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

Comparison of Rating Scale and Pairwise Comparison Methods for Measuring Human Co-worker Subjective Impression of Robot during Physical Human-Robot Collaboration

Qiao Wang[1,*], Ziqi Wang[1,*], Marc G. Carmichael[1], Dikai Liu[1], and Chin-Teng Lin[2]

***Abstract*— The *Rating Scale* method has been long deemed the standard for measuring subjective perceptions. However, in the field of physical human-robot collaboration (pHRC), its aptness should be put under scrutiny due to inherent challenges such as response bias, between-subject variations, and the granularity nature.**

Individual variances can introduce significant bias in the rating scale results. A high granularity in the scale could overwhelm participants, leading to unclear and biased responses, while a low granularity may gloss over the fine nuances of human feelings. Additionally, there's a notable risk of receiving careless responses, which compromise data reliability. Recognizing these challenges, this paper proposes the application of *Pairwise Comparison (PC)* in pHRC — an alternative survey technique that emphasizes direct comparisons between items on the defined criteria. By using the NASA Task Load Index (NASA-TLX) as a template, RS and PC questionnaires are designed and used in a series of pHRC experiments. Our preliminary findings suggest that PC is more precise and robust than the rating scale method. Compared to RS, PC fosters authentic participant interests in the experiment by intuitive question design and reducing the experimental duration. Besides, the accuracy and reliability of PC are also found to be consistent regardless of the variations in our experimental procedure design.

I. Introduction

With the emergence of Industry 4.0, physical human-robot collaboration (pHRC) using collaborative robots (cobots) has become increasingly popular. As a result, considerations of how the human user perceives and actions to ensure the collaboration is desirable should be taken when developing cobot controls. Unfortunately, the perceptions of human users are often neglected, for example, [1]–[5]. This may be attributed to the lack of standardized procedures on how to conduct subjective evaluations and the difficulty in measuring nuanced differences in human feelings.

Rating Scale (RS), for example the Likert scale, is widely used to evaluate the user impressions of robotic systems [6]–[8]. It is simple to implement as it just requires participants to provide scores against some defined attributes or criteria on a given scale. However, criticism of this method in relation to its limitations has been persistently suggested. Kieruj [9] has discovered that the length of response scales influences the responses. It is also addressed that the cultural factor results in differences in scale completion rates and familiarity with scales [10], [11]. Another challenge associated with Rating Scale is the risk of careless or disengaged responses. Participants who are not fond of the survey or find it tedious might rush through the questions, leading to haphazard results with diminished reliability and validity. All of these limitations of the Rating Scale method lead to bias and noise in its results.

An alternative approach that addresses the aforementioned deficiencies of the RS method is Pairwise Comparison (PC). This method requires the participant to choose between two given options the one that better expresses some characteristic. By counting the number of wins of each option or through sophisticated mathematical models, a ranking that reflects participants' preferences can be obtained. PC has several advantages over Rating Scale: potential response bias due to the rating scales can be avoided. PC is less cognitively demanding because the choices do not require an understanding of any numbered scales, and subjects do not need to track responses from previous decisions [12]. Furthermore, PC allows short experiment designs. For example, Carmichael [2] used PC to rank users' preferences for 6 pHRC control settings by asking a large cohort of participant to individually compare between 2 settings only.

Although both RS and PC have been applied in human-robot interaction studies, when typically utilizing them in pHRC, a side-by-side systematic comparison between them has never been carried out. In this paper, we fulfill this research gap by using both questionnaire styles in a series of pHRC experiments to validate the reliability and robustness of the results against a ground truth. To ensure our results are statistically significant, we employed the one-way analysis of variance (ANOVA). Moreover, Tukey's HSD multiple comparison test also applied to allow us to determine specifically between which groups the differences existed.

The remainder of this paper is structured as the following: In Section II a review of applications of different questionnaire styles in pHRC studies is provided. Section III presents the experimental protocol. In Section IV the experimental results and extensive analysis are provided. Discussions and conclusions are presented in Section V and Section VI respectively.

* indicates equal contributions.

[1] Qiao Wang, Ziqi Wang, Marc G. Carmichael and Dikai Liu are with the Robotics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, Ultimo NSW 2007, Australia. Email: {qiao.wang-1; ziqi.r.wang}@student.uts.edu.au; marc.carmichael@uts.edu.au

[2] Chin-Teng Lin is with the Australian Artificial Intelligence Institute, School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway, Ultimo NSW 2010, Australia
Email: chin-teng.lin@uts.edu.au

II. Applications of Rating Scale and Pairwise Comparison in pHRC

A. Rating Scale

Rating scale assessment is widely used by robotic researchers to explore the views of participants regarding robots' appearance, interaction experience and overall satisfaction [13]. Commonly used rating scale questionnaires include NASA Task Load Index (TLX) [14], Negative Attitude towards Robot Scale (NARS) [15], and Godspeed Questionnaire [16].

