 ISBN: 978-1-61208-177-9

differences and can be of help to see whether higher levels of robot autonomy and advanced situation awareness support can help to decrease problems of some users with current robot control and perception.

A. Observations

An interesting observation concerns the performance of participant 6, who consistently showed a deviation from the performance patterns of the other participants. He performed average on the test battery tasks, but clearly below average in the scenario. His perception of his own performance proved to deviate from his actual performance: he most often did not notice the collisions. Probably, he became somewhat overreliant, overestimated his own capabilities, and, consequently, performed worse in the scenario. Without participant 6, the main results of this experiment showed the same pattern, but the level of significance of the effects proved to increase (i.e., the correlations were significant at $p < 0.5$ without participant 6).

When executing the scenario, several participants believed, after about 12 minutes, that they had explored the whole environment well. After being told that they could go on for some more minutes (the execution time for the scenario was set to 15 minutes), all of them continued. Several of them still found some objects that they had not seen before. This indicates that their situation awareness was less good than they believed it to be.

Some operators complained about the head-mounted display - after some time, it was not comfortable to wear anymore. Most operators liked the situatedness of telepresence, although some complained that they could not see the extensions of the robot, and thus felt could not maneuver well.

B. Future outlook

The results of the evaluation will be used to refine the requirements baseline and the use cases, e.g., the robot needs to be able to notify the operator when having collided with an object. This will eventually lead to a better performance, as the operator will have a better performance perception and can learn from his mistakes.

Furthermore, another evaluation of the methodology will be done, with refined metrics for the unit tasks and scenarios (among other things to improve the comparison), and larger numbers of end-users. In this way the data-set increases to convey systematic correlations between unit tasks and scenario operations, and the effects of individual differences. We do this by

- evaluating whether the test battery is predictive for the performance and situation awareness in a real scenario for another robot (i.e., the NIFTi robot);
- extending the evaluation mentioned above by having more participants execute the test battery tasks and the scenario;
- determining for which aspects of performance and situation awareness, the test battery task results can be used reliably as a standardization measure.

In addition, we will do further research on the general expressiveness of the unit task performances. We will especially look into for which questions the performance evaluation with unit tasks can be used and the advantages that lie in the performance of unit tasks. In particular, we are planning to apply unit test results for

- determining how much and in which way do individual operator differences play a role in the interacting with the robot and the human-robot performance;
- evaluating whether a robot is adequate for executing a particular task;
- determining whether robot-operator cooperation is clearly unsatisfactory, which might lead to either
  - determining whether an operator needs extra training in operating the robot, or
  - determining which components (hardware, software, and interaction possibilities) of a robot need to be improved.

ACKNOWLEDGMENT

We would like to thank the fire fighters of the city of Dortmund, Germany, and of SFO in Italy for their support. This research is supported by the EU FP7 ICT Programme, Project #247870FP7 (NIFTi), and by the Netherlands Defense UGV research program V923.

REFERENCES

[1] J. van Diggelen, R. Looije, T. Mioch, M. A. Neerincx, and N. J. J. M. Smets, "A usage-centered evaluation methodology for unmanned ground vehicles," in *Proceedings of the Fifth International Conference in Computer-Human Interactions (ACHI 2012)*, Valencia, Spain, 2012.

[2] M. A. Neerincx and J. Lindenberg, "Situated cognitive engineering for complex task environments," in *Naturalistic Decision Making and Macrocognition*, J. M. C. Schraagen, L. Militello, T. Ormerod, , and R. Lipshitz, Eds. Aldershot, UK: Ashgate, 2008.

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

[4] 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 (PerMIS) Workshop*, Mexico City, Mexico, 2001.

[5] C. Jansen and J. B. F. van Erp, "Telepresence control of unmanned systems," in *Human-Robot Interactions in Future Military Operations*, M. Barnes and F. Jentsch, Eds. Ashgate Publishing Limited, 2010, pp. 251–270.

 ISBN: 978-1-61208-177-9