2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Usability Evaluation Framework for Close-Proximity Collaboration With Large Industrial Manipulators

Kasper Hald[1], Matthias Rehm[1]

Abstract— **Our goal is to design a framework for holistic evaluation of human-robot collaboration systems. To this end we utilize several standardized questionnaires administered while participants perform collaborative tasks in robot work cells. We used Standard Usability Scale and the Usability metric for user experience questionnaires to access usability, NASA Task-Load for workload, two questionnaires for human-robot trust as well as the Unified theory of acceptance and use of technology questionnaires. We performed two pilot tests of our framework with human-robot collabotation work cells at two test sites as part of the DrapeBot project. The goal of the project is to enable human-robot collaboration in the process of carbon fiber draping the production of outer parts. After utilizing the evaluation framework at the two test sites we found that the collection of questionnaires were easy to adapt to each work cell and the practical limitation around running the experiments. Both work cells scored high in usability, expected increase of productivity, as well as high trust and low anxiety, but both work cells scored low on expectancy of use for work in the future at their current state of development.**

I. INTRODUCTION

Several fields of production cannot yet be fully automated using robots. This can be due to low production volume not justifying the development cost of the automation or due to the difficulty or variability of the tasks making robot-only production unfeasible. For example, in meat production the variation and soft structure of the meat makes the processing difficult to automate. Our research is done in the context of carbon fiber draping for production of outer parts, where the production volume and complexity of certain parts do not justify the development for automated robot production lines[2].

In cases like these we can benefit from human-robot collaboration (HRC), using the power of the robot to relieve strenuous and repetitive motions while benefiting from the experience and fine motor skills of the human worker. In order to enable this collaboration, we need a holistic method of evaluation HRC systems. Holistic in the sense that it must evaluate not only single aspects like usability, but focus on requirements in terms of usability, work-related effort, human trust in the robot as well as the acceptance and potential use of the system in everyday work. In addition, we must consider the time required to administer the test, considering the cost and scheduling required to recruit representative users for an experiment. This means we must weight the importance of each measurement taken, dependant on the number on the conditions we are comparing and what we want to know about the robot system as a whole.

In this paper we describe our framework for evaluating HRC cells, utilizing several questionnaires relevant to the worker's perception of the robot and its utility. We present two HRC work cells developed for the DrapeBot project, focusing on collaborative carbon fiber draping on complex moulds. We report our pilot tests of the evaluation framework on each work cell, testing each with 20 participants and evaluating different natural user interfaces (NUI) for communicating with the system.

II. BACKGROUND

Usability and task effort have previously been explored within HRC. A common method is having participants perform tasks with a collaborative robot or part of the system followed by administering questionnaires, such as the Standard Usability Scale (SUS) [1][2], the NASA Task-Load Index (TLX) [3][4] or both [5][6]. While these questionnaires are not developed for HRC, specifically, the questions are widely applicable and valid for robot systems with human operators. Alternatively to the SUS questionnaire, others have used the Usefulness, Satisfaction, and Ease of use (USE) questionnaire [7] for assessment of robots, while this questionnaire is significantly longer than the SUS questionnaire at 30 items as opposed to ten. For objective measurements specific to particular HRC system, it is also common to measure performance metrics, such as task completion time [3], and analyse it in combination with usability questionnaires.

In the DrapeBot project, we are working in the context of large production robots in close-proximity collaboration with human workers. Trust towards the robot plays thus an inherent role regarding safety and subjective evaluation of the collaboration. While trust in robots can be defined and accessed in many ways, we use the definition that trust is influenced by the operator's perception of the robot's reliability, consistency of behaviour [8][9], as well as their perception of their own safety around the robot [10][2]. A commonly used tool for trust assessment is the Trust Perception Scale-HRI by Schaefer [11], either the full 40-item questionnaire [12] or the shorter 14-item version for repetitive tasks [4]. Alternative scales are the Negative Attitude towards Robots (NARS)[13] and the Multi-Dimensional-Measure of Trust (MDMT) [14], which consists of four five-question components regarding robot capability, ethics, sincerity and reliability. These elements are less relevant to industrial manipulators.

