Initial Face Validity Study of a Tool Measuring User Experience in Human-Robot Collaboration

Bachelor Thesis: Human Factors and Engineering Psychology

Rufaro Hoto

s2540320

Supervisors

1st supervisor: Dr. Simone Borsci

2nd Supervisor: Dr. Cesco Willemse

January 26, 2024

University of Twente

BMS Faculty

Department of Psychology

Abstract

The evolution of robotics and AI, driven by substantial global investments, has significantly transformed human-machine interactions, integrating robotics into diverse sectors of daily life and industry. This technological advancement reflects a growing trend towards automating and enhancing various aspects of modern society. However, as these investments grow, and technologies become more involved with people it's necessary to investigate how to measure subjective experiences with robots. This study presents the development and validation of a scale for Human-Robot Collaboration (HRC), focusing on the user experience (UX) with collaborative robots (cobots). Initiated by the identification of 15 dimensions of HRC by Borsci et al. (2024), this research aimed to refine and validate these dimensions through two primary studies. The first study (Study 1: Consensus Study) involved a Delphi consultation which had 21 experts to achieve consensus on five previously contentious HRC dimensions. The analysis revealed a lack of consensus on their removal, leading to their retention as optional dimensions in the scale, highlighting their context-specific relevance in different HRC scenarios. The second study (Study 2: Item Generation and Card Sorting for Face Validity) focused on developing specific items for each dimension and assessing their face validity through card sorting with 43 participants. 71 items were developed across the 15 dimensions. This initial scale underwent face validity testing which resulted in a refined 38-item scale that captures diverse aspects of HRC. The study contributes significantly to HRC research by providing a comprehensive and validated tool for assessing the UX with cobots. It bridges the gap between technical performance and user-centred factors in HRC.

Table of Contents

1. Introduction

1.1. Human Robot Collaboration

In recent years, the field of robotics and artificial intelligence (AI) has undergone a transformative evolution, reshaping the dynamics of human-machine interactions (Boyd & Holton, 2017). This revolution has been primarily propelled by key players in the AI sector. According to a Stanford University Human-Centered Artificial Intelligence Institute report from 2021, the United States has emerged as the global hub for private AI-focused funding, with a staggering investment of over $23 billion in 2020—more than double China's amount spent during the same period (Stanford HAI, 2021). This significant financial commitment reflects a dedication to advancing AI and robotic technologies. The impact of these investments is evident in the development of robotics, which has become an integral part of our daily lives. Robots are increasingly being employed across various sectors, such as industry and social domains, showcasing their versatility and adaptability (Davenport, 2018).

In everyday life, we encounter robotics in diverse applications, from automated vacuum cleaners and smart home devices to robotic assistants in healthcare. Industries leverage robots for tasks ranging from manufacturing and logistics to hazardous jobs in environments unsafe for humans. The integration of robotics into social settings is exponential, which leads to collaborative scenarios between humans and robots.

Humans' collaboration with robots in society has gained traction, which has implications for various aspects of our lives (Zhang et al., 2023). This collaboration has happened in different ways, including assistance in household chores, companionship for the elderly, and even creative partnerships in the arts. As we continue to integrate robots into our daily routines, the concept of Human-Robot Collaboration (HRC) becomes more important (Papetti et al., 2022). As such it is imperative to develop systems that measure HRC well.

The development of successful HRC systems requires an approach that incorporates both objective and subjective evaluation metrics. Objective metrics, such as task completion time and accuracy, provide essential data on the system's performance. However, Borsci et

al. (2024) highlighted the equal importance of subjective evaluations, that investigate human experiences and reactions post-interaction with robots. These evaluations, often conducted through standardized tools or scales, offer insights into user satisfaction and the ergonomic comfort of the HRC systems. By tracking these subjective reactions, designers can gain a holistic understanding of the system's impact, ensuring that HRC systems are not only efficient but also resonate with human needs and preferences (Coronado et al., 2022). This integrated approach, which aligns with established usability standards like ISO 9241-11, is critical for the iterative improvement of HRC, fostering systems that are both effective and user-centric (Cheng et al., 2022).

Building on this, User Experience (UX), as defined in ISO 9241-210, encompasses the entirety of users' interactions with robotic systems. It focuses on their perceptions and responses resulting from both the use and anticipated use of these systems. UX extends beyond the traditional bounds of usability—a fundamental component outlined in ISO 9241-11, which emphasizes efficiency, effectiveness, and satisfaction in a specific context of use. It goes into the experiential, affective, and meaningful aspects of these interactions, evaluating how comfortable, engaging, and fulfilling they are for users (Cheng et al., 2022). UX therefore assesses whether the robotic systems meet or exceed the users' needs and expectations. This comprehensive approach to UX, grounded in both ISO standards, ensures that the design and development of robotic systems are not only focused on functional aspects but also on creating a holistic and positive experience for the user (Hartson & Partha Pyla, 2012). This emphasis on a positive UX underlines the importance of designing robotic systems that meet user needs and expectations, creating a holistic and engaging experience (Gervasi et al., 2022).

However, the task of assessing UX in HRC presents its unique set of challenges, as detailed by the insights of Borsci et al. (2024), where he notes that HRC systems should be

tailored to human preferences and designed to minimize complex cognitive demands, such as attention and/or memory usage. This approach underscores the necessity to evaluate how well these systems fulfill users' needs for autonomy, competence, and relatedness, ultimately advocating for a shift towards "humanity-centered design" in technology (Norman, 2023). This holistic view of UX in HRC is inherently complex, primarily due to the lack of standardized approaches for assessing UX in this context. Traditional usability evaluations, with their established metrics, do not fully capture the emotional and cognitive dimensions of HRC interactions (Coronado et al., 2022). The subjective nature of these aspects, compounded by the influence of cultural factors and individual preferences, makes developing a universal assessment framework challenging. This complexity highlights the need for innovative and adaptable frameworks to evaluate UX in HRC, ensuring that these systems are not only technically proficient but also deeply attuned to the nuanced human experience (Borsci et al., 2024). In this context, the role of subjective evaluation tools becomes crucial to bridging the gap between technical proficiency and the complex human aspects of HRC.

1.2. Creation of tools in the evaluation of HRC systems.

Subjective evaluation tools play a pivotal role in addressing the complex challenges of assessing UX in HRC. While objective metrics are valuable for quantifying specific aspects of performance, they may not capture the full spectrum of user perceptions and experiences. Subjective evaluation tools provide a means to dive into the emotional and psychological dimensions of HRC, offering insights that objective metrics alone cannot achieve (Borsci et al., 2023). As such the relationship between objective and subjective metrics in assessing HRC is inseparable. These tools, therefore, help in highlighting the emotional and psychological facets that are essential yet often overlooked in objective assessments.

Borsci et al. (2024) presented research that emphasized the need for a tool to evaluate HRC, particularly focusing on the subjective experience of users. This need arose from the inherent diversity and complexity of collaborative robots (cobots), which, despite being designed for similar tasks, could vary substantially in features across different domains. Borsci et al. (2024) noticed that a key to the success of these systems lie in their ability to satisfy human users, thus highlighting the importance of understanding and evaluating the human side of the interaction.

Borsci et al.'s (2024) in their research, proposed a five-step approach for creating a tool to evaluate HRC. The process involves defining the framework (creating dimensions), generating items based on specified dimensions, reviewing them for content validity, evaluating through factor analysis with user interactions, and finally revising the validated scale by eliminating underperforming items.

In the research they covered the first step, defining a framework, and their research led to the identification of 15 essential dimensions, representing a diverse range of aspects related to the interaction between humans and cobots. An overview of the dimensions under these groups can be seen in table 1.

Table 1

The 15 dimensions along with their descriptions, representing a diverse range of aspects related to the interaction between humans and cobots.

Dimension name	Dimension description
1. Easiness of robot regulation.	The easiness of the robot's physical regulation (e.g., robot's components positioning).
2. Robot physical appearance.	How the physical features of the robot can affect the user's judgment. In particular, the dimension considers aspects such as e.g., Level of Anthropomorphism (e.g., Machinelike, Humanlike), Dimension of the robot (i.e., High, Width, Length, Weight), Type of robot (e.g., robotic arm, humanoid robot), Form and Material, Perceived robustness.

3.	Robot's emotional appearance.	How the robot's physical and behavioral characteristics delineate the "robot's emotional profile" and how it can affect the user's judgment. It considers e.g., Robot's Likeability (e.g., happy, kind), Warmth (e.g., social, friendly), Disturbance (e.g., creepy, scary), Discomfort (e.g., awkward, dangerous), Attractiveness.
4.	Robot's competence features.	The user's judgment of the robot's competencies (e.g., reliability, responsiveness) and perceived intelligence (e.g., knowledgeable, responsible) based on its behavior during the interaction.
5.	Robot's physical behavior.	The user's judgment of the robot's physical behavior during the interaction, considering parameters such as, e.g., Movement mode (e.g., rigid, elegant), Autonomy (e.g., no autonomy, full autonomy), Noise produced while it is moving, Adaptability, Animacy (e.g., alive, natural), Interactivity (e.g., no causal behavior, fully causal behavior).
6.	Robot's social behavior.	The user's judgment of the robot's social behavior considering parameters such as e.g., Companionship, Initiative (e.g., not giving orders, not being intrusive), Social relationship (e.g., telling its story, having a real exchange of opinion), Social norms (e.g., no knowledge, full knowledge), Communication.
7.	Robot task performance.	The user's judgment of the robot during a specific performance, considering the efficiency (e.g., time on task), Effectiveness (e.g., task completeness, number of errors), and Utility.
8.	Human judgment before the interaction with a cobot.	The user's perception of the robot before the interaction, based on. Perception and effect, anxiety (e.g., toward communication capability, toward behavioral characteristics), Attitudes toward use, Expectation (e.g., performance expectancy, effort expectancy), Acceptance, Perceived safety (e.g., speed), Trust (e.g., Reliability), Intention to use.
9.	Human judgment of the performance with a cobot.	The user judgment of the robot during a specific performance task, considering. This includes aspects such as e.g., Acceptance, Perceived Safety, Trust, Control (e.g., the robot always listens), Comfort, Intention to use again, Enjoyment (e.g., pleased, bored), Satisfaction, Usability, Frustration, Stress, Cognitive workload.
10.	Human-Factors personality-based.	The user's self- description regarding their own personality characteristics, like e.g., ethics (e.g., social impact, social acceptance), Personality traits, Self-confidence, and Personality to trust.
11.	Human-Factors ability-based.	The user's self- description regarding their own work characteristics, like e.g., self-efficacy (e.g., a robot setup, technology familiarity), Expertise, and Competence.
12.	Task performed.	The characteristics of the specific performed task during the interaction e.g., Type of task, Perceived usefulness of the robot, Physical effort, Task difficulty, and Task criticality.

13. The environment of interaction.	The specific characteristics of the environment where the task was performed. This dimension refers, for instance, to the Workstation layout, Workstation elements, Environment aspects (e.g., illumination, noise, dust), and Application context (e.g., industry, healthcare).
14. Team involved during the task performance.	The members involved in the specific task performed, considering e.g., the number of humans and robots, Members' roles.
15. Interaction aspects.	The interaction aspects of the specific performed task, in terms of, for instance, knowledge of the robot's status, Situation awareness (e.g., feedback), Functionality, Ease of use, Learnability, Memorability, Interface type (e.g., physical-based interface, graphical-based interface, vocal-based interface, gesture-based interface)

Note. *Adapted from "Quantifying the Subjective Experience in Human Robot Collaboration: Towards a Validated Framework" Borsci et al. (2024)*

In their study, Borsci et al. (2024) investigated the dimensions to evaluate the subjective experience in HRC via a Delphi study as used by Borsci et al. (2024) which involved 81 experts who rated the importance of each dimension on a 9-point scale, providing a comprehensive understanding of how these human-related factors impacted the overall user experience in HRC settings.