NASA Task Load Index evaluates subjective workload from six dimensions: *Mental Demand, Physical Demand, Temporal Demand, Performance, Effort* and *Frustration*. A 21-point scale is used where 0 means Very Low (workload) and 21 means Very High (workload). By calculating the weighted average of ratings on these 6 subscales, the overall workload of the task can be derived [14]. NASA TLX was applied in [7] and [17] where in the former the cognitive workload during a heavy object manipulation task and the user's satisfaction with the proposed control schemes was evaluated. In the latter, ratings of human performance, robot performance, rushedness and calmness levels during exercising were evaluated for a rehabilitation robot.

Other aforementioned Rating Scale questionnaires have also had applications in pHRC studies [15], [18]. As these are pre-experiment questionnaires that investigate how users' perceived judgments about the robot affect the results of the experiment, it is beyond the scope of this paper, therefore such type of questionnaire is not considered in this paper.

B. Pairwise Comparison

Pairwise Comparison is performed by simply asking participants to conduct a series of comparisons on the specified criteria between pairs of options. A ranking of all the options can be obtained by using all of the comparison responses together. In the context of pHRC, PC methods have seen relatively little utilization. Carmichael [2] used Pariwise Comparison to study users' preferences for different pHRC control algorithms. The comparison results were processed with the Bradley-Terry (BT) model to rank users' favorite damping settings of the designed singularity handling method where the difference between the tested damping settings was narrow. [19], [20] also made use of the Pairwise Comparison method to obtain users' preferences between slightly different robot gaze behaviors during a handling task. Depending on the criteria that researchers want to investigate, the questions in the Pairwise Comparison questionnaire should be customized.

III. Experiment Protocol

A. The Clock Game

The Clock Game is shown in Fig. 1. The position of the cobot end-effector is projected onto the TV through coordinate frames transformation and it is represented by the white dot. The red dot is the moving target. It starts from the top left corner of the yellow rectangular trajectory and moves in a clockwise direction at a constant speed. The task of the experiment is to control the white dot to follow the red moving target as closely as possible. To stimulate the engagement and focus of participants during the experiment, a competitive element is added to the task. A circular boundary is designed to move synchronously with the target dot along the trajectory. The experimental subjects are asked to keep the white dot within the boundary and the radius of the circular boundary shrinks over time, making the task more challenging and requiring higher concentration. Based on how long the participant accommodates the white dot within the boundary, the *Score* is calculated in real-time at $45Hz$ and displayed to the participant. While the white dot is accommodated within the boundary, one point is assigned to the *Score*. Contrarily, it will be penalized by 10 points.

Fig. 1. The experimental task requires participants to follow the red target dot with the white dot while remaining within the red circular boundary. The task becomes more intense as the radius of the boundary reduces with time. A *Score* is calculated based on the time that the participant could stay within the boundary. It increases if the participant is able to remain within the boundary. Otherwise, it reduces.

B. Collaborative Robot (Cobot) Setup

The cobot used in this experiment is the ANBOT which is a physically collaborative robot system designed for industrial abrasive blasting [21]. The ANBOT consists of the 6-DoF Universal Robot UR10 manipulator with a custom-made handle at its end-effector that incorporates a 6-DoF force/torque sensor. The ANBOT is controlled by a mass-damping admittance controller [22] as shown in (1). $M_d \in \mathbb{R}^{6\times6}$ and $D_d \in \mathbb{R}^{6\times6}$ are the virtual inertia and damping matrix respectively; $\dot{x} \in \mathbb{R}^{6\times1}$ is the desired velocity of the end-effector in the Cartesian space and $\ddot{x} \in \mathbb{R}^{6\times1}$ is the desired acceleration; $F \in \mathbb{R}^{6\times1}$ is the collaborative wrench applied to the cobot end-effector by the human operator, it is measured directly by the force/torque sensor; $\tau \in \mathbb{R}^{6\times1}$ is the artificial noise deliberately injected into F.

$$M_d\ddot{x} + D_d\dot{x} = F + \tau \tag{1}$$

Considering the task introduced in the section above, the ANBOT is restricted to planar motions by setting the corresponding elements in M_d and D_d to the desired values. For F and τ, only those elements that contribute to the planar motion are used by the admittance control law (1).