[1]All author are with the Faculty of IT and Design, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark kh@create.aau.dk, matthias@create.aau.dk

[2]https://drapebot.eu

III. USABILITY FRAMEWORK

For a holistic evaluation of a HRC system we use multiple surveys to access the participants' perception of and attitude towards the robot and the system. The four main elements of our assessment are usability, work-related effort, expected use and human-robot trust. With the questionnaires we have chosen we cannot only gain a holistic impression the users' perception of a HRC work cell, but the overlap on some of the questionnaires also allow us to verify answers.

A. Usability

Standard Usability Scale (SUS): The SUS [15] is a simple ten-item questionnaire that evaluates the overall usability of a system. It measures the usability of a system against a benchmark data base and thus provides a quick first insight into the usability of the system.

Usability metric for user experience (UMUX): Like the SUS, the UMUX aims at measuring the overall usability of a system, but focuses specifically on the user experience as the determining dimension [16]. Thus, it provides an addition to the result of the SUS and allows to derive a more detailed assessment of the system's usability. The UMUX consists of 4 items that are measured with 7 point Likert scales.

B. Work-Related Effort

The **NASA Task-Load index (NASA-TLX [17])** is used to evaluate the different types of demands involved in a task, both in terms of degrees and importance. The demands evaluated are mental, physical and temporal demands as well as performance, effort and frustration levels. For each distinct task being evaluated its demands are tested in two steps. First, the participant rates the task on six 20-step scales according to the six different categories of demands mentioned. Next, every command is compared pairwise, a total of 15 comparisons, where the participant states which of the two was the most influential on their experience. The number of times a demand was chosen in a comparison determines its weight. The rating of each demand is multiplied by its weight for the demand score and the total workload score is the sum of the demand scores divided by 15, that being the sum of the weights.

C. Expected Use

The DrapeBot project develops innovative and so far unknown types of interaction with large industrial manipulators. Thus, it is important to assess if the potential users (and stakeholders) would be open to using the technology. The **unified theory of acceptance and use of technology (UTAUT)** is a suitable tool for this assessment (e.g. [18], [19]). The different measurement constructs include performance expectancy (PE, 4 items), effort expectancy (EE, 4 items), social influence (SI, 3 items), facilitating conditions (FC, 8 items), attitude towards using technology (ATT, 4 items), behavioral intention (BI, 7 items), self-efficacy (SE, 4 items), and anxiety (ANX, 4 items). Items are measured on 7 point Likert scales where 1 denotes negative (fully disagree) and 7 positive (fully agree) answers.

D. Human-Robot Trust

Trust Perception Scale-HRI: We can assess the operator's trust towards the robot using post-interaction questionnaires. Schaefer [11] developed a human-robot trust (HRT) scale, where the final score is measured based on a series of questions regarding the operator's perception of the robot. These questions pertain to the robot's capability, predictability, expected error rates and more. The full questionnaire consists of 40 questions, but we will use the shortened version consisting of 14 questions for repetitive tasks. The final trust score is on a scale between 0 and 100, where the operator's agreements to statements suggesting trust in the robot are weighted positively, and statements suggesting expectations of errors are weighted negatively.

Trust in Industrial HRI: Whereas Schaefer's scale is a general instrument for measuring trust in HRI, and can be seen as a standard measurement instrument, the scale developed by Charalambous, Fletcher, and Webb [20] concentrates on contexts with large industrial manipulators and collaborative robots in industrial settings. Thus, it is in principle a good match for the DrapeBot scenario but it has not been used extensively until now. It consists of ten statements that are each rated on a five point Likert scale. The 10 statements consist of three major components that are used in the analysis to calculate a single trust score: the robot's motion and pick-up speed (two questions), safety of the collaboration (four questions), and reliability of robot and gripper (four questions). The trust scores of Schaefer and Charalambous questionnaires are referred to as general trust and industrial trust, respectively, for the rest of the paper.

IV. DRAPEBOT SCENARIOS

The goal of the DrapeBot project is enabling HRC in the context of carbon fiber draping. Where normally a cut piece of carbon fiber has to be placed on the mould by one or more workers before it is draped to follow the shape of the mould, the goal of the project is enabling collaborative transport of the cut pieces between one worker and a robot. The robot holds onto the cut piece using and array of adjustable suction units, allowing it to roughly conform to the mould, holding it down while the worker does the fine draping. We tested the evaluation framework in collaborative draping tasks in two HRC cells in development at two sites while comparing NUI for interacting with the robot system. The HRC work cells are hosted by project partners Profactor and DLR. At both sites all participant started with a session of using a non-NUI method for interaction with the robot system, i.e. a button or pedal. This was to allow the participants to get a first impression of the system with least likely errors, before comparing it to the NUI interaction methods. All participants were recruited from the production or research staff at the hosting companies, and no participants were involved in the DrapeBot project.