Firstly, each dimension required a median score of 6 or higher on a 9-point Likert scale, ensuring that the average expert opinion indicated a positive agreement on the dimension's importance. Additionally, there needed to be a substantial level of expert agreement, with at least 75% of experts concurring on the importance of each aspect within the dimension. Lastly, the Interquartile Range (IQR) for the dimension's scores had to be 2 or less, signifying a narrower spread in the responses and thus indicating a higher level of agreement among the experts. These criteria were used to assess expert consensus on the dimensions.

The results of the Delphi study achieved consensus among experts on ten of the fifteen proposed dimensions for evaluating HRC. These dimensions are D1, D2, D4, D5, D7, D9, D11, D12, D13, and D15. Notably, dimensions D4 (Robot competence features) and D15

(Interaction aspects) received the highest level of consensus. However, the dimension D2 (Robot physical appearance), while included among the ten, showed a relatively lower consensus, with only 65% of experts considering it relevant.

Notably, 'Robot's Social Behavior' (D6) had a median score of 6, with 62% agreement among experts and an IQR of 3. 'Robot's Emotional Appearance' (D3) and 'Human-Factors Personality-Based' (D10) both received a median score of 6, but only 60% and 53% agreement, respectively, and an IQR of 3. For 'Human Judgment Before the Interaction with a Cobot' (D8), the median score was 6, with 65% expert agreement and an IQR of 2. Lastly, 'Team Involved During the Task Performance' (D14) showed a median score of 6, 80% agreement, and an IQR of 2. There was no agreement on the importance of these five dimensions in accordance with the criteria for agreement as such these dimensions were considered less relevant, indicating a need for further investigation or possible re-evaluation.

Borsci et al.'s (2024) findings emphasize the crucial role of subjective evaluation in HRC and acts as a guiding light for further research. The gaps and disagreements identified in their study indicate the evolving nature of UX assessment in HRC, where both subjective and objective tools must be continuously refined. The study's identified dimensions act as a comprehensive checklist, aiding experts in the thoughtful design of robots and ensuring key user interaction elements are considered. However, the utility of these dimensions extends beyond guidance for experts and designers; it marks the starting point for creating practical assessment tools. Translating these dimensions into specific, user-focused assessment items is a critical step. This process involves developing questions or statements that users can employ to evaluate their experiences with robots, with each item directly linked to a specific dimension. This approach enables the capture of detailed data from users that reflect their mental models, providing insights into their perceptions, beliefs, and expectations regarding robotic systems.

1.3. Mental Models

Building upon the established dimensions for HRC, there needs to be a focus on understanding how these dimensions align with and influence the mental models of users. As previous stated it is crucial to enhance the design and evaluation of robotic systems, ensuring that they are in sync with user, mental models' expectations, and experiences. Mental models are the ways in which individuals internally represent and understand the external world. These internal representations act as cognitive blueprints, shaping our expectations, beliefs, and assumptions about how the world works (Tabrez et al., 2020). In the context of HRC this means that a user's mental model of a robot would encompass their understanding and expectations of how the robot behaves, functions, and interacts in various situations (Rosén et al., 2022).

In HRC, the theoretical construct for measuring experience is designed to encapsulate key dimensions that mirror the aspects of robots as perceived and understood by users. It is not merely a tool for data collection but a means to bridge the gap between the robotic system's design and the users' cognitive framework. Therefore, aligning this construct with users' mental models is essential (Vázquez-Ingelmo et al., 2021). This alignment ensures that the construct accurately reflects the nuanced ways users interact with, perceive, and respond to robots. It also ensures that the dimensions within the construct are relevant with the users' real-world experiences and expectations understanding (Carley & Palmquist, 1992).

Moreover, understanding and incorporating these mental models into the theoretical construct can significantly enhance the design and evaluation of robotic systems. It enables designers and evaluators to view the robot through the users' eyes, leading to more empathetic and user-centric designs (Mustapha Mouloua & Hancock, 2019). This approach also ensures that evaluations conducted using the construct are grounded in the actual

experiences of users, providing more meaningful insights into the effectiveness and user acceptance of robotic systems (Piras, 2023).

In essence, the theoretical construct for measuring robot experience in HRC must be informed by and aligned with the users' mental models. This alignment is critical for capturing the full spectrum of user experience, from objective performance metrics to subjective perceptions and interactions, thereby ensuring that robotic systems are designed and evaluated in a manner that truly resonates with the end users. Therefore, as previously discussed, its essential to move towards the creation of assessment items. This step is vital for understanding how users categorize various elements within the identified dimensions, thereby revealing their mental models. Such an approach enables us to gain insights into how users perceive and interpret different aspects of HRC, reflecting their cognitive frameworks and expectations interactions (Mustapha Mouloua & Hancock, 2019).

1.4. Research Objectives

Building on the previous studies (Prati et al. 2022; Borsci et al 2024) this research aims to achieve two primary objectives.

The first objective of this study is to decide whether to keep or remove the dimensions of HRC that were identified in previous research that lacked a consensus among experts regarding their relevance. These dimensions were investigated in previous research, but experts couldn't agree on whether they were important/relevant. The aim is to see whether these dimensions should be included in the comprehensive 15-dimensional model of HRC, or not.

Given that that these dimensions lacked consensus on the previous work of Prati et al. 2022 and Borsci et al 2024. It is reasonable to control as a sub-objective if there are differences caused due to experience level. Experts with varying level of experience could

have different judgments about the importance of aspects that should be considered in the assessment of HRC. In general, in various industries experience definitely shapes different views on subject matters (Cooke & Goossens, 2008). Typically people with less experience follow a more theoretical point of view whereas those with experience follow a more practical point of view as noted by Van Barneveld and Strobel (2018) and HRC is no exception to this, as such it would be interesting to see if we can account for a lack of consensus by assessing the differences in opinions about these 5 dimensions that lack consensus by level of experience.

In line with these studies of Cooke & Goossens, (2008) & Van Barneveld and Strobel (2018) we expect that the years of experience of an expert to influence the rating concerning the importance of dimensions to assess HRC, therefore we would like to test if years of experience of experts makes for a significant difference on the ratings of the 5 dimensions under investigation.

The second objective is to generate specific items for each of the fifteen HRC dimensions for scale development and establish face validity of these items via card sorting which is the second step in the creation of an inventory as stated by Borsci et al. (2024). This objective serves to determine if individuals, especially those without expert knowledge, can correctly match the developed items to their corresponding HRC dimensions, this can indicate alignment with a user's mental model defined as the way in which a person categorizes items in their mind (Schmettow & Sommer, 2016; Ntouvaleti & Kastanos, 2022), establishing face validity. Face validity, as defined by Beerlage-de Jong et al. (2020), is crucial in ensuring that the scale's items are comprehensible to individuals likely to interact with collaborative robots in real-world scenarios.

These, two objectives work together to both refine the conceptual understanding of HRC and provide practical means for its evaluation and application. By achieving these goals, the

study aims to take a significant step forward in creating the first version of an evaluation scale to assess experiences with cobots. These 2 objectives will be addressed in parallel, namely, study 1: Consensus Study and study 2: Item Generation and Card Sorting for Face Validity.

2. Study 1: Consensus study

2.1. Methods

2.1.1. Participants

To achieve our research goals, experts from the previous study were contacted for a new Delphi consultation, focusing on the five dimensions lacking consensus. In addition to re-engaging experts from the initial study, additional experts in the field of HRC were contacted for participation in the new Delphi consultation. A total of 21 experts participated in this second-round consensus study. The gender distribution among the participants was predominantly male, accounting for 76.2% (n = 16), followed by female participants at 19% (n = 4), and 4.8% (n = 1) preferring not to answer. The distribution of experience levels revealed that most participants had 1 to 5 years of experience (47.6%, n = 10), followed closely by those with 5 to 10 years of experience (33.3%, n = 7). A smaller group had more than 10 years of experience (9.5%, n = 2) or less than 1 year (9.5%, n = 2). Regarding previous study participation, 61.9% (n = 13) of the participants confirmed their participation in the first study, while 28.6% (n = 6) did not participate, and 9.5% (n = 2) preferred not to say. Geographically, the participants were diverse, with the highest representation from Italy (42.9%, n = 9), followed by Portugal (19%, n = 4). Other countries like Denmark, Spain, Mexico, Jamaica, Greece, Norway, the United States of America, and France each had one participant representing them (4.8% each). The domains of expertise varied significantly among participants, with the most common being "Cobot for Industry" (23.8%, n = 5). Other domains included a mix of cobot applications in areas such as Warehouse, Education,

Healthcare, Domestic, and Social Interactions, showcasing a wide range of interests and specializations within the HRC field.

2.1.2. Materials

The survey (see Appendix A) was designed to gather expert evaluations on the 5 dimensions of HRC that lacked consensus and will be hosted on Qualtrics, a survey software.

- *Informed Consent:* Prior to beginning the survey, participants were presented with an informed consent form. This form outlined the purpose of the study, what participation involved, data confidentiality, and the voluntary nature of their involvement. Having given consent participants be able to proceed to the survey.
- *Task Description:* Following the informed consent**,** participants received a clear description of the task at hand. That included an overview of the study's objectives, the significance of each dimension under review, and instructions on how to complete the survey.
- *Review of Dimensions:* Participants were then provided with a link to review the 15 dimensions of HRC identified in the previous round of the study, including those that resulted in consensus and those that did not. This ensured that participants have a comprehensive understanding of the context and scope of each dimension.
- *9-Point Likert Scale*: For each of the five dimensions that previously resulted in disagreement, participants were asked to rate if they agree with the disagreement of the 5 dimensions in the previous study. The rating was done using a 9-point Likert scale, ranging from 1 (strongly disagree) to 9 (strongly agree). This quantitative measure was designed to gauge the perceived relevance of each dimension.
- *Open-ended Questions:* In addition to the Likert scale ratings, there was an optional open-ended question for each of the five dimensions. These questions aim to gather qualitative insights into why there might be disagreements regarding a certain

dimension and how it might be improved. This qualitative component provided depth to the understanding of each dimension and the reasons behind any discordance in ratings.

- *Demographic Questionnaire:* Information was gathered at the end of the survey about the experts, such as but not limited to; gender (male, female, and non-disclosed), experience levels (ranging from less than 1 year to over 10 years), and if they participated in the previous study, which country they are from and the domains of expertise.

2.1.3. Procedure

The procedure for the second round of the Delphi study was carefully structured to comply with ethical standards and to enable effective participation from both the returning and new experts. Prior to the study, consent was obtained from the Behavioral, Management, and Social Sciences (BMS) Ethics Committee of the University of Twente. The ethics committee reviewed the study's objectives, methods, and participant engagement strategies, and awarded approval. Following ethical approval, emails were sent out to both the returning experts from the previous survey and the newly identified experts. These emails detailed the objectives of the second round of the study and expressed appreciation for their participation in advancing HRC research (see Appendix B). Included in the email was a direct link to the Delphi study hosted on Qualtrics. Upon clicking the survey link, participants were first presented with an informed consent form. This form outlined the study's purpose, procedures, voluntary nature of participation, confidentiality measures, and contact information for any queries. Participants were required to read and provide consent before proceeding to the main survey. Once consent was given, participants were directed to the main survey. Here, they encountered the five dimensions that required further evaluation. Each participant was asked to rate the importance of these dimensions using a 9-point Likert

scale and had the option to provide additional qualitative feedback through open-ended questions. At the end of the survey, participants were asked to provide demographic information.

2.1.4. Data Analysis

The data analysis involved a quantitative approach utilizing R for data analysis, the R code can be seen in Appendix C.

- *Descriptive Statistical Analysis:* For each of the five dimensions evaluated in the survey, we calculated the mean, standard deviation, median, interquartile range (IQR), percentage of agreement. These measures provided a foundational understanding of the expert opinions, revealing the central tendencies and variability of the responses.
- *Criteria for Establishing Consensus:* To systematically assess consensus among experts, we followed specific criteria drawn from the initial Delphi study by Prati et al. (2022) and Borsci et al. (2024). These criteria were:
    - Mean score: We considered a mean score >5 as indicative of a consensus on the removal of a dimension.
    - Percentage Agreement: There should be 75% agreement among experts on the decision to remove any of the 5 dimensions to ensure a significant level of consensus. i.e., the percentage of experts that agree (score > 5) to the fact that the dimension should be removed.
    - IQR: An IQR of 2 or less was set as a threshold to confirm that the responses were not overly dispersed, indicating a tighter consensus.
- *Difference in the agreement levels of experts on the five contentious dimensions of HRC based on years of experience in HRC:* A linear regression was implemented to see if there was significant differences in agreement scores were present among experts differing years of experience in HRC.