C. Artificial Noise Design

In (1), the artificial noise τ is used to establish the ground truth for the experiment based on the hypothesis that a noise

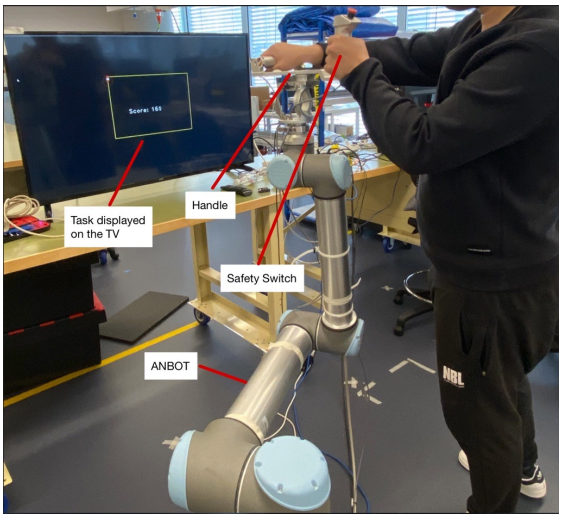


Fig. 2. A participant conducting the experiment using the custom-made handle attached at the end-effector of ANBOT. The safety switch is used to allow the participant to activate ANBOT and control the start of the experiment.

added to F will deteriorate the user experience: the larger the noise is, the worse the experience should be.

τ is randomly generated every 0.5s and injected into a random quadrant of F. The magnitude of τ is randomized following the Gaussian distribution between the lower limit L and the upper limit U. Four noise levels are specified through experimental trials based on the principle that the difference between two successive noise levels is so subtle that a pHRC expert cannot distinguish them based on trials:

- τ_1: L=0N, U=0N. *No disturbance.*
- τ_2: L=1N, U=2N. *Little disturbance.*
- τ_3: L=3N, U=4N. *Medium disturbance.*
- τ_4: L=5N, U=6N. *Large disturbance.*

Throughout the entire experience, the noise levels used are kept unbeknownst to participants.

D. Questionnaires Design

The NASA TLX questionnaire is used in this experiment. With considerations of the experiment length and the relevancy of the original NASA TLX questions to our experiment design, the following two questions and the standard 21-point scale [14] are used in the experiment:

- *RS Q1: How physically demanding was the task?*
- *RS Q2: How frustrated did you feel? E.g. insecure, discouraged, irritated, stressed, and annoyed were you?*

For the PC questionnaire, we created the following questions by referring to the NASA TLX questionnaire:

- *PC Q1: Which mode required less physical demand?*
- *PC Q2: Which mode was less frustrating to use?*

Participants are asked to compare the two noise levels given to them on a 5-point scale and points are assigned to the noise levels accordingly:

- **A$\gg$B:** Performance of A is much better than B. 2 points for A, 0 point for B.
- **A$>$B:** Performance of A is a little better than B. 1 point for A, 0 point for B.
- **A=B:** Performance of A and B is about the same. 0 point for both A and B.
- **A$<$B:** Performance of B is a little better than A. 0 point for A, 1 point for B.
- **A$\ll$B:** Performance of B is much better than A. 0 points for A, 2 points for B.

The first noise level that the participant is given is marked as A and the subsequent one is marked as B.

E. Experimental Procedure

Two experimental procedures are designed which correspond to conducting all the experimental conditions and several experimental conditions (Experimental procedures 1 and 2). Both RS and PC questionnaires are carried out in the following experiments:

1) Experimental Procedure 1: For the first experimental procedure, participants need to complete all the combinations of comparisons, that is: τ_1&τ_2, τ_1&τ_3, τ_1&τ_4, τ_2&τ_3, τ_2&τ_4 and τ_3&τ_4 with single-order matchups (e.g, A versus B and no B versus A). Using balanced Latin Square [23], six distinctive sequences of conducting the comparisons are generated to reduce the order effect.

For the NASA TLX questionnaire, participants take trials of the noise levels in the sequence generated by the balanced Latin Square above and give scores after every trial.

From this experimental procedure, twelve NASA TLX results and six pairwise comparison results are collected from each participant. The order of completing the questionnaires is also alternated. For example, if the previous participant completes the NASA TLX questionnaire first, the subsequent participant will be given the PC questionnaire first.

2) Experimental Procedure 2: In the second experimental procedure, the experimental sequence generated in Experimental Procedure 1 is used likewise. Except, participants are only required to conduct the experiment 4 times for either questionnaire instead of 12. As a result, only 2 PC questionnaire results and 4 NASA TLX results are collected. The order of completing the questionnaires is also altered.