A. Profactor

At the Profactor work cell the draping task was simulated with reusable cut pieces. The participant would occupy a

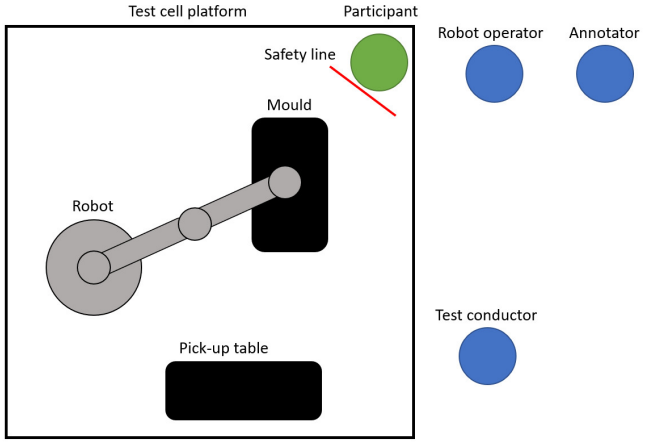

Fig. 1. To-down layout of the simulated collaborative draping task performed at Profactor. The test cell is on a platform with the robot, the mould, the participant and the pick-up table. The robot retrieves the cut pieces from the pick-up table, then brings and holds them down on the mould. New cut pieces are placed on the pick-up table and the old cut pieces are removed from the mould by the test conductor as the tasks proceed. The test is also observed by a robot operator and an annotator. The robot operator always has the emergency stop button within reach. The robot operator always has the emergency stop button within reach.

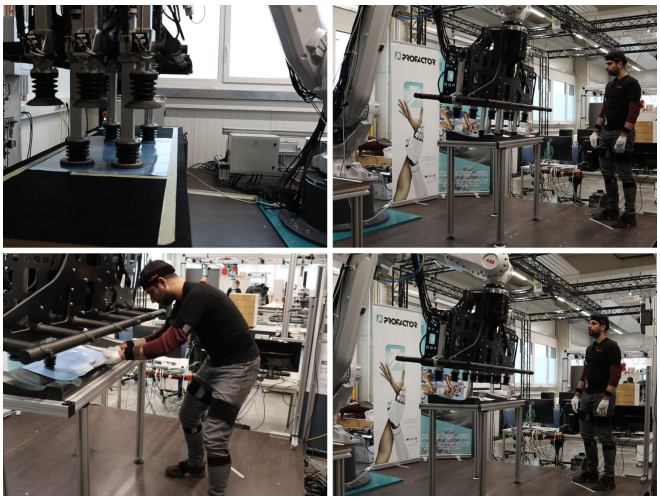

Fig. 2. The collaborative draping task at the Profactor test site. Top-left: The robot picks up the cut piece from the pick-up table. Top-right: The robot places the cut piece on the mould. Bottom-left: The participant approaches the mould and performs the simulated draping task. Bottom-right: The participant moves away from the mould to the safe area and steps on the pedal to signal for the next cut piece.

safe zone outside the reach of the robot, when the robot was in motion. One task repetition consisted of the robot retrieving a large and narrow cut piece from a table and placing it on the mould, holding it at the seeding point. The participant would then approach the robot and perform the draping motions. The draping paths were marked on the cut pieces with erasable whiteboard marker. The participant was tasked to erase the markings using gloved hands. When the participant had finished the draping motions, they retreated to the safe area of the HRC cell and signaled to proceed to the next repetition, and the robot would retrieve the next cut piece. The cut pieces and markings were prepared by a test conductor between tasks.