- *Qualitative Feedback Analysis*: It was decided to analyze the open-ended question, "Why do you think there was disagreement on this dimension?" for sentiment analysis. This question could yield sentiment data. In contrast, the second question, asking how the dimension could be improved, tends to generate positive responses and is less suitable for sentiment analysis and finding out why there's a lack of consensus on these dimensions. We started by counting the comments for each dimension and then used the "syuzhet" sentiment analysis tool from the R library to categorize comments as either "Positive" or "Negative" based on their sentiment. Positive comments are comments that the sentiment analysis tool identified as having a positive emotional tone. They might include expressions of approval, satisfaction, or optimism related to the specific dimension they are addressing. Negative comments, are identified as having a negative emotional tone, potentially containing criticisms, concerns, or pessimism regarding the dimension in question. Each comment was then counted based on whether they were positive or negative and the rest would be considered neutral.

2.2. Results

The opinions of experts were sought regarding the possible exclusion of dimensions D3, D6, D8, D10, and D14. The results show the average scores and the consensus among these experts which can be seen in table 2. According to the established criteria for expert consensus, none of the dimensions achieved a mean score greater than 5, nor did they meet the criteria of at least 75% agreement among experts or an interquartile range (IQR) of 2 or less. This indicates a general lack of consensus among the experts regarding the removal of these dimensions. Specifically, Dimension D10 (Human-Factors Personality-Based) had the highest mean score of 5.14 but still did not satisfy the necessary criteria for consensus.

Table 2

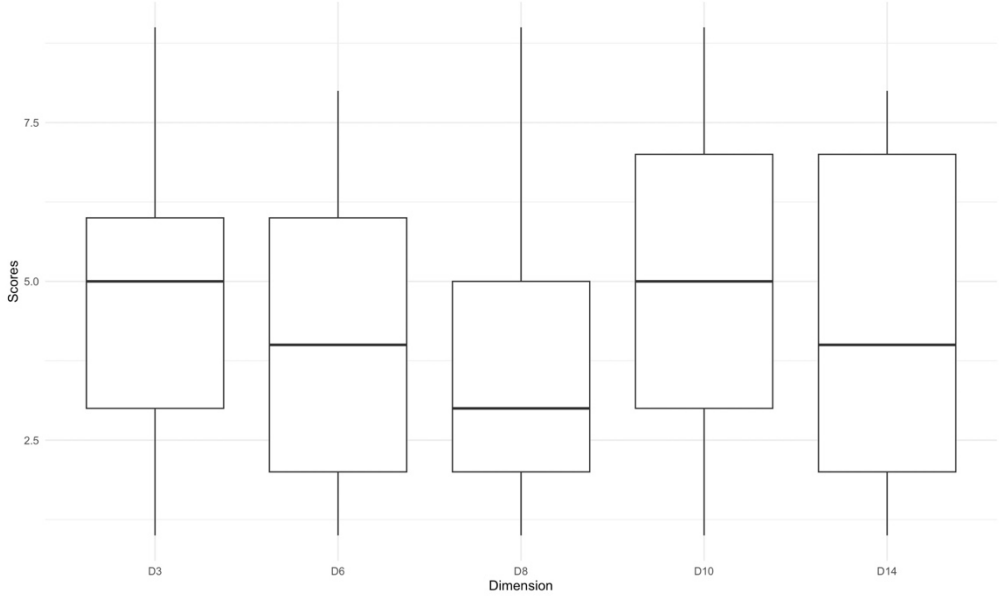
Agreement of the experts in the panel toward each dimension of HRC. Average of the rating, Standard Deviation (SD) Median (Mdn) Interquartile range (IQR), percentage (%) of experts that agreed about the removal of the dimension and Consensus was achieved (Y) or not achieved (N).

Dimensions	Experts' opinions					Consensus
	Mean	Sd	Mdn	IQR	% of experts	
D3	4.71	2.17	5	3	33%	N
D6	4.05	2.38	4	4	29%	N
D8	3.76	2.36	3	3	24%	N
D10	5.14	2.37	5	4	48%	N
D14	4.29	2.49	4	5	38%	N

For D3 (Robot's Emotional Appearance), the mean score was 4.71 (SD = 2.17), with a median of 5.0, an IQR of 3.0, and a 33.33% agreement. D6 (Robot's Social Behavior) had a mean score of 4.05 (SD = 2.38), a median of 4.0, an IQR of 4.0, and a 28.57% agreement. D8 (Human Judgment Before Interaction with Cobot) recorded a mean of 3.76 (SD = 2.36), a median of 3.0, an IQR of 3.0, and a 23.81% agreement. For D10 (Human-Factors Personality-Based), the mean was 5.14 (SD = 2.37), with a median of 5.0, an IQR of 4.0, and a 47.62% agreement. Lastly, D14 (Team Involved During Task Performance) showed a mean of 4.29 (SD = 2.49), a median of 4.0, an IQR of 5.0, and a 38.10% agreement. A visual representation of the distribution, central tendency, and variability of these findings can be seen in Figure 1.

Figure 1

Boxplot representations of the distribution, central tendency, and variability.



In Figure 1, the boxplots portray the five dimensions with boxes representing the IQR range and the line within each box represents the median. The boxes, vary in size, indicating the extent of opinion variability which is also representative of the SD among experts. This clear difference in box sizes shows there are differences in opinion on whether these dimensions should be removed or not.

Additionally, the results of the linear regression indicated no significant differences in agreement levels among the 4 different experience groups, less than a year, 1 to 5 years, 5 to 10 years, and more than 10 years across all dimensions.

Finally, each dimension in the survey was accompanied by two open-ended questions. Participants were asked to explain why there were disagreements within the dimension and suggest potential rephrasing for the dimension. As previously stated, our focus was on understanding the reasons behind disagreements related to the dimensions. The qualitative feedback from experts was analyzed for the 5 dimensions of HRC under investigation. The

results, summarized in Table 3, include the number of comments and the distribution of those comments in numbers under positive, negative, or neutral categorization per dimension.

Table 3

Comments of the experts on thoughts why there is disagreement on these dimensions in the survey. Number of comments in total and frequency of positive negative and neutral comments per dimension with the use of a sentiment analysis

	Number of Comments	Positive Comments (%)	Negative Comments (%)	Neutral Comments (%)
D3	8	37.5	50	12.5
D6	9	77.8	22.2	0
D8	9	88.9	11.1	0
D10	9	88.9	11.1	0
D14	7	57.1	28.6	14.3

These findings reveal a range of sentiments regarding the reasons for disagreement on these dimensions. In the case of D3, there was an almost equal distribution of positive and negative comments, reflecting a mixed opinion. For example, Participant 4 provided a positive comment, stating, "Cannot be ignored as we move into real human-robot collaboration," while Participant 1 expressed a negative view, saying, "Sometimes it is felt that, in specific contexts, robots' appearance is irrelevant for the purpose." Dimensions D6, D8, D10, and D14 also exhibit varying levels of sentiment, indicating a diversity of opinion on why these dimensions lacked consensus with the common theme in the comments talking about varying contexts where these dimensions apply or do not apply.

The results, consistent with prior research, reveal a lack of consensus among experts regarding the five dimensions. None of the dimensions met the removal criteria. While expert opinions varied, a common theme emerged – the importance of these measures depends on the context. As a result, it was decided to keep all dimensions but present them as optional for assessing the user experience with cobots.

2.3. Discussion

The study aimed to assess expert consensus on the inclusion or removal of specific dimensions within a 15-dimensional model for evaluating HRC whilst also assessing whether there are differences in levels of agreement due to years of experience of experts. Our focus was on five dimensions: D3, D6, D8, D10, and D14. An expert survey was used to gather expert opinion, a linear regression was used for assessing differences amongst groups based on years of experience. A sentiment analysis was used on the open-ended questions within the survey, providing nuanced insights into expert opinions. Based on the results we are going to keep D3, D6, D8, D10, and D14 and treat them as optional dimensions in HRC evaluations depending on the context of use and the HRC experts' opinion on their relevance.

The data presents a dilemma, while there is a slight lean towards D10's removal, the lack of a strong consensus argues for its retention. This decision is underpinned by the need for a comprehensive understanding of personality factors in HRC settings. The other dimensions investigated (D3, D6, D8, D14) exhibited a similar trend, with varying levels of agreement but none reaching the established threshold for consensus on removal. The absence of a clear consensus on these dimensions suggests their potential importance in capturing the multifaceted nature of HRC. The divided opinion among experts across all 5 dimensions further indicates that these dimensions may have context-specific relevance or importance for particular user demographics. Additionally, the sentiment expressed by experts on why there is disagreement on these 5 dimensions revealed a mix of positive, negative, and neutral sentiments across the dimensions. For example, some dimensions might have received almost equal distribution of positive and negative comments. This indicates a balanced perspective, where experts see both strengths and areas for improvement. Such mixed feedback suggests that while some aspects of these dimensions are well-received or deemed essential, others may be controversial or less understood.

Interestingly, as there was no difference in opinion due to varying levels of experience this indicates that the concerns and perspectives regarding these dimensions transcend experience levels. Both novice and experienced professionals in HRC share similar views, highlighting core issues recognized across the board. For D8, the absence of significant opinion differences based on experience levels suggests that apprehensions or judgments about cobots prior to interaction are common, regardless of one's familiarity or tenure in the field. This could imply that preconceived notions about cobots are not necessarily mitigated by increased years of expertise, underscoring the need to address these judgments irrespective of expertise. The same could be said for D3, D6, D10 and D14.

Ultimately, the lack of clear consensus on these 5 dimensions points to their potential significance in capturing the complex nature of HRC. Maintaining these dimensions could lead to a more comprehensive and nuanced understanding of HRC in certain contexts hence why they have been made optional, ensuring that the assessment model remains attuned to various interaction dynamics and contexts.

3. Study 2: Item Generation and Card Sorting for Face Validity

3.1. Methods

3.1.1. Design

From the 15 dimensions 10 were confirmed by experts in the study by Prati et al. 2022 and Borsci et al 2024 to be relevant and there was a lack of consensus on 5 namely D3, D6, D8, D10 and D14. These 5 dimensions were reinvestigated in study 1 and a lack of consensus remained, this means these dimensions could still lead to a comprehensive understanding of HRC. As such these 5 dimensions were retained. For this design this means that all 15 dimensions will be utilized for study 2. After reaching these conclusions, the next step was to generate items per dimension for an initial scale to assess experience with cobots.

In the initial phase, two experts in Human Factors drafted a preliminary version of the items, adhering to specific, well-defined criteria to ensure relevance and accuracy. This preliminary set of items was then subjected to a rigorous review process by three additional experts specializing in both Human Factors and Human-Robot Collaboration. These experts carefully revised the items wording and evaluated their alignment with the characteristics of the defined dimensions, ensuring each item was optimally structured for the HRC assessment context. This resulted in the generation of 71 items across 15 dimensions as seen in table 4. The 71 items are listed with their statements in their respective dimensions in Appendix D.

Table 4

The 71 items and the dimensions they are characterized under.

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
1	4	10	15	18	22	27	30	36	46	51	54	59	64	66
2	5	11	16	19	23	28	31	37	47	52	55	60	65	67
3	6	12	17	20	24	29	32	38	48	53	56	61		68
	7	13		21	25		33	39	49		57	62		69
	8	14			26		34	40	50		58	63		70
	9						35	41						71
								42						
								43						
								44						
								45						

To assess the face validity of these generated items, employed was a closed card sorting methodology to participants had to sort items into predefined categories (the dimensions). This process helps in assessing the face validity of the scale by determining whether the items subjectively appear to be relevant to the dimensions they are intended to measure. The extent to which participants group each item in the intended construct (dimension) provides insight into how well the items represent these dimensions at first glance (Beerlage-de Jong et al., 2020). As such through closed card sorting, we wanted to test if participants could match the 71 items to any of the 15 dimensions created in previous

studies by testing their ability to put an item correctly or incorrectly into the dimensions defined by experts. This is a measure of distance between the designers of the scale and the user's mental model.

