Developing and Evaluating Ecologically Valid Study Tasks to Investigate Human-Robot Teaming

Jasmin J. Chadha, Patrick K. Pischulti, Ciara Hume, Zachary Selleck, William Roessmann, and Katya Arquilla
University of Colorado Boulder Smead Aerospace Engineering Sciences, 3775 Discovery Dr., Boulder, CO, USA 80303

Background & Motivation

Gap 1: Characterizing Psychological Safety

Psychological Safety: "the shared belief held by members of a team that the team is safe for interpersonal risk taking" [3]

→ Studied extensively in large human-human teams and associated with higher personal engagement and performance

→ **Gap in studying psychological safety in operational environments and human-robot interaction (HRI)**

Gap 2: Current Methods in Study Tasks

Researchers often employ computer-based tasks, simplified physical tasks, or completely in-the-wild tasks [2,4,5]

Ecological validity: the degree to which proximal cues in the laboratory are correlated to the environment [1]

→ Study setup elicits responses directly relevant to intended environments

→ **Gap in development of ecologically valid study tasks that are both controlled and relevant to operational environments**

Objective

Develop and evaluate **ecologically valid tasks** that:
1) mimic spaceflight surface **operations** and
2) can be implemented in an experiment investigating team dynamics constructs that contribute to **psychological safety** in **human-robot collaboration**

Task Development & Study Methods

Task Development Process

Study Task Criteria	
Ecological Validity	Psychological Safety
Derived from real human spaceflight surface operation tasks	Induces interpersonal risk-taking
Mimics parameters of spaceflight surface operations	Induces some form of creativity
Feasible to execute in the laboratory	Has quantifiable performance metrics
Mimics human-robot team dynamics during operations	Can be accomplished in collaboration with a robotic agent

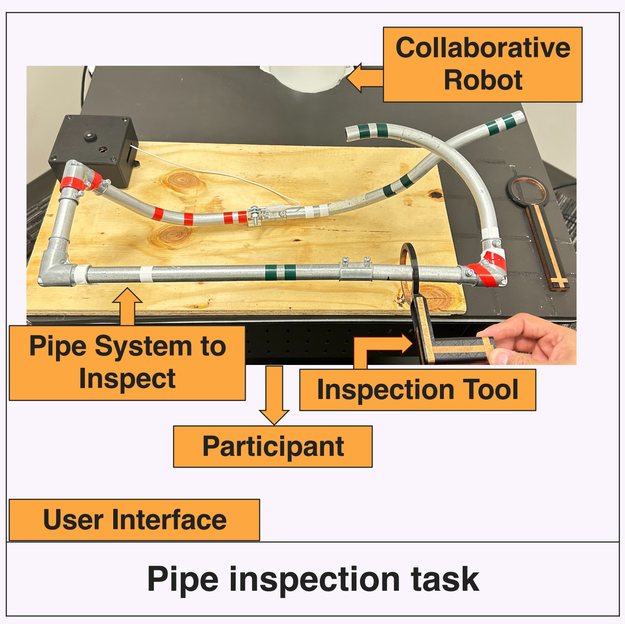
Experimental Design

Within-participants (n = 33)	**Gender** = 15 female, 16 male, 2 transgender **Sex assigned at birth** = 16 female, 17 male **Mean age** = 29.7±10.9 years

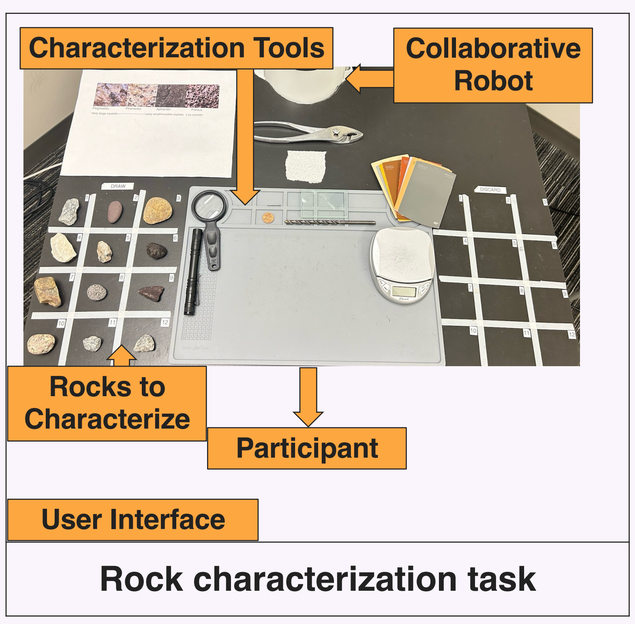
12 trials (3 minutes each)

