Development of an Evaluation Method for Acceptable Usability

Brian Stanton
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD, USA
brian.stanton@nist.gov

Brian Antonishek
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD, USA
brian.antonishek@nist.gov

Jean Scholtz[i]
Pacific Northwest Laboratories
P.O. Box 999
Richland, WA 99352
jean.scholtz@pnl.gov

***Abstract*—**The National Institute of Standards and Technology (NIST) conducted three workshops with First Responders to determine requirements for robots used in Urban Search and Rescue (USAR). These requirements [1] were further prioritized by the responders. NIST has now undertaken the task of developing evaluation methods and metrics for the requirements deemed to be the highest priority. Of these high priority requirements, a number of these addressed the human-system interface and interaction. In this paper, we explain the pilot testing that has led us to the current evaluation design and outline the next steps.

I. Introduction

Accepted measures of usability [2] are effectiveness, efficiency, and user satisfaction. Effectiveness is often defined as the percentage of users that can carry out a particular task. Efficiency is the time it takes users to complete the task. User satisfaction refers to users' ratings on one of a number of standard satisfaction questionnaires.

Acceptable usability was defined as one of the requirements for the human-system interaction in USAR robots. To measure this, it is first necessary to define the tasks that users need to carry out. Ground robots were selected as the focus for defining the first set of evaluations. Once we have defined evaluation methods for ground robots, we will need to define similar evaluations for aerial vehicles, underwater vehicles, and wall climbing vehicles.

II. Definition of Tasks

We selected navigation and camera manipulation as the first two essential tasks that an operator must be able to do using the human-system interface. These tasks are independent of the degree of autonomy that a particular robot has. We are only concerned with testing the interface – not the capabilities of the robot. Our evaluation of the user interface is but one piece of a package of evaluation methods, many of which are measuring the capabilities of the robot. The acceptable usability evaluation is not meant as a comparison between robots but as a comparison of how well a novice user does as compared to an expert user.

III. Acceptable Usability Test

A. Version One of the Acceptable Usability Test

The first version of the test was designed as a slalom course. This is shown in Figure 1. We set up a number of gates, some were wide and some were narrow gates. In addition, the gates differed in the distance they were apart and the offset of the gates. Our thought was that the gate width, distance between gates, and offset of the gates would differ depending on the size and turning radius of the robots.

We added camera manipulation to the evaluation by posting numbers and letters on the various gates and asking the operator to read as many as possible to us before going through the gate.

The measures we collected were: the time it took to navigate through the gates; the number of gates cleanly navigated through; and the number of letters and numbers the operator viewed. The latter was adjusted based on the capabilities of the robot. For example if a robot did not have a pan tilt camera and was low to the floor, then the operator was not expected to be able to see a letter posted on the very top of a cone.

This evaluation would be conducted after novices had completed the vendor suggested training. We piloted this test at a USAR workshop with both novices and experts in August, 2005. We asked both novices and experts to try out this first test. It should be noted that operators were not able to view the course while running the robots through. While we did not have enough users and robots present for any statistical data collection, we did discover that the novices could complete the course in about twice the time it took the experts to do this.

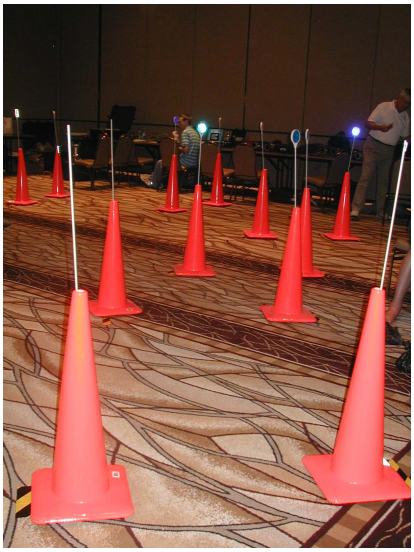

Figure 1. Version one of the acceptable usability test

There were several problems that we noted with the evaluation. One issue was the base on the cones. While the robot operators were able to navigate successfully through the upright portion of the cones, they often did not see the base of the cone and ran over it or moved the cone as they pushed up against the base. In addition it was difficult to place letters and numbers on different portions of the cones for the camera manipulation portion. Moreover, the operators could read some of the numbers and letters if they were far enough away from the cone. So deciding on the "correct" number of letters and numbers depending on a robot's cameras was difficult.

B. Version Two of the Acceptable Usability Test

In March 2006 we tried out a second design for the acceptable usability test. In place of the cones we previously used, we decided to use cardboard boxes to make rows. We did this to give us more flexibility in placing markings to use in the camera manipulation exercise and to prevent operators from being able to view the entire course at one time. We marked gates using red and white hazard tape and told the operators that they were only to go through the marked gates. We tried to place tape at various locations on the boxes so that all robots could see this using their cameras. Figure 2 shows the setup for this version of the test. The cones shown in the photo were used to designate the end of a row as we had to set this up in a large hotel space.