3.1.2. Participants of the Card Sorting

Participants were recruited through various channels to ensure a diverse representation. Once identified, they were sent a link to participate in the study, which includes a link to the Qualtrics card sorting platform. A total of 43 participants were included in the study. The age of participants ranged from 18 to 44 years, with a mean age of 25.21 years (SD = 5.62). The distribution of age was relatively skewed towards younger adults. Regarding sex, most of the participants were female (n = 29, 67.4%), while male participants constituted 32.6% of the sample (n = 14). Participants' experience with collaborative robots varied: a small portion had experience with collaborative robots (n = 3, 7%), the majority had no experience (n = 31, 72.1%), a significant number were not sure about their experience (n = 9, 20.9%), and none of the participants chose the 'prefer not to say' option. The analysis of demographic data was conducted in R and the code can be seen in Appendix F.

3.1.3. Materials

The card sorting activity was conducted using Qualtrics. The platform enabled participants to interactively drag and drop items into designated categories, facilitating the card sorting experience. The card sorting survey can be seen in Appendix E. The materials used in the card sorting activity are as follows:

- *Items for Sorting:* Participants were presented with 10 items, each representing a distinct aspect of HRC. The process of item creation was systematic and thorough. Initially, the items were crafted by the research team and then refined based on input

from three experienced Human Factors experts, all of whom had significant knowledge in HRC assessment as aforementioned in the design.

- *Dimensions for Categorization:* Fifteen dimensions were provided, each corresponding to a specific aspect of HRC as identified in prior research. This guided participants in categorizing the items based on their understanding of these dimensions. In addition, there was a "None of the above" dimension, where participants placed items that they believe do not fit into any of the provided dimensions.
- *Informed Consent:* Participants were presented with an informed consent form via Qualtrics. This form will detail the purpose of the study, the procedures involved, the voluntary nature of participation, and the confidentiality of their responses.
- *Task:* Each participant was assigned a subset of 10 items from the total of 71 items. This sequence was arranged so that after one participant categorized their set, the next participant received a subsequent set of 10 items. This rotation continued until all 71 items were sorted by different participants. And then begins again from the first 10 for the next set of participants.

3.1.4. Procedure

The participants individually followed this online procedure to perform the card sorting. Upon accessing the link, participants viewed the informed consent form first. This form provided detailed information about the study, including its purpose, what participation involves, the voluntary nature of participation, and the confidentiality of responses. Participants had to read and agree to the terms of the informed consent before proceeding. After consenting, participants received an introduction to the card sorting task. This included instructions on how to use the Qualtrics interface for dragging and dropping items into dimensions, and what the task entails. Before beginning the actual card sorting, participants

engaged in a practice task. This was designed to familiarize them with the interface and the process of sorting items into dimensions. Participants proceeded to the main card sorting activity. They were presented with 10 items, one at a time, and asked to categorize each item into one of the 15 provided dimensions, or into the "None of the above" dimension if they feel the item doesn't belong in any of the provided categories. Participants were encouraged to take their time to thoughtfully consider where each item should be placed. The entire activity is expected to take approximately 15 to 20 minutes. After completing the card sorting task, participants were asked to provide some basic demographic information. Upon completion of the survey, participants received a thank you message, along with a debrief about the study. They were also be provided with contact information should they have any questions or wish to receive information about the study results.

3.1.5. Data Analysis

To evaluate the face validity of this initial scale, the focus was on item-level agreement, how many times each item was categorized within the expected dimension as a percentage by the participants. The criteria for whether an item should be removed or kept is if it has an agreement level of at least 50% as discussed with HRC experts.

3.2. Results

3.2.1. Face Validity

The findings show that 43 items out of the 71 were sorted into a dimension at an item level agreement of 50% or higher. Of these 43, 35 were sorted into the expected dimension and 7 were not sorted into the expected dimension as illustrated in table 5.

Table 5

The table contains the items that were assigned to a certain dimension by 50% or more of the participants. It shows the percentage of participants that assigned an item to certain

dimension. The item numbers marked with an asterix (*) are sorted to an unexpected dimension.

Dimension	Item-Level Agreement					
D1	i1 83%	i2 80%	i3 57%	i75* 75%		
D2	i4 100%	i5 75%	i7 50%	i13* 86%		
D3	i10 100%	i11 60%	i14 57%			
D4	i16 50%	i17 86%				
D5	i18 86%	i20 80%				
D6	i22 88%	i23 71%	i24 100%			
D7	i8* 57%	i28 63%	i40* 60%			
D8	i30 50%	i31 67%	i32 50%	i33 50%	i34 60%	i35 100%
D9	i15* 50%	i29* 67%	i42 50%	i45 71%		
D10	i49 83%	i50 83%				
D11	i46* 50%	i51 71%	i52 60%			
D12	i55 50%					
D13	i59 100%	i60 83%	i61 88%			
D14	i64 57%	i65 50%				
D15	i41* 50%					

In line with the discussion with experts and the need to make a scale of reasonable size it was decided to retain 2-3 items per dimension were appropriate except for D14 as this

dimension only had 2 items. This will result in the creation of a scale of maximum 44 items. The items kept per dimension are as follows:

Dimension 1

In Dimension 1, participants attributed four items based on their perceptions. While all items were associated with the dimension, Item 1 (83%) and Item 2, (80%) received the highest attribution percentages and directly addressed the concept of easiness in robot regulation. Additionally, Item 37, "I realized while collaborating with the [System/collaborative robot name] that it was pleasing to use and easy to control during the task" (75% attribution), explicitly related to the dimension's essence. Consequently, these three items were retained for the scale, while Item 3, "I believe that from a physical point of view it appears to be easy to manipulate and put the robot into position" (57% attribution), was excluded. Item 3 also seems by wording to relate to Dimension 2 robots' physical appearance where it could be more appropriate.

Dimension 2

Item 4, with a unanimous agreement of 100%, and Item 5, with a high agreement level of 75% and Item 7 with an agreement level of 50%. These items are kept for further scale development due to their strong consensus among participants. Item 13 "I did not consider the [System/collaborative robot name]'s appearance disturbing." was not retained, although it had a higher agreement of 86% it fails to address the dimension robots physical appearance as a whole and is probably more appropriate in its allocated dimension D3 Robots Emotional Appearance.

Dimension 3

Item 10, with a unanimous agreement of 100%, is retained, reflecting a consensus on the robot's likable and attractive emotional expression. Items 11 with agreement levels of 60, was retained. Additionally, Item 14, which assesses the robot's capacity to convey warmth

and receives a 57% agreement, was retained. These agreement levels indicate that Items 10, 11 and 14 are kept for further scale development in HRC.

Dimension 4

Item 17, with an 80% agreement, is retained for further scale development, suggesting a strong categorization of responsiveness and transparency. Item 16, with a 50% agreement is also retained.

Dimension 5

Item 20, with an 80% agreement, is retained. Item 18 also demonstrates high agreement at 86%, and is therefore retained.

Dimension 6

Item 24, with a 100% agreement rate, is retained, reflecting a unanimous view of the robot's meaningful social interactions. This consensus indicates that the item aligns well with participants' understanding of social interaction in robotics. Items 22 and 23 also exhibit strong agreement levels at 88% and 71%, respectively as such they are retained.

Dimension 7

Item 8 and Item 28 are retained for further scale development within D7, with item 28 having 63% agreement and being correctly placed in line with expectations. Whilst item 8 originally belongs to D2 its wording can definitely be associated with D7s Robots Task Performance as it states, “The [System/collaborative robot name]'s perceived robustness (e.g., its ability to withstand physical stress, challenges etc.) met the specific requirements for the task and context of usage.” For these reasons with the adequate item level agreement of 57% it is retained for further scale development. While item 40 also had high item level agreement the wording of the item does not align with the wording of the dimension or its sub-factors, as item 40 states, “I was highly satisfied with the [System/collaborative robot name]‘s performance,” and D7 does not include satisfaction within its description so it was decided not to retain it.

Dimension 8

Item 35, with a 100% agreement, is retained, while Items 31 and 34, each with agreement levels of 67% and 60% respectively, are also retained. Conversely, Items 30 31 and 32, had a 50% agreement level and do meet the criteria for inclusion however as they have the lower percentages among the 6 and we must retain 3 they will not be retained.

Dimension 9

Item 45, with a 71% categorization by participants, is retained. Similarly, Item 42, with a 50% categorization, is also retained. These 2 items were sorted correctly in line with expectations as well as such they are retained. It was also decided to retain item 29 because of the 67% item level agreement within this dimension. Item 29, "I believe that the collaboration with the [System/collaborative robot name] was useful by enabling tasks to be completed in an efficient and effective manner," aligns more closely with the description provided for the human judgment of the robot during a specific performance task. This item directly addresses the concept of collaboration efficiency and effectiveness, which is in line with aspects like Acceptance, Perceived Safety, Trust, Control, Comfort, Intention to use again, Satisfaction, Usability, Frustration, Stress, and Cognitive workload. Item 15 although it also fits dimension 9 it had 50% item-level agreement and was sorted into D9 instead of D4. As such it is the least suitable comparably to the others to be retained for further scale development.

Dimension 10

Item 49, with an 83% categorization, and Item 50, with an 83% categorization, were both retained, indicating a strong association with Dimension 10 as they were categorized into their expected dimension.

Dimension 11

Item 51, with a 71% categorization by participants, and Item 52, with a 60% categorization by participants, are retained. Additionally, item 46 is retained as it has a 50% item level agreement, and the wording, "I feel confident in my ability to use the [System/collaborative robot name] to achieve key tasks," aligns well with the user's confidence in their ability to effectively utilize the robot for important tasks. It reflects self-

efficacy, which is the belief in one's own capability to perform a specific task or achieve specific goals, which heavily aligns with D11s description.

Dimension 12

Within this dimension, Item 55, with a 50% item-level agreement is retained, being the only item in D12 to meet the criteria for inclusion. It was also decided to retain item 57, "I believe the [System/collaborative robot name] was useful in accomplishing the task." As the wording in this dimension heavily associates with the description in of D12 as it aligns with perceived usefulness of the task in HRC context. Additionally, whilst item 57 did not meet the criteria of 50% item level agreement it was relatively close at 40% item level agreement.

Dimension 13

Item 59 had a 100% agreement among participants, Item 60, had am 83% agreement, and Item 61, had an 88% agreement, as such these items were all retained due to their strong item level agreement. These high agreement percentages signified their alignment with the dimension's focus as well as it was sorted into the expected dimension.

Dimension 14

In this dimension, both Item 64, with a 57% item level agreement, and Item 65, with a 50% item level agreement, were retained as they both items meet the criterion for inclusion.

Dimension 15

Item 41 although not expected to be in D15 it was the only item that was sorted into this dimension with an item level agreement of 50% which meets the criteria for inclusion. This association makes sense because the wording of item 41 says, "I found the [System/collaborative robot name]'s interface or interaction methods (e.g., touch panel, voice commands, haptic feedback) highly usable" aligns with the description of D15. In the sense that both the description of D15 and Item 41 focus on the usability and effectiveness of the robot's interface and interaction methods during a specific task. Additionally, to meet the criterion of having at least 2 items per dimension it was decided to retain item 67 as it at had

the 2nd highest item-level agreement at 40%. And is already an item that was classified under D15 by experts in HRC and human factors.

Items Retained

Across the 15 dimensions, a total of 38 items were retained for use in a scale for HRC, with each dimension having a varying number of items meeting the criteria for inclusion. The selected items reflect a strong alignment with the respective dimensions, ensuring that the developed scales effectively capture the nuanced aspects of HRC from the user's perspective. The retained items can be seen in Table 6.