At this test site each participant performed the task repeatably in two sessions where we compared two methods of signaling the robot. In the first session they repeated the task ten times, communicating to the robot to retrieve the next cut piece by stepping on a pedal in the safe area of the work cell. In the second session the participants performed five task repetitions, signaling the robot using a gesture input by reaching up towards the robot. The motion recognition system required the participant to hold the reaching pose for 2.5 seconds before recognizing the command and getting the next cut piece. To provide feedback the gripper was equipped with three lights, one red, one yellow and one green. When the green light was on the robot would remain stationary, meaning it was safe to leave the safe zone and approach the robot for draping. When the participants performed the reaching gesture and it was recognized by the system, the yellow light would turn on. After holding the reaching pose for 2.5 seconds the robot would to get the next cut piece and the red light would turn on, meaning it was not safe to approach the robot. When the new cut piece had been retrieved and the robot had come to a stop, the green light would turn on again. The scenario is shown in Figure 2 and a top-down illustration is shown in Figure 1.

The usability questionnaires, SUS and UMUX, as well as trust questionnaires were administered after each session, after both gesture and non-NUI conditions. The NASA TLX and UTAUT were only administered once after both sessions were concluded due to the time required to complete them. Participants were told to answer these questionnaires based on the experience with the work cell as a whole.

B. DLR

At the DLR work cell one task repetition consisted of the robot retrieving a 30x30 cm cut piece from a table and placing it on the mould, holding it at the seeding point. The participant would then approach the robot from the safe zone and drape the cut piece on the mould, deforming it. When the participant had finished the draping, they retreated to the safe zone and signaled to proceed to the next repetition, and the robot retrieved the next cut piece. With 10 repetitions each piece was positioned and draped at different positions along the mould, starting at one end and evenly spread along the length of the mould.

Participants repeated the draping task throughout three sessions. In the first session they did the draping task ten times and signaled the robot by pressing a button mounted to their hip. In the second and third session they signaled the robot five times using one of two NUI in counter-balanced order. Due to time constraints and the robot's movement speed, in these sessions the robot did not retrieve a cut piece and the participants did not have to approach it to do any draping. Rather, the robot would perform a short motion, moving the gripper from side to side, to signal that the command was accepted. In one condition the participants would raise their

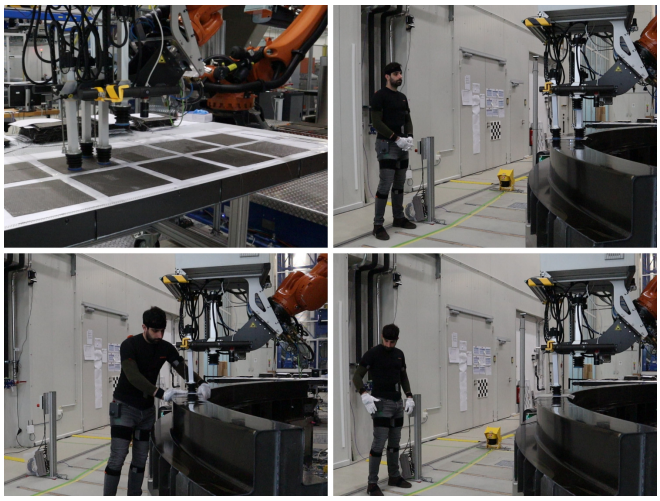

Fig. 3. The collaborative draping task at DLR. Top-left: The robot picks up the cut piece from the pick-up table. Top-right: The robot places the cut piece on the mould. Bottom-left: The participant approaches the mould and performs the draping task. Bottom-right: The participant moves away from the mould to the safe area and presses the button mounted on their hip to signal for the next cut piece.