Again, we measured the time it took to navigate the course, the number of correct gates that the operator went through and whether the operator was able to traverse through the gate cleanly.

We did not have enough robots or users for any statistical analysis. One problem with this setup was that it was easy for the robots to move the boxes. We needed to devise a way to easily replace the boxes in the correct position.

Figure 2. Version two of the acceptable usability test

C. Version Three of the Acceptable Usability Test

We had another opportunity to try out a version of the acceptable usability test during a responder workshop at Disaster City, Texas in April, 2006. This time, we decided to add the notion of situation awareness to the evaluation. We devised some gates that the robots would not fit through. We used cardboard boxes to make rows. In each row we had two gates. We marked the gate that the robots were supposed to go through using red and white hazard tape. In some instances these gates were too narrow for that robot. We asked the operator to determine this and if this was the case to traverse to the other gate and go through it.

As in the first version of the test, we measured the time it took to traverse the course (compared to the expert time), the number of gates that were cleanly navigated and the number of gates that the operator recognized as marked.

We had many more robots this time. In fact, due to the logistics of the workshop, we were unable to customize the gate width and the distance between rows to every robot. We decided to design the course for categories of robots, rather than using a percentage of the robot size as the criteria for the size and placement of the gates. Based on the robots taking part in this evaluation, we used two sizes of gates. For robots wider than 19 inches (48.3 cm), we used gates of 2 feet (61 cm) and narrow gates of 18 (45.7 cm) inches. For robots under 19 (48.3 cm) inches, we used gates of 20

inches (50.8 cm) and 14 inches (35.6 cm). In both configurations the rows were 48 inches (121.9 cm) apart.

Figure 3. Rows of boxes for the third version of the acceptable usability test

Figure 4. Markings on gates for third version of acceptable usability test

Figures 3 and 4 show the setup for the third version of the acceptable usability test. We conducted the test in low light. Although we were unable to measure the actual light in this instance, we plan to propose three lighting conditions for the final version of the acceptable usability test: low light with a range specified; complete darkness; and direct sunlight. For the direct sunlight case, it may only be necessary for the operator to be positioned in the sun and for the actual navigation maze to be in the low light condition.

Figure 4 shows how we marked the gates for the robots. We varied the markings so that they were low, medium, and high on the boxes. This gave all robots a good chance of seeing the markings.

Figure 5 shows the actual configuration we used. We did not have enough room for all eight rows of the maze so we made four rows and had the robots traverse it twice; once from the start and then returning through the maze. The second time through we marked different gates.

To ensure that we could easily reposition the boxes should they be moved by a robot, we taped the boxes together and we marked the various configurations on the floor so we could quickly replace boxes as well as change configurations easily. The space available to us for testing was in the theater building in Disaster City. Thus, like an actual theater, the floor was sloped. In this case the slope was 10°.

In addition one problem with the previous tests was that responders did not have a adequate training on the robots. While we did not have access to vendor supplied training, we did give the responders an opportunity to practice in the environment in a slalom course we set up next to the box course. Figure 5 shows the practice course responders used. We allowed the responders to traverse this course several times "eyes on." When they felt comfortable, they traversed the course using the video feed only.

We collected two sets of data: the performance measures for navigating the course and ratings to a questionnaire that responders filled out after they completed the course. We do not have enough data for any statistical significance at this point, but we have enough data to help us refine the tests.

We asked the responders to provide information about their expertise with operating robots in general, and the robot used in the particular evaluation specifically. We also asked them to rate the difficulty of the exercise, how well they felt this evaluation would predict their performance with the user interface, and how well they felt this evaluation measured the user interface. We also asked them to rate specifics about the user interface and interactions with the robot they used for the specific test. We wanted to determine if there was a relationship between this and their actual performance. Table 1 shows the averages of the ratings concerning the evaluation methodology.

Figure 5. Practice course for the third version of the acceptable usability test

TABLE 1. Rating Questions

Question	**Rating (1 is low; 7 is high)**
Difficulty of the task	3.6
Predictor of how well you can perform US&R tasks	4.6
Indicator of the ease of use of the user interface?	5.2

The data from the questionnaire indicates that the responders feel we are on the right track with the evaluation. The ratings for an indicator of ease of use of the user interface are especially encouraging. The responders did not feel that the task was extremely difficult, despite the fact that six of the seven responders had no training on the specific robots other than the practice time we provided. One responder had several hours – most likely this occurred in other exercises earlier in the week. Two responders had no general training or practice with robots. Of the five others, the amount of training/practice ranged from several hours to 3-4 years.