Table 6

Items Retained for Scale Use Per Dimension

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
i1	i4	i10	i16	i18	i22	i8	i31	i29	i46	i51	i55	i59	i64	i41
i2	i5	i11	i17	i20	i23	i28	i34	i42	i49	i52	i57	i60	i65	i67
i37	i7	i14			i24		i35	i45	i50			i61		

It must be noted, for some items to be added to ensure each dimension has at least 2 items, re-investigation into items with item-level agreement less than 50% were looked at to see their eligibility for the scale.

3.3. Discussion

The second objective of the research focused on two key aspects: firstly, developing specific scale items for each of the fifteen dimensions identified in the HRC framework, and secondly, verifying the face validity of these items through a card sorting technique. After the card sorting was conducted there was a significant reduction in the number of items in the scale, with most items being initially representative of the dimension they were put into by experts earlier. The final scale can be seen in Table 7.

Table 7

The final version of the scale developed numbered from 1 – 38.

Dimension name	Items
1. Easiness of robot regulation.	**Item 1.** It was easy to physically regulate the [system/collaborative robot name]. **Item 2**. I found it/ it appears to be easy to position the [System/collaborative robot name] components correctly. **Item 3.** I realized while collaborating with the [System/collaborative robot name] that it was pleasing to use and easy to control during the task.
2. Robot physical appearance.	**Item 4.** I had a positive impression of the [System/collaborative robot name]'s physical appearance. **Item 5.** I had a positive impression of the [System/collaborative robot name]'s dimensions i.e., high, width, length, weight. **Item 6.** The level of anthropomorphism (machinelike or humanlike) of the [System/collaborative robot name] was appropriate for the intended purpose.
3. Robot's emotional appearance.	**Item 7.** I believe that the [System/collaborative robot name]'s emotional appearance was likable/attractive. **Item 8.** I believe that the [System/collaborative robot name]'s design does not cause emotional discomfort. **Item 9.** I felt that the [System/collaborative robot name] displayed a sense of warmth during our collaboration e.g., it was social, friendly.
4. Robot's competence features.	**Item 10.** I believe the [System/collaborative robot name] is reliable and trustworthy in terms of competencies. **Item 11.** I found the [System/collaborative robot name] to be responsive and transparent in terms of competencies.
5. Robots physical behaviour.	**Item 12.** I perceived the [System/collaborative robot name] movements to be smooth and flexible. **Item 13.** I believe that the [system/collaborative robot name]'s physical behavior (e.g., noise, movement, autonomy, interactivity) during the interaction was suitable.
6. Robots social behaviour.	**Item 14.** I believe that the overall social behavior of the [System/collaborative robot name] was appropriate. **Item 15.** I believe that the [System/collaborative robot name] acted and communicated according to social norms. **Item 16.** I felt that the [System/collaborative robot name] engaged in meaningful social interactions during our collaboration.
7. Robots task performance.	**Item 17.** The [System/collaborative robot name]'s perceived robustness (e.g., its ability to withstand physical stress, challenges etc.) met the specific requirements for the task and context of usage. **Item 18.** I believe that the [System/collaborative robot name] was useful by enabling a correct (without error) performance.

8.	Human judgement before the interaction with a cobot.	**Item 19.** I expected the [system/collaborative robot name] to be reliable and quick prior to collaborate with it. **Item 20.** I accepted the idea of using the [System/collaborative robot name] for the task prior to use. **Item 21.** Before interacting with the [System/collaborative robot name], I had the intention to use it for similar tasks or interactions in the future.
9.	Human judgement of the performance with a cobot.	**Item 22.** I believe that the collaboration with the [System/collaborative robot name] was useful by enabling tasks to be completed in an efficient and effective manner. **Item 23**. I experienced no frustration while working with the [System/collaborative robot name]. **Item 24.** I would like to use the [System/collaborative robot name] again based on my experience during the task.
10.	Human factors personality based.	**Item 25.** I feel confident in my ability to use the [System/collaborative robot name] to achieve key tasks. **Item 26.** I have a trusting personality. **Item 27.** I believe my personality traits have a significant influence on my collaboration with robots in general.
11.	Human factor's ability based.	**Item 28.** My level of expertise contributed to a successful interaction with the [System/collaborative robot name]. **Item 29.** I believe that my general understanding of robotics contributed to a positive collaboration with the [System/collaborative robot name].
12.	Task Performed.	**Item 30.** The task I performed with the [System/collaborative robot name] did not require too much physical effort. **Item 31:** I believe the [System/collaborative robot name] was useful in accomplishing the task.
13.	The environment of Interaction.	**Item 32.** The environmental conditions (e.g., lighting, noise, dust) were disturbing the task. **Item 33.** The workstation layout facilitated a positive interaction with the [System/collaborative robot name]. **Item 34:** The interaction took place in a workstation with a layout that facilitated the completion of the task.
14.	Team involved during the task performance.	**Item 35.** I believe that it is possible for multiple operators (a team) to collaborate proficiently to use the [System/collaborative robot name] to achieve the task. **Item 36.** When multiple operators (a team) have to collaborate interacting with the [System/collaborative robot name] all the operators can understand their roles.

15.	Interaction Aspects	**Item 37.** I found the [System/collaborative robot name]'s interface or interaction methods (e.g., touch panel, voice commands, haptic feedback) highly usable. **Item 38.** I had good knowledge of the [System/collaborative robot name] status during the task performance.

The number of items retained varied across dimensions, reflecting a tailored approach to scale development. While the general aim was to keep 2-3 items per dimension, exceptions were made based on the specific characteristics of each dimension and the items' relevance. For example, in Dimension 14, only two items were available, and both were retained. In contrast, other dimensions had a higher number of potential items, leading to more selective retention based on agreement levels and relevance.

This variability in the number of items retained per dimension highlights the nuanced approach taken in the scale development. It acknowledges that not all dimensions require an equal number of items to capture their essence effectively. Some dimensions might be adequately represented with fewer items, especially if those items are highly relevant and have strong face validity. Ultimately the scale developed can be used to measure different aspects of HRC effectively and can be developed in further research.

4. Conclusion

This research has produced a 38-item evaluation scale spanning 15 dimensions of HRC. This scale, with a focus on UX, serves as an initial tool for assessing how humans perceive and interact with cobots in real-life scenarios. It bridges the gap between technical performance and user-centric aspects, making it a valuable asset for evaluating HRC systems to ensure that they are efficient, user-friendly, safe, and well-received.

Originating from the groundwork laid by Borsci et al. (2024) in identifying 15 HRC dimensions. Study 1 focused on the five dimensions where expert consensus was lacking. In both the Borsci et al. (2024) study and this research, consensus could not be reached

regarding the inclusion or exclusion of these dimensions. This highlights the substantial variability in expert opinions regarding these specific dimensions within HRC. It highlights the importance of recognizing their context-specific relevance and suggests that retaining them as optional dimensions was correct. For instance, in D3 regarding robots' emotional appearance would be more important for a robotic nurse aid assisting the elderly, or a conversational agent in a home setting, D3 would be a useful dimension to include because its factors can be measured in these instances. However, D3 may not be so useful in industrial settings, where robots are primarily used for precision tasks, manufacturing, or automated processes, the emotional appearance of the robot may not have any practical significance or impact on the task's efficiency and safety. In such cases, assessing D3 may not be necessary, and it can be considered less relevant or even unnecessary for the evaluation of those specific HRC systems. This could be extended to apply to D6, D8, D10 and D14, in the sense that for these 5 dimensions context is important for their applicability in evaluating HRC systems.

Next was study 2 which led to the formulation of 71 items in collaboration with HRC and human factors experts, covering all 15 dimensions. This was a critical step in creating an initial evaluation scale. The next phase involved assessing the face validity of these items through closed-ended card sorting, resulting in a scale of 38 items as seen in Table 7. These selected items were deemed the most representative according to face validity, making this scale suitable for real-world applications in assessing HRC interactions.

Aforementioned by Zhang et al. (2023) is that HRC is gaining traction, which means subsequently scales for HRC are needed now more than ever (Papetti et al., 2022). As such the development of this scale is particularly important in HRC evaluations as it provides a structured and reliable method to measure and understand how humans perceive and interact with cobots. In HRC, the quality of interaction between humans and robots directly impacts productivity, safety, and user satisfaction (Brondi et al., 2021). This scale enables a more

nuanced evaluation of these interactions, going beyond the technical. Which is important because often human centered factors are overlooked to investigate more technical aspects when it comes to robots (Prati et al., 2021).

These two studies had some limitations. Firstly, in the initial phase, there was a lower number of participants compared to the initial Delphi study, approximately 21 experts participated which is considerably less compared to the previous phase involving 81 experts. This discrepancy in participant numbers could potentially constrain the generalizability of the findings. In the card sorting study, the unequal exposure of items during the card sorting task could have influenced the results as each item was probably sorted into a dimension approximately 6 times across the 43 participants. While an optimal scenario would have involved each participant sorting all items, the extensive item list carried the risk of inducing participant fatigue. This consideration prompted us to adopt a more concise approach.

For future research this ready to use scale needs to be applied in various real-world settings. Following the steps outlined in development of an inventory by Borsci et al. (2024) the next step is to do an exploratory factor analysis on the 38-item scale which should be conducted to understand the scale's underlying structure and identify latent variables, ensuring alignment with the 15 dimensions. This should be followed by a confirmatory factor analysis to test and validate the hypothesized structure derived from the exploratory factor analysis solidifying the scale's validity and its representation of HRC dimensions. This real-world application would test the scale’s relevance and applicability across different contexts. Subsequently, further analysis on the experimentally validated items is essential, incorporating techniques like item response theory to examine each item's properties, such as difficulty and discrimination, thereby refining the scale for more accurate measurement of human-robot interaction quality (So Young Song et al., 2023). Lastly, given the diversity in

HRC environments globally, cross-cultural validation of the scale is paramount, ensuring its effectiveness and applicability across various cultures and industries.

In conclusion, this 38-item scale represents a foundational step in assessing UX with cobots. While acknowledging its current limitations, it is anticipated that with ongoing research and refinement, this scale will emerge as an asset to the field of HRC. It promises to bring enhanced clarity and efficacy to the design and evaluation of HRC systems, as well as the instruments used to measure their performance.

References

Baratta, A., Cimino, A., Longo, F., & Nicoletti, L. (2023). Digital Twin for Human-Robot Collaboration enhancement in manufacturing systems: literature review and direction for future developments. *Computers & Industrial Engineering*, 109764–109764. https://doi.org/10.1016/j.cie.2023.109764

Beerlage-de Jong, N., Kip, H., & Kelders, S. M. (2020). Evaluation of the Perceived Persuasiveness Questionnaire: User-Centered Card-Sort Study. *Journal of Medical Internet Research*, *22*(10), e20404. https://doi.org/10.2196/20404

Biermann, H., Brauner, P., & Ziefle, M. (2020). How context and design shape human-robot trust and attributions. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 74–86. https://doi.org/10.1515/pjbr-2021-0008

Borsci, S., Prati, E., Landwehr, J., & Peruzzini, M. (2024). *Quantifying the Subjective Experience in Human- Robot Collaboration: Towards a Validated Framework.*

Boyd, R., & Holton, R. J. (2017). Technology, innovation, employment and power: Does robotics and artificial intelligence really mean social transformation? *Journal of Sociology*, *54*(3), 331–345. https://doi.org/10.1177/1440783317726591

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social Robotics. *Springer Handbook of Robotics*, 1935–1972. https://doi.org/10.1007/978-3-319-32552-1_72

Brondi, S., Pivetti, M., Di Battista, S., & Sarrica, M. (2021). What do we expect from robots? Social representations, attitudes and evaluations of robots in daily life. *Technology in Society*, *66*, 101663. https://doi.org/10.1016/j.techsoc.2021.101663

Cai, M., Ji, Z., Li, Q., & Luo, X. (2023). Safety evaluation of human–robot collaboration for industrial exoskeleton. *Safety Science*, *164*, 106142. https://doi.org/10.1016/j.ssci.2023.106142

Carley, K., & Palmquist, M. (1992). Extracting, Representing, and Analyzing Mental Models. *Social Forces*, *70*(3), 601–636. https://doi.org/10.1093/sf/70.3.601