F. Hypothesis

It is reasonable to anticipate that participants' subjective sensations for physical demand and frustration will be ranked from the best to the worst in line with the magnitude of the noise: τ_1 should result in the least physical demand and frustration, followed by τ_2, τ_3 and τ_4 in sequence. Based on this presumption, the following hypotheses are made for different experimental procedures:

- *Hypothesis 1*: For Experimental Procedure 1, both questionnaire methods should return the correct result that matches the anticipation. The NASA TLX results should show that τ_1 has the lowest scores for physical effort demand and frustration, with the scores increasing for τ_2, τ_3 and τ_4 sequentially. For the PC results, τ_1 should receive the most points, followed by τ_2, τ_3 and τ_4 in order.

- *Hypothesis 2*: For Experimental Procedure 2, it is hypothesized that only the PC results will match with the anticipation. NASA TLX will fail to rate the noise levels correctly because of inter-rater variations and insufficient data from each participant.

G. Participant Recruitment:

The experiment was performed under the University of Technology Sydney (UTS) Human Research Ethics Committee Approval ETH18-3029. Participants are recruited through personal connections and incidental engagements. Participants are given more details about the experimental task through an information sheet and verbal explanation. A demonstration of the experiment is also provided. After this, participants are given the consent form to read and sign.

The experiment was performed at the UTS Open Day (OD) and Orientation Week (OW) with age group between 16 and 40. During these events, a cohort of 36 inexperienced volunteers (16 females and 20 males) who had no prior pHRC experience were recruited. Nine of them followed Experimental Procedure 1 and the rest performed Experimental Procedure 2. In addition, eighteen volunteers with pHRC experience (2 females and 16 males) were recruited within the UTS Robotics Institute (RI) in the weeks following and they were split in half for the experimental procedures. For the 18 participants who completed Experimental Procedure 1, their first 8 experimental runs and the corresponding questionnaire results are extracted and combined with the Experimental Procedure 2 results. For either experimental procedure, the balanced Latin Square was completed in full for 3 times.

Before commencing the experiment, the questionnaire questions are shown to participants and they will practice with ANBOT under τ_1 (no artificial noise) until the score reaches 500 as shown in Figure 1 to guarantee the same proficiency level of starting the experiment to reduce the learning effect between individuals.

IV. Results

For the NASA TLX results, the mean and standard deviation of each noise level are computed from participants' rated scores directly, and the results are presented in Fig.3a and Fig.4a for Experimental Procedure 1 and Experimental Procedure 2 respectively. From the PC questionnaire results, points for the noise levels are obtained from the comparison results as per Section III-D. The mean and standard deviation of points are computed, and the results are shown in Fig. 3b and Fig. 4b for two experimental procedures respectively.

One-way analysis of variance (ANOVA) is carried out to test whether there was a statistically significant difference in the means between different noise level scores. The *p*-value tests whether the true difference of means for two noise levels is equal to zero, in other words, whether two noise levels resulted in similar levels of physical demands and frustration in our experiment. Typically, a *p*-value < 0.05 indicates a significant difference between the means of two groups [24]. The results are shown in Table I.

A. Hypothesis 1: Both RS and PC will return agreeing and statistically significant results in Experimental Procedure 1

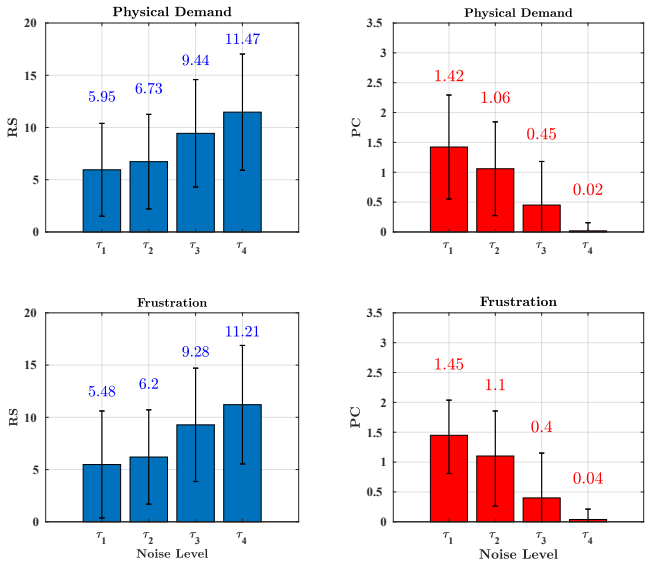
Fig. 3 shows the results of Experimental Procedure 1 and is used to test Hypothesis 1. The RS mean scores are presented in Fig.3a. For both Physical Demand and Frustration, RS returned a result agreeing with our anticipation. The average scores are rated the lowest for τ_1 and start to climb for τ_2, τ_3 and τ_4. This indicates that participants found τ_1 the least physically demanding and resulted in the least frustration, followed by τ_2, τ_3 and τ_4 in sequence. From the second and fourth columns of Table I, the *p*-values for τ_1 vs τ_3, τ_1 vs τ_4, τ_2 vs τ_3, and τ_2 vs τ_4 can be viewed to be less than 0.05, indicating RS showed statistically significant differences between these noise levels in the Physical Demand and Frustration categories. However, the *p*-values for τ_1 vs τ_2 and τ_3 vs τ_4 are all *greater* than 0.05 (cells highlighted in yellow) which are statistically insignificant, implying that within these pairs, participants found these noise levels resulted in similar levels of physical demand and frustration.