Table 2 shows the performance data from the responders along with their ratings of the particular robot interface used. The times for completion were quite variable between robots. The times for navigating the maze were less variable between users. The times it took the users of R2 to complete the maze were considerably more than the times it took the users of R3 to complete the task. The user ratings for ease of navigation also reflected this. The user who failed to accurately identify one gate as too narrow (R2) had a lower rating for ease of assessing gate width and camera manipulation than did the other user. Again, while we lack enough data for any statistical analysis at this time, we do think the trend is promising.

In addition to the novice performance we also had experts (representatives from the robot company or owners of the robots) drive the robots. This was done to determine if we could place a lower bound on the performance, such as novices should be able to complete the maze in three times the time required by the expert. Table 3 shows the comparison of the expert times and the average of the novice times. We had either two or three novices complete the maze for four different robots.

TABLE 3. Comparison of expert and novice times

Robot	**Expert time**	**Average of novice time**
Robot 1	11:05	5:07
Robot 2	6:14	5:06
Robot 3	4:20	16:04
Robot 4	6:04	15:27

TABLE 2. Data collection

Data	R1-user 1	R1-user 2	R2-user 1	R2-user 2	R3- user 1	R3- user 2	R3-user 3
Time	NA	NA	18 min	11 min	4:47 min	5:28 min	NA*
Number of gates called	NA	NA	6/6	5/6	6/6	6/6	4/4*
Ease of use	5	4	3	5	4.5	5	5
UI info	6	6	5	-	6	5	NA
Navigation	5	4	2	4	5	6	5
Camera manipulation	6	5	7	6	Camera not working – eyes on run		
Assessing Gate Width	5	5	6	4	Camera not working – eyes on run		

*This user did not complete the entire test due to a lack of time.

The results of the comparison of expert and novice times were rather surprising. There are several issues here. First, we need to ensure that the "experts" are really "experts." We will need to develop requirements for operation time or some test of operational skill to ensure that we have true experts. However, we can also infer that Robot 1 and Robot 2 were relatively easy to operate and that Robot 3 and Robot 4 were much less intuitive.

IV. Next Steps

We have not discussed how we will score the evaluation as it combines navigation, camera manipulation and situation awareness (determining which gates the robot will fit through). Our first thoughts were to award one point for each gate that the robot went through cleanly and another point for traversing each row cleanly. This would be a navigation score. The camera manipulation score would be computed based on the number of marked gates correctly identified by the operator. The situation awareness score would be the number of gates that the operator identified correctly as fitting through. In theory there is no reason that any of these tests could not be completed by an autonomous robot, assuming that the robot could automatically identify the hazard tape markings. There is an issue of the use of time in navigating the maze as this includes not only the time to drive, but the time to manipulate the cameras as needed and any time needed to make an assessment of the gate width. However, this is more of a true world measure that just time to navigate. Our biggest concern at this time is the identification of classes of robots by size and steering mechanism. We would like to identify a number of classes and design the test for those classes rather than using a percentage of each robots' dimensions to configure the gates, row width and gate displacement.

Currently, we found no tele-operated robots with any additional features for situation awareness so we can consider eliminating this part of the test as it depends only on the operator's ability to estimate the size of the robot. We are also considering developing a more rigorous camera manipulation test. This would consist of having the robot placed in an appropriately sized box with numbers and letters on the sides, top and bottom of the box. Without moving, the operator would be asked to move the camera(s) to read as many of the symbols as possible as quickly as possible. For each robot, we could calculate the total possible given the cameras. The metric would be computed using the ratio of those correctly read to those possible factoring in the time taken.

We will also vary the lighting conditions both in the maze (low light and completely dark) and the lighting conditions of the operator (direct sunlight with glare and low light). We need to find a method for specifying the exact lighting conditions and duplicating those in another environment.

We plan to run this evaluation with many more robots and with both expert and novice users, preferably who have had some basic training. We will collect the same type of data to determine if the performance measures correlate with the subjective ratings of the user interfaces. We are also considering having experts in human-robot interaction rate the quality of the interfaces and to determine if those ratings have any correlation with the performance measures.

Acknowledgement

The authors would like to thank Elena Messina, Adam Jacoff, Brian Weiss, and Ann Virts for all their help and support. Our thanks also to the Responders who participated in all our pilot evaluations and to the Department of Homeland Security for providing the funding for this research.

References

[1] Prelimiary Report
http://www.isd.mel.nist.gov/US&R_Robot_Standards/
accessed May 18, 2006

[2] ISO 9241 part 11

[i] This research was conducted while Dr. Scholtz was at NIST.