Chammas, A., Quaresma, M., & Mont'Alvão, C. (2015). A Closer Look on the User Centred Design. *Procedia Manufacturing*, *3*, 5397–5404. https://doi.org/10.1016/j.promfg.2015.07.656

Cheng, C. Y. M., Lee, C. C. Y., Chen, C. K., & Lou, V. W. Q. (2022). Multidisciplinary collaboration on exoskeleton development adopting user-centered design: a systematic integrative review. *Disability and Rehabilitation: Assistive Technology*, 1–29. https://doi.org/10.1080/17483107.2022.2134470

Ciccarelli, M., Papetti, A., & Germani, M. (2023). Exploring how new industrial paradigms affect the workforce: A literature review of Operator 4.0. *Journal of Manufacturing Systems*, *70*, 464–483. https://doi.org/10.1016/j.jmsy.2023.08.016

Cooke, R. M., & Goossens, L. L. H. J. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety*, *93*(5), 657–674. https://doi.org/10.1016/j.ress.2007.03.005

Coronado, E., Kiyokawa, T., Ricardez, G. A. G., Ramirez-Alpizar, I. G., Venture, G., & Yamanobe, N. (2022). Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *Journal of Manufacturing Systems*, *63*, 392–410. https://doi.org/10.1016/j.jmsy.2022.04.007

Davenport, T. H. (2018). The AI Advantage: How to Put the Artificial Intelligence Revolution to Work. In *Google Books*. MIT Press. https://books.google.nl/books?hl=en&lr=&id=QzNwDwAAQBAJ&oi=fnd&pg=PR5&dq=As+these+nations+continue+to+invest+in+cutting-edge+technology

Faccio, M., Granata, I., Menini, A., Milanese, M., Rossato, C., Bottin, M., Minto, R., Pluchino, P., Gamberini, L., Boschetti, G., & Rosati, G. (2022). Human factors in cobot era: a review of modern production systems features. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-022-01953-w

Fornasiero, R., Marchiori, I., Pessot, E., Zangiacomi, A., Sardesai, S., Barros, A. C., Thanous, E., Weerdmeester, R., & Muerza, V. (2021). Paths to Innovation in Supply

Chains: The Landscape of Future Research. *Lecture Notes in Management and Industrial Engineering*, 169–233. https://doi.org/10.1007/978-3-030-63505-3_8

Gervasi, R., Khurshid Aliev, Mastrogiacomo, L., & Franceschini, F. (2022). User Experience and Physiological Response in Human-Robot Collaboration: A Preliminary Investigation. *Journal of Intelligent and Robotic Systems*, *106*(2). https://doi.org/10.1007/s10846-022-01744-8

Gervasi, R., Mastrogiacomo, L., & Franceschini, F. (2020). A conceptual framework to evaluate human-robot collaboration. *The International Journal of Advanced Manufacturing Technology*, *108*(3), 841–865. https://doi.org/10.1007/s00170-020-05363-1

Hartson, R., & Partha Pyla. (2012). *The Ux Book : Process and Guidelines for Ensuring a Quality User Experience*. Morgan Kaufmann Pub.

Meissner, A., Trübswetter, A., Conti-Kufner, A. S., & Schmidtler, J. (2021). Friend or Foe? Understanding Assembly Workers' Acceptance of Human-robot Collaboration. *ACM Transactions on Human-Robot Interaction*, *10*(1), 1–30. https://doi.org/10.1145/3399433

Mukherjee, D., Gupta, K., Chang, L. H., & Najjaran, H. (2022). A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings. *Robotics and Computer-Integrated Manufacturing*, *73*, 102231. https://doi.org/10.1016/j.rcim.2021.102231

Mustapha Mouloua, & Hancock, P. A. (2019). *Human Performance in Automated and Autonomous Systems*. CRC Press.

Papetti, A., Ciccarelli, M., Scoccia, C., Palmieri, G., & Germani, M. (2022). *A human-oriented design process for collaborative robotics*. 1–23. https://doi.org/10.1080/0951192x.2022.2128222

Parvez, M. O., Arasli, H., Ozturen, A., Lodhi, R. N., & Ongsakul, V. (2022). Antecedents of human-robot collaboration: theoretical extension of the technology acceptance model. *Journal of Hospitality and Tourism Technology*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/jhtt-09-2021-0267

Piras, B. (2023, October 27). *Effects of dynamic planning on the adaptability of a cobotic system to human constraints for a cooperative HRI assembly task*. Webthesis.biblio.polito.it. https://webthesis.biblio.polito.it/28584/

Rosén, J., Lindblom, J., & Billing, E. (2022). The Social Robot Expectation Gap Evaluation Framework. *Lecture Notes in Computer Science*, 590–610. https://doi.org/10.1007/978-3-031-05409-9_43

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Simone, V. D., Pasquale, V. D., Giubileo, V., & Miranda, S. (2022). Human-Robot Collaboration: an analysis of worker's performance. *Procedia Computer Science*, *200*, 1540–1549. https://doi.org/10.1016/j.procs.2022.01.355

So Young Song, Lee, J., & Woong Yeol Joe. (2023). *Scale Development of Anxiety Toward Robots in Consumer Robotics: An Approach Using Item Response Theory*. https://doi.org/10.1109/ro-man57019.2023.10309588

Stanford Institute for Human-Centered Artificial Intelligence. (2019). Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/

Tabrez, A., Luebbers, M. B., & Hayes, B. (2020). A Survey of Mental Modeling Techniques in Human–Robot Teaming. *Current Robotics Reports*, *1*(4), 259–267. https://doi.org/10.1007/s43154-020-00019-0

Unhelkar, V. V. (2015). *Introducing mobile robots on the automotive final assembly line : control, sensing and human-robot interaction*. Dspace.mit.edu. https://dspace.mit.edu/handle/1721.1/98814

Van Barneveld, A., & Strobel, J. (2018). Engineering educators' perceptions of the influence of professional/industry experience on their teaching practice. *Proceedings of the Canadian Engineering Education Association (CEEA)*. https://doi.org/10.24908/pceea.v0i0.10221

Vázquez-Ingelmo, A., Alonso-Sánchez, J., García-Holgado, A., José, F., Jesús Sampedro-Gómez, Sánchez-Puente, A., Víctor Vicente-Palacios, P. Ignacio Dorado-Díaz, & Sánchez, P. L. (2021). Bringing machine learning closer to non-experts: proposal of a user-friendly machine learning tool in the healthcare domain. *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*. https://doi.org/10.1145/3486011.3486469

Wang, Y., & Wang, Z. (2013). Artificial Emotion Engine Benchmark Problem Based on Psychological Test Paradigm. *Telkomnika: Indonesian Journal of Electrical Engineering*, *11*(8). https://doi.org/10.11591/telkomnika.v11i8.3067

Willems, K., Verhulst, N., De Gauquier, L., & Brengman, M. (2022). Frontline employee expectations on working with physical robots in retailing. *Journal of Service Management*. https://doi.org/10.1108/josm-09-2020-0340

Zhang, C., Wang, Z., Zhou, G., Chang, F., Ma, D., Jing, Y., Cheng, W., Ding, K., & Zhao, D. (2023). Towards new-generation human-centric smart manufacturing in Industry 5.0: A systematic review. *Advanced Engineering Informatics*, *57*, 102121–102121. https://doi.org/10.1016/j.aei.2023.102121

Appendix A: Consensus Study

Start of Block: Introduction

Q8 **Participants' information sheet**
Before you decide to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take a couple of minutes to read the following information carefully. A member of the team can be contacted (see below) if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part in this consultation.

Purpose of the research
This research aims to contribute to the development of a new instrument to assess the experience of the users in the context of Human-Robot interaction and or Collaboration (HRI/HRC).

Results from the previous research phase
In the first consultation, we asked a group of 81 international experts on HRI/HRC about 15 dimensions that can be considered relevant in order to evaluate the User Experience in Human-Robot Interaction/Collaboration (7 robot-related aspects, 4 human-related aspects, and 4 context-related aspects). The prior survey led to a common consensus regarding 10 of the factors. However, it also revealed disagreement on 5 dimensions (2 robot-related aspects, 2 human-related aspects; 1 context-related aspect) which need further research.
In the following survey, we will ask you once again for your expert opinion on these 5 factors as well as ask some follow-up questions.
To have a look at the 15 dimensions and their descriptions, please click here and open the link in a new tab.

What we are asking you to do
In this second consultation, we would like you to look at the 5 dimensions that resulted in a disagreement to gain more insights about the usefulness of these aspects in assessing user experience with robots. Specifically, for each dimension, we will ask you to perform the following three actions:

- Mandatory: Rate how much form 1 (not important at all) to 9 (very important) do you believe that the dimension contributes (or it is important for) the evaluation of the user experience after the interaction/collaboration with a robot?
- Optional: Why do you think there is disagreement with a certain dimension?
- Optional: Is there a way the dimension can be improved?

Expected time for the survey
To perform the mandatory actions, we do not expect you to invest more than **15-20 minutes** of your time. Of course, if you would like to provide us with additional insights and suggestions by filling in the optional fields this might increase the time of your consultation.

How will we use your data?
Your participation to the present study is voluntary and you can decide to quit at any time.

Your personal data are going to be anonymised and used in the form of aggregated statistics for scientific purposes e.g., journal publications, conference presentations, etc. Only the researcher team will have access to your data and the data will be stored in a secure server in line with GDPR. This research project has been reviewed and approved by the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente. For questions regarding this study please contact the research team: Ásthildur Stefánsdóttir (a.l.stefansdottir@student.utwente.nl), Rufaro M. Hoto (r.m.hoto@student.utwente.nl) and Dr Simone Borsci (s.borsci@utwente.nl)

Consent **Consent form**
I have read and understood the participant information sheet above. I voluntarily consent to be a participant in this study and understand that I can refuse to answer questions, and I can withdraw from the study at any time, without having to give a reason. I understand that personal information collected about me will not be shared beyond the research team.

- ○ I understand and agree to participate voluntarily (1)
- ○ No, I would like to end this session (2)

End of Block: Introduction

Start of Block: Factors

Q13 In the next page we will show you individually each one of the 5 dimensions and their descriptions.

For each dimension we would like you to answer this question:
How much do you agree with the result of the previous consultation i.e., we should not consider this dimension among the main dimensions for assessing the UX after the interaction/collaboration with a robot?

Please, answer considering the dimensions and their descriptions and rate each factor on a 9-point scale ranging from 1 (Not important at all) to 9 (Extremely important).

Page Break

Q8 (Dimension name) **Robot's emotional appearance**
(Description) This aspect refers to how the robot's physical and behavioral characteristics, that delineate the "robot's emotional profile", can affect the user's judgment. In particular, it considers the following sub-factors: *Robot's Likeability* (e.g., happy, kind), *Warmth* (e.g., social, friendly), *Disturbance* (e.g., creepy, scary), *Discomfort* (e.g., awkward, dangerous), *Attractiveness*.

Results of previous consultation
Experts in the previous consultation **moderately disagree** about the relevance/importance of this dimension suggesting removing or do not consider such Dimension in the assessment of UX in HRI/HRC context.

Question:
How much do you agree with the result of the previous consultation i.e., we should not consider the <<Robot's emotional appearance>> among the main dimensions for assessing the UX after the interaction/collaboration with a robot?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	
Strongly disagree	○	○	○	○	○	○	○	○	○	Strongly agree

Q27 Why do you think there is disagreement with the importance of the factor: **Robot's emotional appearance**?

_____

Q25 Do you think there is a way that the description of the dimension **Robot's emotional appearance** could be improved?

_____

Page Break

Q5 **Robot's social behavior.**
This dimension is described as: the user's judgment of the robot's social behavior considering parameters such as e.g., *Companionship*, *Initiative* (e.g., not giving orders, not being intrusive), *Social relationship* (e.g., telling its story, having a real exchange of opinion), *Social norms* (e.g., no knowledge, full knowledge), *Communication*.