As can be seen in Fig. 3b, PC has returned perfectly agreeing results for Physical Demand and Frustration. The average points obtained by τ_1 for the two tested criteria are substantially higher than τ_2 by 25.6% and 23.9% respectively. The average points obtained by τ_3 are much lower compared to τ_1 and τ_2, and they are approaching 0 for τ_4. These reviews that participants found τ_1 resulted in the least physical demand and frustration, followed by τ_2, τ_3 and τ_4 in sequence. To test the statistical significance of the PC results, *p*-values in the third and fifth columns of Table I should be studied. As can be seen, all the *p*-values in these columns are all than the standard significance level. For the noise level pairs where RS failed to differentiate (τ_1 vs τ_2 and τ_3 vs τ_4), PC has successfully shown the statistically significant differences between them (cells highlighted in green).

These results provide evidence to reject the null hypothesis 1: RS results agree with our anticipation shown in Figure 3a. However, it is not statistically significant. On the other hand, PC was capable of returning statistically significant results that matched our anticipation.

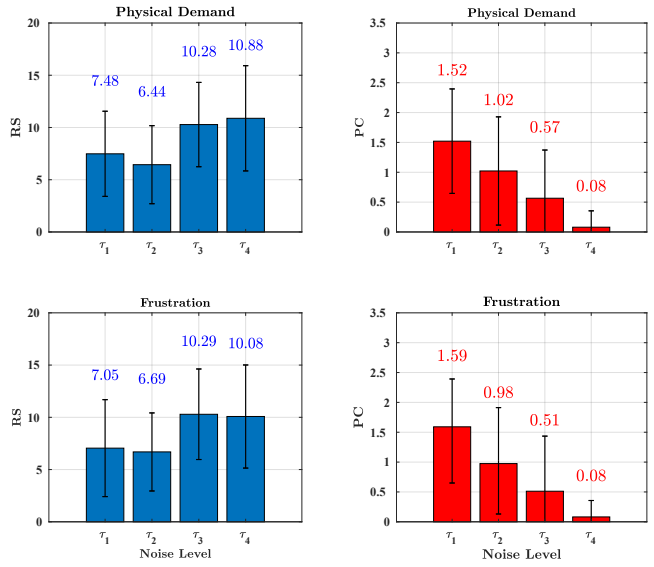
B. Hypothesis 2: PC can return statistically significant and agreeing results under Experimental Procedure 2 and RS will fail

The Experimental Procedure 2 results are shown in Fig. 4 and they are used to test Hypothesis 2. From Fig. 4a, it can be observed that RS has performed poorly for both the Physical Demand and Frustration categories. For the former, the error happened between τ_1 and τ_2, the RS average scores show participants found τ_1 more physically demanding than τ_2. For Frustration, noise levels were rated from the best performance to the worst performance in the order of τ_2, τ_1, τ_4 and τ_3. These results violate the outcomes of Experimental Procedure 1 as well as our anticipation. Moreover, the sixth and the eighth columns of Table I show the statistical significance. Similar to the Experimental Procedure 1 results,



(a) Rating Scale (b) Pairwise Comparison

Fig. 3. Experimental Procedure 1 results for (a) Rating Scale Questionnaire (b) Pairwise Comparison Questionnaire



(a) Rating Scale (b) Pairwise Comparison

Fig. 4. Experimental Procedure 2 results for (a) Rating Scale Questionnaire (b) Pairwise Comparison Questionnaire

the mean scores are tested to be statistically insignificant for τ_1 vs τ_2 and τ_1 vs τ_2.