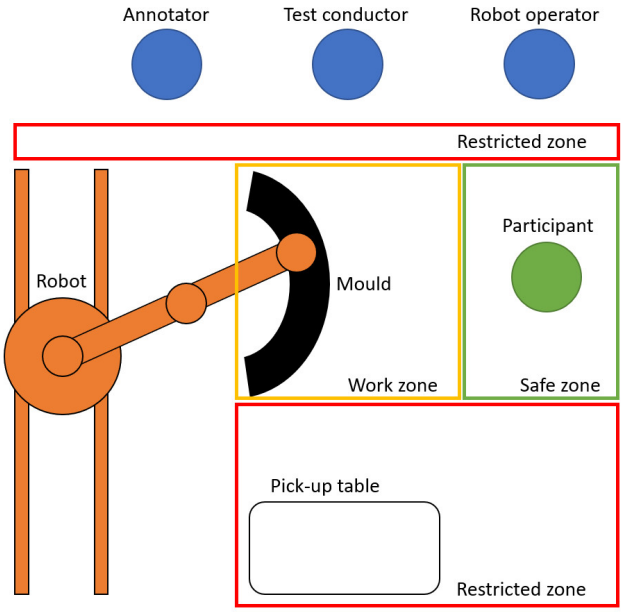

Fig. 4. Top-down layout of the collaborative draping task performed at DLR. While the robot retrieves cut pieces, the participant waits in the safe zone until the robot comes to a stop, placing the cut pieces along the length of the mould. The robot moves along rails to pick up the ten cut pieces laid out on the pick-up table. If the participant is detected in the shared work zone while the robot is moving at a speed over the safety threshold, the robot stops moving. If anyone crosses the laser fences at the edges of the work cell or are detected in the restricted zone, the robot also stops. The test is observed by the test conductor, the robot operator and the data annotator. The robot operator always has the emergency stop button within reach.

arm to signal the robot. In the other condition they would wear a headset with microphone and signal the robot using voice commands. Before giving commands the participants had to say the code word ”Porcupine” for the system to get ready. The participants then had to say ”Continue draping”. The scenario is shown in Figure 3 and a top-down illustration is shown in Figure 4.

As with the Profactor work cell the usability questionnaires were administered after each session, and the NASA TLX and UTAUT were administered once after all sessions were concluded. Due to the time constraints, and because participants did not move close to the robot, the trust questionnaires were not administered after the NUI sessions.

V. Test Results

In this section we are presenting the results of the evaluation of the HRC cells. The two HRC cells are evaluated separately due to the differences between them (see previous section). The participant gender distribution at Profactor was 16 males and 4 females and the ages ranged from 20 to 56 with median age 31. The distribution at DLR was 17 males and 3 females and ages ranged from 21 to 57 with median age 36. At Profactor 18 participants stated they had previous experience working with industrial robots. At DLR this was 13 participants. The distributions of the SUS and UMUX scores for each UI case at each test site is shown Figure 5. While the median SUS scores vary between the test condition, they are all above 68, which is considered above average. Similarly, the UMUX scores, average above 68 for all test conditions. While the UMUX does not have an established benchmark for what is average usability, we consider this an acceptable score.

The trust scores from both trust questionnaires are shown in Figure 6 with similar median scores between test sites and UI conditions. All conditions scored high average trust from both the general trust questionnaire and the questionnaire on trust in industrial manipulators. The median scores for industrial robot trust are slightly higher than the general trust scores, which is likely due to the questions regarding communication and feedback from the robot posed in the general HRI trust questionnaire. For both cells median trust levels are above 80%, showing high trust in the systems. This is a difference to a previous study in the same context [21] that showed median trust levels of about 50% for naive users, which indicates a strong influence of expertise and prior experience on trust levels.

The distribution of effort scores from the NASA TLX surveys are shown in Figure 7. The median effort scores are very low for most categories for both robot systems. There is, however, a high spread of scores for performance effort, meaning that some participants suspect poor performance from the system. Some participants scored high physical demand for the system at Profactor. This is likely due to the need to bend forward and reach underneath the gripper to drape the cut pieces where this was not needed at DLR.

The distributions of the UTAUT scores are shown in Figure 8. Looking at the scores of the individual qualities assessed, some have similar median scores between the test sites. For both test sites participants have differing opinions regarding the performance and productivity of the system, which could be due to the slow movement speed of the robot