6 pipe inspection trials (randomized)		6 rock characterization trials (randomized)	
3 trials w/ human	3 trials w/ robot	3 trials w/ human	3 trials w/ robot

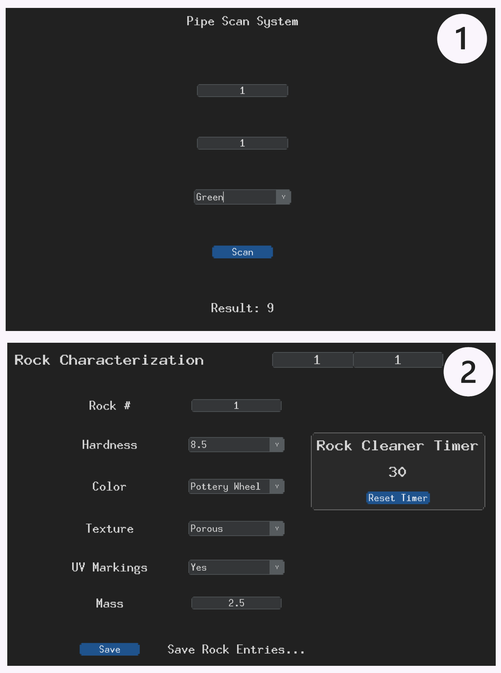
Task Setup



Pipe inspection task



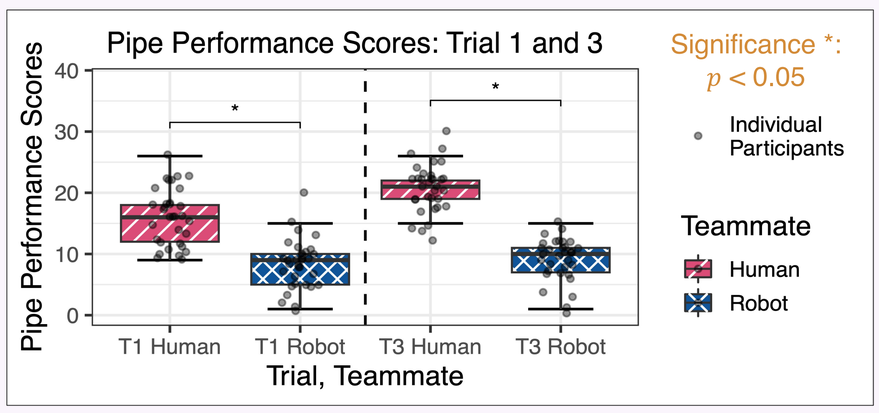
Rock characterization task



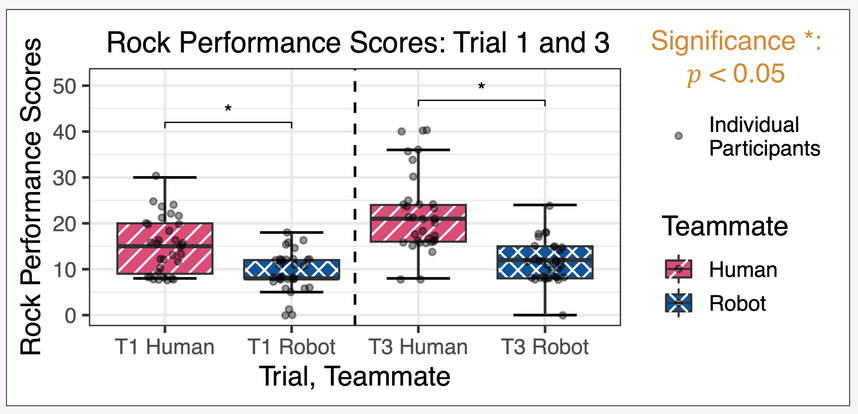
User interface for
1) pipe task and 2) rock task

Robot teammate: Franka Research 3
7 degrees of freedom robotic arm

Task Evaluation: Results



Task performance scores from Trial 1 and Trial 3 for pipe tasks.