Results of previous consultation: Experts in the previous consultation **strongly disagree** about the relevance/importance of this dimension suggesting removing or do not consider such Dimension in the assessment of UX in HRI/HRC context.

Question: How much do you agree with the result of the previous consultation i.e., we should not consider the <<Robot's social behavior>> among the main dimensions for assessing the UX after the interaction/collaboration with a robot?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	
Strongly disagree	○	○	○	○	○	○	○	○	○	Strongly agree

Q28 Why do you think there is disagreement with the importance of this dimension: **Robot's social behavior**?

_____

Q26 Do you think there is a way that the factor **Robot's social behavior** could be improved?

_____

Page Break

Q9 **Human judgment before the interaction with a cobot.**
This dimension is described as: the user's perception of the robot before the interaction, based on. *Perception and effect*, *anxiety* (e.g., toward communication capability, toward behavioral characteristics), *Attitudes toward use*, *Expectation* (e.g., performance expectancy, effort expectancy), *Acceptance*, *Perceived safety* (e.g., speed), *Trust* (e.g., Reliability), *Intention to use*.

Results of previous consultation: Experts in the previous consultation **moderately disagree** about the relevance/importance of this dimension suggesting removing or do not consider such Dimension in the assessment of UX in HRI/HRC context.

Question: How much do you agree with the result of the previous consultation i.e., we should not consider the <<Human judgment before the interaction with a cobot>> among the main dimensions for assessing the UX after the interaction/collaboration with a robot?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	
Strongly disagree	○	○	○	○	○	○	○	○	○	Strongly agree

Q30 Why do you think there is disagreement with the importance of the factor: **Human judgment before the interaction with a cobot**?

_____

Q27 Do you think there is a way that the factor **Human judgment before the interaction with a cobot** could be improved?

_____

Page Break

Q10 **Human-Factors personality-based.**
This dimension is described as: the user's self- description regarding their own personality characteristics, like e.g., *ethics* (e.g., social impact, social acceptance), *Personality traits*, *Self-confidence*, and *Personality to trust.*

Results of previous consultation: Experts in the previous consultation **strongly disagree** about the relevance/importance of this dimension suggesting removing or do not consider such Dimension in the assessment of UX in HRI/HRC context.

Question: How much do you agree with the result of the previous consultation i.e., we should not consider the <<Human-Factors personality-based>> among the main dimensions for assessing the UX after the interaction/collaboration with a robot?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	
Strongly disagree	○	○	○	○	○	○	○	○	○	Strongly agree

Q29 Why do you think there is disagreement with the importance of the factor: **Human-Factors personality-based**?

_____

Q28 Do you think there is a way that the factor **Human-Factors personality-based** could be improved?

_____

Page Break

Q7 **Team involved during the task performance.**
This dimension is described as: the members involved in the specific task performed, considering e.g., *the number of humans and robots, Members' roles*.

Results of previous consultation: Experts in the previous consultation **moderately disagree** about the relevance/importance of this dimension suggesting removing or do not consider such Dimension in the assessment of UX in HRI/HRC context.

Question: How much do you agree with the result of the previous consultation i.e., we should not consider the <<Team involved during the task performance>> among the main dimensions for assessing the UX after the interaction/collaboration with a robot?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	
Strongly disagree	○	○	○	○	○	○	○	○	○	Strongly agree

Q31 Why do you think there is disagreement with the importance of the factor: **Team involved during the task performance**?

_____

Q30 Do you think there is a way that the factor **Team involved during the task performance** could be improved?

_____

End of Block: Factors

Start of Block: Personal information

TEXT PERSONAL INFO In order to better categorise your answers, please provide the following information:

Q28 Did you participate in the first phase of the research project?

- ○ Yes (1)
- ○ No (2)
- ○ Prefer not to say (3)

CONTRY In which country do you currently reside? *

▼ Afghanistan (1) ... Zimbabwe (1357)

SEX What is your sex (as assigned at birth)? *

- ○ Female (1)
- ○ Male (2)
- ○ Prefer not to answer (3)

ROLE How do you describe your job in relation to HRC? Multiple answers are possible. *

- [ ] Robotics engineer (who design & built robots) (1)
- [ ] Workstation layout designer (e.g, selection of layout based on the production requirements, selection of hardware) (2)
- [ ] Software developer (e.g., robot programming, controller programming & development) (3)
- [ ] Hardware designer (e.g., design of new components, integration of multi-brand instrumentation) (4)
- [ ] Robot assembly worker (e.g., assembly of robot's mechanical components) (5)
- [ ] Human-factors specialist (e.g., user interface designer, ergonomist, phycologist) (6)
- [ ] Researcher (please specify the research area) (7)
_____
- [ ] Robot user (i.e., if you use the robot for its final scope, e.g., assembly, physical support) (8)
- [ ] Other (please indicate) (9)
_____

EXPERTISE How many years of experience do you have in HRC? *

- ○ Less than 1 year (1)
- ○ From more than 1 to 5 years (2)
- ○ From more than 5 to 10 years (3)
- ○ More than 10 years (4)

DOMAIN In which HRC application domain(s) is your experience? Multiple answers are possible. *

- [ ] Cobot for Industry (1)
- [ ] Cobot for Warehouse (2)
- [ ] Cobot for Healthcare (3)
- [ ] Cobot for Domestic (4)
- [ ] Cobot for Entertainment (5)
- [ ] Cobot for Military and police (6)
- [ ] Cobot for Space expedition (7)
- [ ] Cobot for Surgery (8)
- [ ] Cobot for Social (e.g., waitress, information support) (9)
- [ ] Cobot for Education (10)
- [ ] Cobot for Agriculture (11)
- [ ] Other (please indicate): (12)

_____

ROBOT TYPE What type of robot(s) do you work on? If possible, please specify the robot model. Multiple answers are possible. *

- [ ] Collaborative robotic arm (1)
- [ ] Humanoid robot (2)
- [ ] Robot pet (3) _____
- [ ] Autonomous Mobile Robot (4)
- [ ] Automated Guided Vehicle (5)
- [ ] Unmanned aerial vehicles (6)
- [ ] Unmanned ground vehicles (7)
- [ ] Unmanned underwater vehicle (8)
- [ ] Toy (9) _____
- [ ] Other (please indicate) (10)

TYPE OF TASK Please provide an example of a task (e.g., assembly, physiotherapy) that the robot you are working with can perform:

_____

_____

_____

_____

_____

DESIGN FLOW In order to improve our research, can you briefly describe what activities you and your team carry out during a HRC design project?

_____

Q31 In case we will need to ask you additional questions we would like to have your contact. Do you agree to be contacted in the future? *

- ○ Yes (please write here your email) (1)

_____

- ○ No (2)

End of Block: Personal information

Appendix B: Email Communication of Second Round Delphi Study

Greetings,

We are reaching out to you once again as valued experts in the field of robotics for the second phase of our Delphi study.

As you may recall, our research project aims to create a new scale that assess the User eXperience (UX) during interaction with cobots.

In the previous phases of our study, we asked you to agree and comment about 15 dimensions we identified by a systematic literature review. The potential dimensions might affect the overall UX during human-robot interaction and collaborative tasks.

The result of the first round of consultation involved more than 100 worldwide HRC experts, we achieved a consensus on 10 out of the 15 dimensions, i.e., varying levels of disagreement among our panel of experts emerged.

Now, we wish to delve deeper into the potential underlying reasons for this divergence of perspectives. For this reason, we would like to invite you to participate in the next phase of our study, which aims to explore the rationales behind the differences in expert opinions regarding these five dimensions. Your insights and contributions will be very valuable in shedding light on these aspects and further advancing our understanding of UX in HRC scenarios.

Clinking the link below you will access the survey, and you will also find more explanation about the 15 Dimensions, and the agreement or lack of thereof per each dimension.

https://utwentebs.eu.qualtrics.com/jfe/form/SV_3aSbPOrRZmHyoui

Thank you for being an integral part of our study, and we look forward to your participation.

Best regards,

Research team:
Ásthildur Stefánsdóttir, Rufaro M. Hoto and Dr. Simone Borsci

Appendix C: R Code For Study 1

```
# Install and load necessary packages
install.packages("tidyverse")
install.packages("lmtest")
install.packages("ggplot2")
install.packages("syuzhet")
install.packages("dplyr")
library(tidyverse)
library(lmtest)
library(ggplot2)
library(syuzhet)
library(dplyr)

# Load the data
data <- read.csv("/Users/rufarohoto/Downloads/UX-Consensus-Study-Data.csv", sep = ";",
header = TRUE)

# Selecting relevant columns for dimensions D3, D6, D8, D10, D14
dimensions <- c("D3", "D6", "D8", "D10", "D14")

# Calculating descriptive statistics
descriptive_stats <- data %>%
  select(all_of(dimensions)) %>%
  summary()

# Adding IQR
iqr <- apply(data[dimensions], 2, IQR)
descriptive_stats <- rbind(descriptive_stats, "IQR" = iqr)

# Percentage of agreement (scores > 5)
percentage_agreement <- colMeans(data[dimensions] > 5) * 100
descriptive_stats <- rbind(descriptive_stats, "Percentage Agreement" =
percentage_agreement)

#Boxplots
# Reshape the data to long format
long_data <- data %>%
  pivot_longer(cols = c("D3", "D6", "D8", "D10", "D14"),
               names_to = "Dimension",
               values_to = "Score")

# Create a combined box plot for all dimensions
ggplot(long_data, aes(x = Dimension, y = Score)) +
  geom_boxplot() +
  labs(title = "Box Plots for All Dimensions", y = "Scores", x = "Dimension") +
  theme_minimal()

# Convert expertise to numeric categories for regression
```

```
data$Experience_Category <- as.numeric(factor(data$EXPERTISE))

# Linear regression for each dimension and extracting p-values
regression_p_values <- list()
for (dim in dimensions) {
  formula <- as.formula(paste(dim, "~ Experience_Category"))
  model <- lm(formula, data = data)
  summary_model <- summary(model)
  regression_p_values[[dim]] <- summary_model$coefficients[2, "Pr(>|t|)"]  # Extracting the
p-value for the Experience_Category predictor
}

# Viewing the p-values
regression_p_values




# Define the columns containing open-ended questions
open_ended_columns <- c("Thoughts_On_Why__There_Is_Disagreement_On_D3",
                        "Thoughts_On_Why__There_Is_Disagreement_On_D6",
                        "Thoughts_On_Why__There_Is_Disagreement_On_D8",
                        "Thoughts_On_Why__There_Is_Disagreement_On_D10",
                        "Thoughts_On_Why__There_Is_Disagreement_On_D14")

# Initialize a list to store results
sentiment_results <- list()

# Loop through each column, perform sentiment analysis, and count positive/negative
comments
for (col in open_ended_columns) {
  # Select non-NA and non-empty comments
  valid_comments <- data %>%
    select(all_of(col)) %>%
    filter(!is.na(!!sym(col)) & !!sym(col) != "")

  num_comments <- nrow(valid_comments)  # Count comments

  # Proceed if comments are available
  if (num_comments > 0) {
    # Extract comments for sentiment analysis
    comments <- valid_comments[[1]]

    # Get sentiment scores using syuzhet
    sentiments <- get_sentiment(comments, method = "syuzhet")

    # Categorize as positive or negative based on sentiment polarity
    positive_count <- sum(sentiments > 0)
    negative_count <- sum(sentiments < 0)
  } else {
```

```
    # If no comments, set counts to zero
    positive_count <- 0
    negative_count <- 0
  }

  # Store results
  sentiment_results[[col]] <- list("Number of Comments" = num_comments,
                                   "Positive" = positive_count,
                                   "Negative" = negative_count)
}

# Viewing the results
sentiment_results
```

Appendix D: List of Generated Items by Research Team and Approved by HRC Experts

The numbers are the associated dimension numbers D1 – D15.

1. **Easiness of robot regulation**
   Item 1. It was easy to physically regulate the [system/collaborative robot name].
   Item 2. I found it/ it appears to be easy to position the [System/collaborative robot name] components correctly.
   Item 3. I believe that from a physical point of view it appears to be easy to manipulate and put the [System/collaborative robot name] into position.