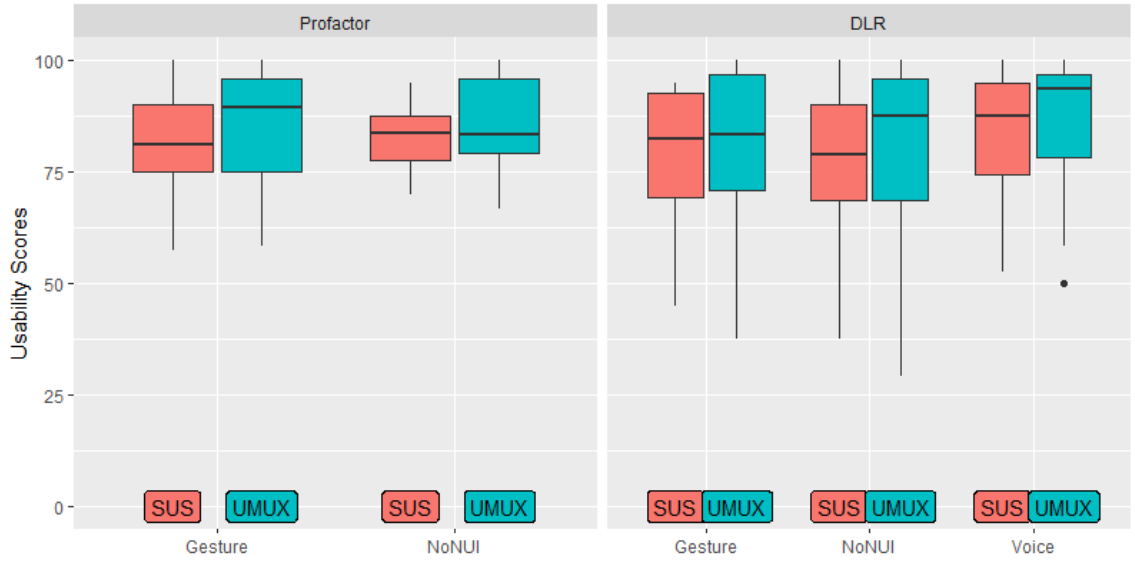

Fig. 5. The distributions of the SUS and UMUX scores between test sites and UI conditions.

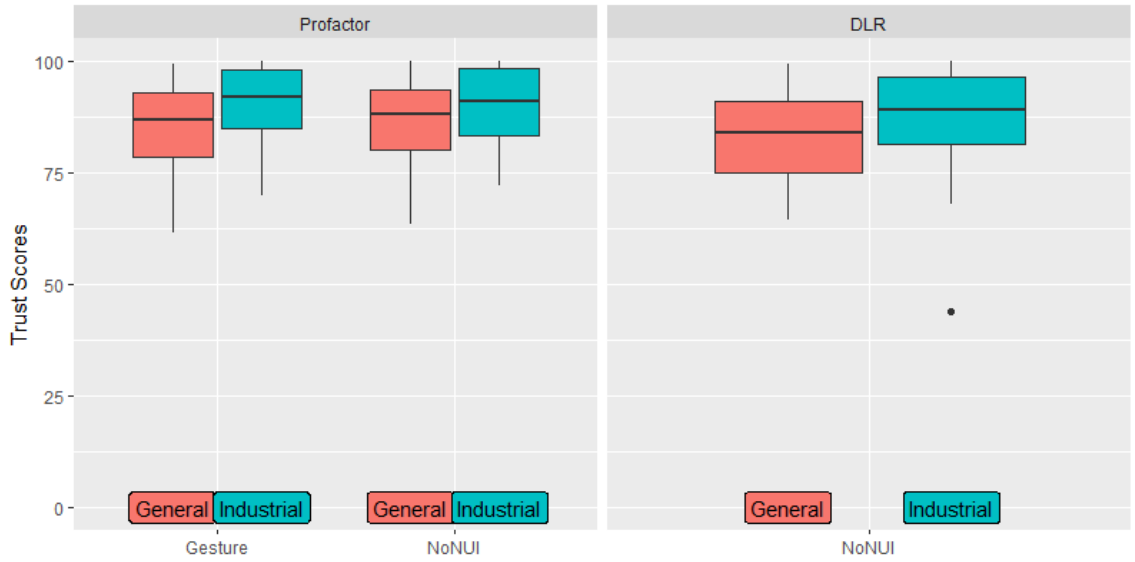

Fig. 6. The distributions of general and industrial robot trust scores between test sites and UI conditions. At Profactor the trust surveys were administered after both the non-NUI condition and the gesture signaling condition. Due to time constraints at the DLR the trust surveys were only administered after the first condition where the robot performed the full retrieval task.

at the current state of the system. This is consistent with the performance expectations from the NASA TLX surveys. Despite this, for both systems we see high median scores regarding low effort of use and positive attitudes towards the systems. Both systems scored means around 50 on social influence, which is understandable for an experimental setting, as there is no social incentive for them to use the systems at this stage of development. Both systems have median scores around 75 for participants' feeling of self-efficacy; their sense that they would be able to use the system on their own. Most participants stated low anxiety around the robot systems, but some scored anxiety around 50 or higher for both systems. The overall low anxiety scores are consistent with the overall high trust scores. Intention to use the system in the future was scored low for the system at Profactor and neutral at DLR with high spreads of scores for both. This can be due to the unfinished state of the systems, meaning that they do not find them suitable for use right now.