On the other hand, as shown in Fig.4b, PC persisted in relating participants' perceptions with the noise levels correctly. The smallest noise level achieved the highest average point; as larger and larger noises were used, lower and lower points were obtained by them. Besides, the PC results are validated to be statistically significant as per Columns 7 and 9 in Table I. All the *p*-values are less than the standard significance level including the noise levels with subtle differences (τ_1 vs τ_2 and τ_3 vs τ_4 as highlighted in green).

Experimental Procedure 2 shows RS fails to return the correct and statistically significant results when the number of repetitions for the experiment is limited for each participant. At the same time, PC accuracy endures. Therefore, Hypothesis 2 should be accepted.

C. Experimental Duration

In Experimental Procedure 1, participants spent an average of *25 minutes* to complete all the NASA TLX questionnaires. It should be noted that this length included 3 repetitions of the noise levels. The average completion time of the PC questionnaire is *14 minutes*.

V. Discussion and Future Work

In this paper, we proposed the utilization of the Pairwise Comparison approach in pHRC studies and carried out a preliminary study to compare the results of PC to the more popular Rating Scale method. Through testing our hypotheses, the following key advantages of PC over RS can be identified:

1) When the tested characteristics are repeated multiple times, both RS and PC mean values provide accurate measurements of participants' subjective impressions. However, only PC results are statistically significant in determining the differences for the characteristics with subtle differences. RS results are not statistically significant.
2) When the length of the experiment is reduced, PC mean values rate human subjective feelings accurately and are statistically significant. RS mean values are noisy and fail to reflect human's true subjective perceptions.

A potential cause of the poor results of RS may be the absence of a mutual understanding of the rating system between participants. Individual participant needs to establish their own rating standards by retrospectively comparing the current result to their previous responses when conducting any rating scale questionnaires [25]. When the number of repeating the experiment is limited as in Experimental Procedure 2, experimental subjects fail to establish the rating standard. Therefore, they provided noisy and inaccurate results using RS. Furthermore, cultural influences [10] [11], the length of the RS scale [9] and careless responses may have also contributed to the poor results of RS.

A limitation associated with PC is the complexity of experiment design. To compare n characteristics, PC requires a total of $(n-1)^2/2+(n-1)/2$ comparisons for a single-order match-up comparison while RS only requires a minimum of n experiments. However, as validated in Experimental Procedure 2, the accuracy of PC does not rely on individual participants completing all the comparisons. The robustness is particularly useful for pHRC studies because the experi-

TABLE I

COMPARISON OF RATING SCALE AND PAIRWISE COMPARISON METHOD BY EMPLOYING THE P-VALUE FROM THE ONE-WAY ANOVA FOLLOWED BY TUKEY'S HSD MULTIPLE COMPARISON TESTS FOR EXPERIMENTAL PROCEDURE 1 AND 2.

Noise Level	Experimental Procedure 1 Physical Demand		Experimental Procedure 1 Frustration		Experimental Procedure 2 Physical Demand		Experimental Procedure 2 Frustration	
	RS	PC	RS	PC	RS	PC	RS	PC
τ_1 vs τ_2	0.822166	0.037120	0.872085	0.027475	0.575070	0.007382	0.972555	0.001132
τ_1 vs τ_3	0.000503	0.000000	0.000304	0.000000	0.002536	0.000000	0.000531	0.000000
τ_1 vs τ_4	0.000000	0.000000	0.000000	0.000000	0.000063	0.000000	0.000955	0.000000
τ_2 vs τ_3	0.013759	0.007265	0.006597	0.000000	0.000014	0.019474	0.000112	0.030415
τ_2 vs τ_4	0.000001	0.000000	0.000001	0.000000	0.000000	0.000000	0.000202	0.000000
τ_3 vs τ_4	0.113456	0.007265	0.179922	0.016261	0.862868	0.008956	0.993072	0.036399

mental task is often repetitive and experimental subjects will get fatigued both mentally and physically after many repetitions. PC reduces the workload of experimental subjects more effectively than RS. However, the increased number of experiments required by PC and its high robustness lead to higher recruitment costs because more experimental subjects are needed.

The main limitation of this study is we only utilized one questionnaire as the template, NASA-TLX. There are various types of questionnaires, such as trust [26] and satisfaction [27] to evaluate different aspects of subjective impression. it may raise the question of whether or not the results presented can be broadly generalized to other types of questionnaires. However, we did not believe that applying different types of questionnaires would have a significant impact on the experimental results. From similar studies in other disciplines [12] [28] where the effectiveness of different Rating Scale assessments and Pairwise Comparisons are compared, results similar to ours were obtained.