2. **Robot physical appearance.**
   Item 4. I had a positive impression of the [System/collaborative robot name]'s physical appearance.
   Item 5. I had a positive impression of the [System/collaborative robot name]'s dimensions i.e., high, width, length, weight.
   Item 6. I had a positive impression of the [System/collaborative robot name]'s features e.g., form, material.
   Item 7. The level of anthropomorphism (machinelike or humanlike) of the [System/collaborative robot name] was appropriate for the intended purpose.
   Item 8. The [System/collaborative robot name]'s perceived robustness (e.g., its ability to withstand physical stress, challenges etc.) met the specific requirements for the task and context of usage.
   Item 9. The type of robot (e.g., Robotic Arm, Humanoid Robot) seems appropriate for the task and context of usage.

3. **Robot's emotional appearance.**
   Item 10. I believe that the [System/collaborative robot name]'s emotional appearance was likable/attractive.
   Item 11. I believe that the [System/collaborative robot name]'s design does not cause emotional discomfort.
   Item 12. I believe that the [System/collaborative robot name]'s behavior does not cause emotional discomfort.
   Item 13. I did not consider the [System/collaborative robot name]'s appearance disturbing.
   Item 14. I felt that the [System/collaborative robot name] displayed a sense of warmth during our collaboration e.g., it was social, friendly.

4. **Robot's competence features.**
   Item 15. I perceived the [System/collaborative robot name] as competent and smart in terms of behavior.
   Item 16. I believe the [System/collaborative robot name] is reliable and trustworthy in terms of competencies.
   Item 17. I found the [System/collaborative robot name] to be responsive and transparent in terms of competencies.

5. **Robot's physical behavior.**
   Item 18. I perceived the [System/collaborative robot name] movements to be smooth and flexible.
   Item 19. I believe that [System/collaborative robot name] is (physically) adaptable and autonomous.
   Item 20. I believe that the [system/collaborative robot name]'s physical behavior (e.g., noise, movement, autonomy, interactivity) during the interaction was suitable.
   Item 21. I believe that the [System/collaborative robot name] movements and behavior seemed lifelike and natural.

6. **Robot's social behavior.**
   Item 22. I believe that the overall social behavior of the [System/collaborative robot name] was appropriate.
   Item 23. I believe that the [System/collaborative robot name] acted and communicated according to social norms.
   Item 24. I felt that the [System/collaborative robot name] engaged in meaningful social interactions during our collaboration.
   Item 25. I think the [System/collaborative robot name] gave me a sense of companionship during our collaboration.
   Item 26. I perceived the [System/collaborative robot name] to be intrusive during our collaboration.

7. **Robot task performance.**
   Item 27. I believe that the [System/collaborative robot name] was useful by enabling a timely efficient performance.
   Item 28. I believe that the [System/collaborative robot name] was useful by enabling a correct (without error) performance.
   Item 29. I believe that the collaboration with the [System/collaborative robot name] was useful by enabling tasks to be completed in an efficient and effective manner.

8. **Human judgment before the interaction with a cobot (collaborative robot).**
   Item 30. I expected the collaboration with the [System/collaborative robot name] to be safe before using it.
   Item 31. I expected the [system/collaborative robot name] to be reliable and quick prior to collaborate with it.
   Item 32. I did not experience any anxiety related to the [System/collaborative robot name] prior to the collaboration with it.
   Item 33. My overall attitude towards using the [System/collaborative robot name] was positive prior to our collaboration.
   Item 34. I accepted the idea of using the [System/collaborative robot name] for the task prior to use.
   Item 35. Before interacting with the [System/collaborative robot name], I had the intention to use it for similar tasks or interactions in the future.

9. **Human judgment of the performance with a cobot** (collaborative robot)
   Item 36. I realized while collaborating with the [System/collaborative robot name] that it is safe and trustworthy in use.
   Item 37. I realized while collaborating with the [System/collaborative robot name] that it was pleasing to use and easy to control during the task.
   Item 38. I felt comfortable during my collaboration with the [System/collaborative robot name].
   Item 39. After using the [System/collaborative robot name], I found myself accepting of its role in the collaboration.
   Item 40. I was highly satisfied with the [System/collaborative robot name]'s performance.
   Item 41. I found the [System/collaborative robot name]'s interface or interaction methods (e.g., touch panel, voice commands, haptic feedback) highly usable.
   Item 42. I experienced no frustration while working with the [System/collaborative robot name].
   Item 43. I felt calm (e.g., no stress) during the interaction with the [System/collaborative robot name].
   Item 44. I perceived an appropriate amount of cognitive workload during the collaboration with the [System/collaborative robot name].
   Item 45. I would like to use the [System/collaborative robot name] again based on my experience during the task.

10. **Human-Factors personality-based.**
    Item 46. I feel confident in my ability to use the [System/collaborative robot name] to achieve key tasks.
    Item 47. I feel a sort of natural tendency to align well with the [System/collaborative robot name] during the collaboration to achieve certain goals.

Items 48. I do not see any ethical, personal, or social issues in collaborating with the [System/collaborative robot name] for my job.
Item 49. I have a trusting personality.
Item 50. I believe my personality traits have a significant influence on my collaboration with robots in general.

11. **Human-Factors ability-based.**
Item 51. My level of expertise contributed to a successful interaction with the [System/collaborative robot name].
Item 52. I believe that my general understanding of robotics contributed to a positive collaboration with the [System/collaborative robot name].
Item 53. I believe that I have enough competence using collaborative robots to be able to properly handle the [System/collaborative robot name].

12. **Task performed.**
Item 54. I believe that I can do this task more efficiently without the [System/collaborative robot name].
Item 55. The task I performed with the [System/collaborative robot name] did not require too much physical effort.
Item 56: I felt that the task I performed with the [System/collaborative robot name] did not require too much mental effort.
Item 57: I believe the [System/collaborative robot name] was useful in accomplishing the task.
Item 58: I felt that it was important to use the [System/collaborative robot name] to perform this critical task.

13. **The environment of interaction.**
Item 59. The environmental conditions (e.g., lighting, noise, dust) were disturbing the task.
Item 60. The workstation layout facilitated a positive interaction with the [System/collaborative robot name].
Item 61: The interaction took place in a workstation with a layout that facilitated the completion of the task.
Item 62. The elements present in the workstation during the interaction were suitable to achieve the task.
Item 63. The [System/collaborative robot name] is well-suited for the demands of the applicational context.

14. **Team involved during the task performance.**
Item 64. I believe that it is possible for multiple operators (a team) to collaborate proficiently to use the [System/collaborative robot name] to achieve the task.

Item 65. When multiple operators (a team) have to collaborate interacting with the [System/collaborative robot name] all the operators can understand their roles.

15. **Interaction aspects.**

Item 66. I was pleased with the overall interaction with the [system/collaborative robot name].

Item 67. I had good knowledge of the [System/collaborative robot name] status during the task performance.

Item 68. I found the [System/collaborative robot name]'s interface and interaction modality easy to use.

Item 69. I found the [System/collaborative robot name]'s interface and interaction modality easy to learn.

Item 70. I found the [System/collaborative robot name]'s interface and interactions modality easy to remember.

Item 71. I found the type of interface used for interaction (e.g., physical-based, graphical-based, vocal-based, gesture-based) was appropriate and effective.

Appendix E: Card Sorting Survey

Introduction

Welcome to this Card Sorting study, and thank you for agreeing to participate!

Task Overview:
First you will be presented with a practise task, where you can get a feel for how the sorting in the experiment will work.

Then you will be presented with 10 items, and your goal is to categorize each of them into one of the 15 groups or, if you believe an item doesn't belong in any of the provided categories, you can place it into a special group labeled "None of the above."

The activity shouldn't take longer than 15 to 20 minutes to complete.

Your participation is anonymous and confidential. Your responses will be used solely for research purposes and will not be linked to your personal information.

Consent form
I voluntarily consent to be a participant in this study and understand that I can refuse to answer questions, and I can withdraw from the study at any time, without having to give a reason. I understand that personal information collected about me will not be shared beyond the research team.

- ○ I understand and agree to participate voluntarily
- ○ No, I would like to end this session

Practise sorting

THIS IS A PRACTISE ROUND AND WILL NOT COUNT.

Task Description:
- On the lower side of the screen, you will find 15 boxes of dimensions/groups that are relevant to the assessment of the user eXperience (UX) when operating collaborative robots.

- On the top of the screen, you will see two items related to this assessment, and your task is to sort the items into the dimension that best aligns with the provided statement.

Sorting Guidelines:
- Drag and drop each item into the dimension that you believe is the most appropriate based on the statement provided. More than one item can go into the same group.
- If you feel an item doesn't fit into any of the 15 dimensions, you can move it to a special group labeled "None of the above."
- You can make adjustments if you realize you initially placed an item in the wrong dimension.

Notes:
- There are no right or wrong sorting choices; trust your judgment and go with what feels correct to you.

Items

Try to put this item under the dimension "Robot's competence features"

Try to put this item under the dimension "NONE of the ABOVE"

Easiness of robot regulation	Robot's physical appearance
Robot's emotional appearance	Robot's competence features
Robot's physical behavior	Robot's social behavior

Robot task performance
Human judgment before the interaction with a cobot
Human judgment of the performance with a cobot
Human-Factors personality-based	Human-Factors ability-based
Task performed	The environment of interaction

Team involved during the task performance	Interaction aspects
NONE of the ABOVE	

Default Question Block

Please drag and drop these 10 items into the category you think is most appropriate. More than one item can go into the same category.

If you hover your mouse over the title of each dimension you will see a description of the dimension. This will be very helpful during the sorting (it can take a few seconds to show up). If you would rather see a full list of the dimensions and their descriptions, click here.

If you feel that the dimensions or the items are not clear in any way, please take note of it. You will be able to express your opinions if you feel there is a lack of clarity on the next page of the survey.

Items

I believe that robot is (physically) adaptable and autonomous

I believe that I can do this task more efficiently without the robot

I had good knowledge of the robot status during the task performance

Easiness of robot regulation	Robot physical appearance

I felt calm (e.g., no stress) during the interaction with the robot

I believe that the robot's behavior does not cause emotional discomfort.

I found the robot's interface and interactions modality easy to remember

I expected the collaboration with the robot to be safe before using it

I have a trusting personality

I realized while collaborating with the robot that it was pleasing to use and easy to control during the task

I believe that from a physical point of view it appears to be easy to manipulate and put the robot into position.

Robot's emotional appearance	Robot's competence features
Robot's physical behavior	Robot's social behavior
Robot task performance	
Human judgment before the interaction with a cobot	
Human judgment of the performance with a cobot	

Qualtrics Survey Software 05/01/2024, 21:45

Participant information

What is your age?

What is your sex (as assigned at birth)?

- Female
- Male

Do you have any previous experience with collaborative robotics?

- Yes
- No
- Im not sure
- Prefer not to say

(Optional) Were the items and dimensions clear to you? If not, please explain why.

Appendix F: R Code Card Sorting Demographic Data

```
install.packages("readr")
# Load the required library
library(readr)

# Read the CSV file
data1 <- read.csv("/Users/rufarohoto/Downloads//Demographic_Data_Card_Sorting.csv", sep
        = ";", header = TRUE)

# View the first few rows of the dataframe to verify its contents
head(data1)

# Validate the data
# Check if all expected columns are present
expected_columns <- c("AGE", "SEX", "EXPERIENCE")
if(!all(expected_columns %in% names(data1))) {
  stop("One or more expected columns are missing")
}

# Transforming 'SEX' column: 1 for male, 2 for female
data1$SEX <- ifelse(data1$SEX == 1, "Male", "Female")

# Transforming 'EXPERIENCE' column: 1 for yes, 2 for no, 3 for 'I'm not sure', 4 for 'Prefer
        not to say'
data1$EXPERIENCE <- factor(data1$EXPERIENCE, levels = 1:4, labels = c("Yes", "No",
        "I'm not sure", "Prefer not to say"))
```

```
summary(data1$AGE)
table(data1$SEX)
table(data1$EXPERIENCE)
```