VI. Conclusion

We outlined a framework for holistic evaluation of HRC systems, consisting of several standardized questionnaires administered while participants perform collaborative tasks with the robot. We used SUS and UMUX questionnaires to access usability, NASA-TLX to access workload, two questionnaires for human-robot trust as well as the UTAUT

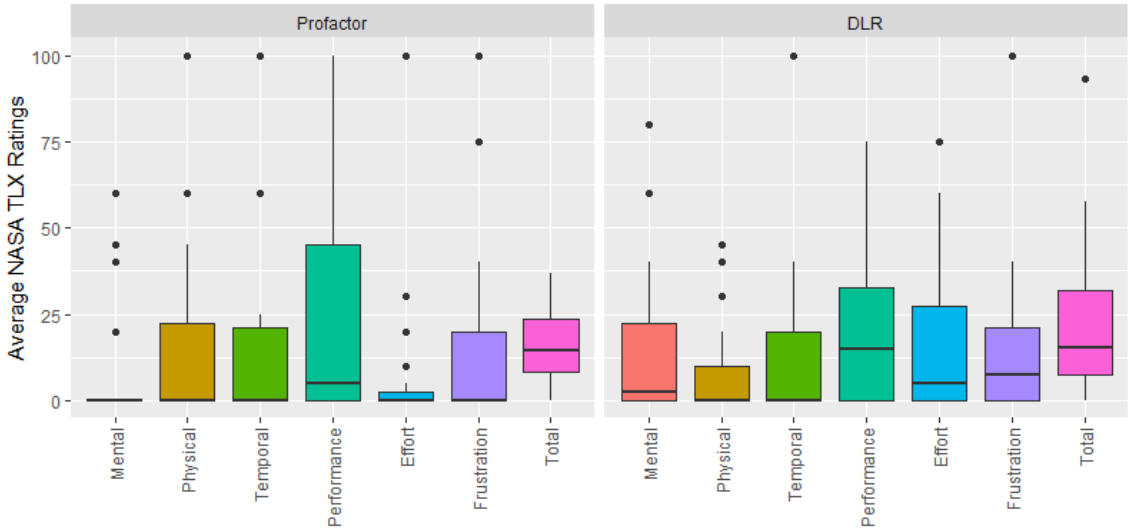

Fig. 7. The distribution of task load scores of the NASA-TLX between test sites.

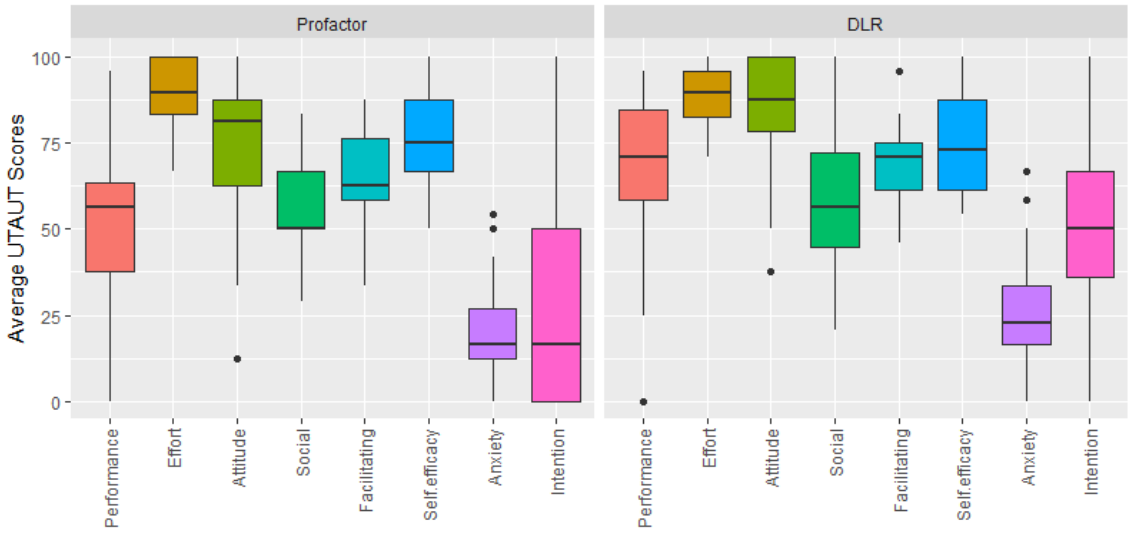

Fig. 8. The distributions scores of the UTAUT questionnaires between test sites.