Task performance scores from Trial 1 and Trial 3 for rock tasks.

Repeated Measures ANOVA Results: Task performance as it relates to teammate type and trial number		
	Pipe Inspection Task	Rock Characterization Task
Main effect: Teammate	$F_{(1,32)} = 307.16$, $p < 0.05$, $\eta^2_p = 0.88$	$F_{(1,32)} = 86.37$, $p < 0.05$, $\eta^2_p = 0.80$
Main effect: Trial	$F_{(1.69,54.23)} = 11.32$, $p < 0.05$, $\eta^2_p = 0.26$	$F_{(1.66,53.09)} = 28.01$, $p < 0.05$, $\eta^2_p = 0.47$
Interaction effect: Teammate x Trial	$F_{(1.96,62.80)} = 7.63$, $p < 0.05$, $\eta^2_p = 0.19$	$F_{(1.87,59.99)} = 6.87$, $p < 0.05$, $\eta^2_p = 0.18$

Post-hoc results: Significant differences between earlier and later trials for human pipe, human rock, and robot rock combinations
Significant differences between human and robot performance scores across trials 1-3 for pipe and rock tasks

Pipe Performance Scores Over Trials 1-3		
Trial	Human Score (M±SD)	Robot Score (M±SD)
1	16.12 ± 4.68	8.33 ± 3.97
2	19.15 ± 4.89	8.82 ± 3.21
3	20.64 ± 3.88	9.12 ± 3.45

Rock Performance Scores Over Trials 1-3		
Trial	Human Score (M±SD)	Robot Score (M±SD)
1	15.18 ± 6.12	9.45 ± 4.37
2	19.33 ± 7.46	11.39 ± 4.21
3	22.15 ± 8.74	11.61 ± 4.48

Discussion

Participants demonstrated higher performance scores when working with a human vs. robot teammate.

Unequal capabilities of teammates + 82% of participants having less experience with robots may explain lower performance with robot.

Performance scores increased over trials (more for human), indicating possible learning effects.
→ Implications for task ordering

Conclusion & Future Work

These **collaborative**, **ecologically valid** study tasks that are **operationally relevant** to human spaceflight elicit significant performance differences that may be relevant to **human-human and human-robot team dynamics.**

Analyze teaming questionnaires, physiological data, and behavioral data

Equalize capabilities between active teammates

References

[1] E. Brunswik. 1955. Representative design and probabilistic theory in a functional psychology. *Psychological Review* 62, 3 (1955), 193–217.
[2] E. Carsenti, A. Manor, A. Oberlender, A. Parush, and H. Erel. 2025. Raising Stars: Influences of Robotic Peer Liking on Emergent Leadership. *Proc. ACM/IEEE HRI '25,* 447–457.
[3] A. C. Edmondson. 1999. Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly* 44, 2 (June 1999), 350–383.
[4] J. Sung, S. Leary, V. S. Hurd, C. Lee, Y. Qin, Z. Kong, T. K. Clark, and A. Anderson. 2024. Operationally Realistic Human-Autonomy Teaming Task Simulation to Study Multi-Dimensional Trust. *ACM/IEEE HRI Companion '24*, 1028–1032.
[5] Y. Hu, M. Benallegue, G. Venture, and E. Yoshida. 2020. Interact With Me: An Exploratory Study on Interaction Factors for Active Physical Human-Robot Interaction. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6764–6771.

Video

Contact & Connect

Acknowledgements

The authors would like to thank the Gerald A. Soffen Memorial Fund for supporting conference travel to present this work.