Another limitation of the study is that only 4 noise levels were designed in this study. More noise levels could be investigated in future research to validate the robustness and accuracy. This study involved 54 volunteers which is an acceptable number for the study. However, a larger number of experimental subjects with variations in genders, age groups, educational backgrounds and cultural backgrounds is desirable as it will refine inter-rater variations and validate the robustness of the results.

The outcomes of this preliminary research show the potential of applying PC to other practical pHRC scenarios. For example, using PC to investigate human operators' experience in industrial tasks such as pick-and-place and grit-blasting; Comparing pHRC control methods, such as the singularity avoidance algorithms in [29] and the role arbitration method in [22]. Moreover, it is also of our interest to generalize the method of applying PC in human-robot interaction (HRI). This requires more types of HRI experiments to be conducted, such as Remote Control Interaction and social interaction. Also, the research outcome demonstrate significant insight in cobot design, including software and hardware, the predictability of the cobot behvaiour is critical to subjective impression, for instance, whether the damping and inertia of admittance/impedance model is desired for the human co-worker.

VI. CONCLUSION

In this paper, a novel comparison between two subjective impression evaluation methods is conducted. The classic NASA-TLX questionnaire was used as the template to create the Rating Scale and Pairwise Comparison questionnaires, respectively. Through two experimental procedures of different lengths, the effectiveness of RS and PC was compared statistically. The experimental results show statistically significant results for both RS and PC when the difference of comparing groups is large. However, when the true difference is small, PC is more robust and accurate than RS regardless of the experimental procedure. Furthermore, compared to the RS, PC also reduced the experimental duration by 44%, resulting in a more enjoyable experimental experience and lower cognitive workloads A summarized comparison between Rating Scale and Pairwise Comparison is shown in Table II. Even though the cost of recruitment is higher for PC, it still offers a more time-efficient, accurate and robust alternative method to pHRC researchers who are interested in capturing nuances in human perceptions during pHRC studies.

TABLE II

COMPARISON BETWEEN RATING SCALE AND PAIRWISE COMPARISON METHODS

	Rating Scale	Pairwise Comparison
Effectiveness with repetition (Experiment 1)	**No**	**Yes**
Effectiveness without repetition (Experiment 2)	No	**Yes**
Averaged Experimental length	25 minutes	**14 minutes**
Robustness	Low	**High**
The level of satisfaction in answering questions	Low	**High**
The level of satisfaction in conducting the experiments	Low	**High**
Recruitment cost	**Low**	High

VII. ACKNOWLEDGMENTS

This work was supported in part by the Australian Research Council (ARC) under discovery grant DP210101093.

REFERENCES

[1] H. Xing, A. Torabi, L. Ding, H. Gao, Z. Deng, V. K. Mushahwar, and M. Tavakoli, "An admittance-controlled wheeled mobile manipulator for mobility assistance: Human–robot interaction estimation and redundancy resolution for enhanced force exertion ability," *Mechatronics*, vol. 74, p. 102497, 2021.

[2] M. G. Carmichael, R. Khonasty, S. Aldini, and D. Liu, "Human preferences in using damping to manage singularities during physical human-robot collaboration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10184–10190, 2020.

[3] Q. Wang, D. Liu, M. Carmichael, S. Aldini, and C.-T. Lin, "Computational model of robot trust in human co-worker for physical human-robot collaboration," *IEEE Robotics and Automation Letters*, pp. 1–1, 2022.

[4] B. Navarro, A. Cherubini, A. Fonte, G. Poisson, and P. Fraisse, "A framework for intuitive collaboration with a mobile manipulator," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6293–6298, 2017.

[5] K. Wakita, J. Huang, P. Di, K. Sekiyama, and T. Fukuda, "Human-walking-intention-based motion control of an omnidirectional-type cane robot," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 1, pp. 285–296, 2013.

[6] E. Rosen, D. Whitney, E. Phillips, G. Chien, J. Tompkin, G. Konidaris, and S. Tellex, "Communicating robot arm motion intent through mixed reality head-mounted displays," in *Robotics Research* (N. M. Amato, G. Hager, S. Thomas, and M. Torres-Torriti, eds.), (Cham), pp. 301–316, Springer International Publishing, 2020.