questionnaires, which overlaps with some of the other questionnaires in addition to accessing intention of future use. This allows us to compare some of the results between questionnaires to ensure consistency of responses, such as comparing the anxiety portion of UTAUT with the trust questionnaires. The questionnaires within the framework were easy to utilize and adapt to each work cell and the practical limitation around running the experiments.

While the data obtained allows to assess the HRC cells in detail, it would be beneficial to accompany the quantitative methods used here with additional qualitative information, e.g. with observations of the work flow and exit interviews to collect additional information for the interpretation of for instance the UTAUT data.

We used the framework for accessing two HRC work cell in development as part of the DrapeBot project. While the work cells and the experimental procedures between them were too different for comparing them directly, the results obtained show similar trends. Both work cells scored high in usability, expected increase of productivity, as well as high trust and low anxiety. Still, both work cells scored low on expectancy of use for work in the future at their current state of development. This is to be expected, as this was an early test demonstrating a single robot action.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101006732 (DrapeBot).

REFERENCES

[1] A. Amaya, D. D. Arachchige, J. Grey, and I. S. Godage, "Evaluation of human-robot teleoperation interfaces for soft robotic manipulators," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 412–417.

[2] G. Tsamis, G. Chantziaras, D. Giakoumis, I. Kostavelis, A. Kargakos, A. Tsakiris, and D. Tzovaras, "Intuitive and safe interaction in multi-user human robot collaboration environments through augmented reality displays," in *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*. IEEE, 2021, pp. 520–526.

[3] H.-S. Song, J. Woo, J. Y. Won, and B.-J. Yi, "Usability test of master devices for robotic vascular intervention procedure," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 914–919.

[4] M. Adamik, A. P. Madsen, and M. Rehm, "Explainability in collaborative robotics: The effect of informing the user on task performance and trust," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1252–1257.

[5] S. Radmard, A. J. Moon, and E. A. Croft, "Interface design and usability analysis for a robotic telepresence platform," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 511–516.

[6] C. Henrichs, F. Zhao, and B. Mutlu, "Designing interface aids to assist collaborative robot operators in attention management," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 264–271.

[7] E. Saad, J. Broekens, and M. A. Neerincx, "An iterative interaction-design method for multi-modal robot communication," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 690–697.

[8] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011.

[9] N. Abe, D. Rye, and L. Loke, "A microsociological approach to understanding the boundary between robot cooperativeness and uncooperativeness in human-robot collaboration," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1085–1092.

[10] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[11] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

[12] R. Savery, R. Rose, and G. Weinberg, "Establishing human-robot trust through music-driven robotic emotion prosody and gesture," in *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[13] M. Romeo, P. E. McKenna, D. A. Robb, G. Rajendran, B. Nesset, A. Cangelosi, and H. Hastie, "Exploring theory of mind for human-robot collaboration," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 461–468.

[14] D. Ullman and B. F. Malle, "What does it mean to trust a robot? steps toward a multidimensional measure of trust," in *Companion of the 2018 acm/ieee international conference on human-robot interaction*, 2018, pp. 263–264.

[15] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[16] K. Finstad, "The usability metric for user experience," *Interacting with computers*, vol. 22, no. 5, pp. 323–327, 2010.

[17] S. G. Hart, "Nasa task load index (tlx)," 1986.

[18] M. Lescevica, E. Ginters, and R. Mazza, "Unified theory of acceptance and use of technology (utaut) for market analysis of fp7 choreos products," *Procedia Computer Science*, vol. 26, pp. 51–68, 2013.

[19] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[20] G. Charalambous, S. Fletcher, and P. Webb, "The development of a scale to evaluate trust in industrial human-robot collaboration," *International Journal of Social Robotics*, vol. 8, pp. 193–209, 2016.

[21] K. Hald and M. Rehm, "Determining movement measures for trust assessment in human-robot collaboration using imu-based motion tracking," in *Proceedings of ROMAN*, 2023.