[7] S. M. M. Rahman and R. Ikeura, "Calibrating intuitive and natural human–robot interaction and performance for power-assisted heavy object manipulation using cognition-based intelligent admittance control schemes," *International Journal of Advanced Robotic Systems*, vol. 15, no. 4, p. 1729881418773190, 2018.

[8] Y. Chen, C. Yang, Y. Gu, and B. Hu, "Influence of mobile robots on human safety perception and system productivity in wholesale and retail trade environments: A pilot study," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 624–635, 2022.

[9] N. D. Kieruj and G. Moors, "Variations in Response Style Behavior by Response Scale Format in Attitude Research," *International Journal of Public Opinion Research*, vol. 22, pp. 320–342, 07 2010.

[10] C. H. Hui and H. C. Triandis, "Effects of culture and response format on extreme response style," *Journal of Cross-Cultural Psychology*, vol. 20, no. 3, pp. 296–309, 1989.

[11] J. W. Lee, P. S. Jones, Y. Mineyama, and X. E. Zhang, "Cultural differences in responses to a likert scale," *Research in Nursing & Health*, vol. 25, no. 4, pp. 295–306, 2002.

[12] A. P. Clark, K. L. Howard, A. T. Woods, I. S. Penton-Voak, and C. Neumann, "Why rate when you could compare? using the "elochoice" package to assess pairwise comparisons of perceived physical strength," *PLOS ONE*, vol. 13, pp. 1–16, 01 2018.

[13] G. Noury, M. Tsekeni, V. Morales, R. Burke, M. Palomino, and G. Masala, *Experiment Protocol for Human–Robot Interaction Studies with Seniors with Mild Cognitive Impairments*, pp. 243–253. 01 2021.

[14] N. Aeronautics and S. Administration, "Nasa tlx: Task load index," 2020.

[15] T. Nomura and T. Kanda, "On proposing the concept of robot anxiety and considering measurement of it," pp. 373 – 378, 01 2003.

[16] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, 01 2008.

[17] N. Fitter, M. Mohan, K. Kuchenbecker, and M. Johnson, "Exercising with baxter: Preliminary support for assistive social-physical human-robot interaction," *Journal of NeuroEngineering and Rehabilitation*, vol. 17, 02 2020.

[18] R. Flook, A. Shrinah, L. Wijnen, K. Eder, C. Melhuish, and S. Lemaignan, "On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based hri experiments trustworthy?," *Interaction Studies*, vol. 20, pp. 455–486, 11 2019.

[19] A. Kshirsagar, M. Lim, S. Christian, and G. Hoffman, "Robot gaze behaviors in human-to-robot handovers," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6552–6558, 2020.

[20] Z. Minhua, M. AJung, E. A. Croft, and M. Q. . Meng, "Impacts of robot head gaze on robot-to-human handovers," *International Journal of Social Robotics*, vol. 7, pp. 783–798, 11 2015. Copyright - © Springer Science+Business Media Dordrecht 2015; Last updated - 2020-07-09.

[21] M. G. Carmichael, S. Aldini, R. Khonasty, A. Tran, C. Reeks, D. Liu, K. J. Waldron, and G. Dissanayake, "The anbot: An intelligent robotic co-worker for industrial abrasive blasting," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8026–8033, 2019.

[22] Q. Wang, D. Liu, M. G. Carmichael, and C.-T. Lin, "Robot trust and self-confidence based role arbitration method for physical human-robot collaboration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9896–9902, 2023.

[23] A. L. Edwards, "Balanced latin-square designs in psychological research," *The American Journal of Psychology*, vol. 64, no. 4, pp. 598–603, 1951.

[24] F. Ferraguti, C. T. Landi, L. Sabattini, M. Bonfè, C. Fantuzzi, and C. Secchi, "A variable admittance control strategy for stable physical human–robot interaction," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 747–765, 2019.

[25] U. Bockenholt, "Thresholds and intransitivities in pairwise judgments: A multilevel analysis," *Journal of Educational and Behavioral Statistics*, vol. 26, pp. 269–282, 09 2001.

[26] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[27] J. R. Lewis, "Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *International Journal of Human–Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

[28] A. S. Phelps, D. M. Naeger, J. L. Courtier, J. W. Lambert, P. A. Marcovici, J. E. Villanueva-Meyer, and J. D. MacKenzie, "Pairwise comparison versus likert scale for biomedical image assessment," *American Journal of Roentgenology*, vol. 204, no. 1, pp. 8–14, 2015. PMID: 25539230.

[29] M. G. Carmichael, D. Liu, and K. J. Waldron, "A framework for singularity-robust manipulator control during physical human-robot interaction," *International Journal of Robotics Research*, 2017.