



Article

Collaborative System Usability in Spaceflight Analog Environments through Remote Observations

Shivang Shelat ¹ , Jessica J. Marquez ^{2,*}, Jimin Zheng ¹ and John A. Karasinski ² 

¹ San Jose State University Research Foundation, San Jose State University, San Jose, CA 95192, USA; sshelat@ucsb.edu (S.S.); jimin.zheng@nasa.gov (J.Z.)

² NASA Ames Research Center, Mountain View, CA 94035, USA; john.karasinski@nasa.gov

* Correspondence: jessica.j.marquez@nasa.gov

Abstract: The conventional design cycle in human–computer interaction faces significant challenges when applied to users in isolated settings, such as astronauts in extreme environments. Challenges include obtaining user feedback and effectively tracking human–software/human–human dynamics during system interactions. This study addresses these issues by exploring the potential of remote conversation analysis to validate the usability of collaborative technology, supplemented with a traditional post hoc survey approach. Specifically, we evaluate an integrated timeline software tool used in NASA’s Human Exploration Research Analog. Our findings indicate that voice recordings, which focus on the topical content of intra-crew speech, can serve as non-intrusive metrics for essential dynamics in human–machine interactions. The results emphasize the collaborative nature of the self-scheduling process and suggest that tracking conversations may serve as a viable proxy for assessing workload in remote environments.

Keywords: human-in-the-loop; user interaction development methodology; collaborative work; conversation analysis; usability evaluation; self-scheduling; remote observation



Citation: Shelat, S.; Marquez, J.J.; Zheng, J.; Karasinski, J.A. Collaborative System Usability in Spaceflight Analog Environments through Remote Observations. *Appl. Sci.* **2024**, *14*, 2005. <https://doi.org/10.3390/app14052005>

Academic Editor: João M. F. Rodrigues

Received: 7 December 2023

Revised: 8 February 2024

Accepted: 15 February 2024

Published: 28 February 2024



Copyright: © This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The conventional design cycle in human–computer interaction (HCI) involves understanding a specific domain, gathering feedback from end users, refining prototypes through usability testing, and ultimately releasing a product developed with significant end-user participation [1]. This process, however, faces significant access challenges when applied to user populations in isolated, confined environments. Our research revolves around enabling human space exploration, with astronauts as our user population. Space is an extreme environment inhabited by only a select few who live in spaceships like the International Space Station (ISS). Astronaut crews heavily rely on numerous computer systems to support critical tasks and ensure mission success during spaceflight. The advanced technologies used in space missions demand meticulous human-in-the-loop design and development, but achieving this is challenging due to the limited number of astronauts, the very limited time they have to be part of the HCI design process, and their presence in remote and inaccessible environments.

To address some of these limitations, NASA has invested resources in developing spaceflight analogs. Analog missions simulate aspects of spaceflight; for instance, crews may experience isolation, an extreme environment, or interact in a physical environment resembling space. These analog missions recruit participants with backgrounds similar to astronauts for simulated long-duration space missions. Analogous user populations provide an opportunity to not only refine prototypes and conduct usability evaluations but also allow exploration of new methods to apply HCI in the spaceflight domain. This paper summarizes how our research team conducted remote usability observations in a

spaceflight analog for an integrated timeline software tool used in operations. We demonstrate that conversation analysis of naturalistic voice recordings significantly contributes to characterizing the usability of collaborative software.

2. Background

2.1. Remote Observations through Conversation Analysis

Usability is defined as how easy and feasible it is to use an interface to accomplish tasks [1]. The think-aloud protocol for testing usability is one of the most widely used methods in HCI design, allowing for the assessment of a user's thought processes and decisions in real time [2]. Participants are required to verbalize their thoughts while using the system under evaluation. This method's popularity stems from its simplicity, cost-effectiveness, and flexibility. However, like all usability research techniques, the think-aloud protocol is not without drawbacks. In the context of crewed space missions, limitations of think-aloud evaluation for software become particularly pronounced. The necessity for verbal collaboration, unnatural conditions created by the think-aloud methodology, the likelihood of participants opting for selected verbalizations instead of the preferred stream-of-consciousness, and potential biases all contribute to limiting the effectiveness of the think-aloud protocol. These constraints often make its application in a space environment less than ideal, prompting the exploration of alternative methods.

Recent studies have found that think-aloud protocols significantly benefit from additional evaluation in analog contexts. For example, Li et al. [3] paired think-aloud testing with a novel, near-live simulation approach. Their user population, primary care providers relying on clinical decision support tools, interacted with simulated patients to provide necessary care. Transcripts and behavioral analyses from these sessions showed where, when, and how providers accessed the support tool to efficiently provide the necessary services; they also revealed workflows that conventional think-aloud testing never caught. Morgan and colleagues [4] assessed the usability of a patient-care guide by coupling think-aloud testing with mock patient discussions. Post hoc interviews and transcript content analyses showed that accessing the guide in "practice" was hindered by several obstacles. From these conclusions, the authors made significant revisions to the guide to make it easier for providers to find relevant information. Testing in analogous settings emerges as a valuable strategy in usability science, especially for tools used during interpersonal interactions and complex workflows.

A recent framework presented by Clinkenbeard [5] outlines a methodology for studying usability by coupling conversation analysis with traditional approaches, such as user experience (UX) questionnaires. This framework asserts that conversation analysis, the study of social interactions through audio and/or video data, can effectively validate the usability of collaborative technologies in complex, social environments. Such environments often carry an intricate ecology of human–human and human–computer interactions that must be characterized with a targeted data collection approach. For example, Nicolini [6] uncovered several issues associated with novel medical software that became only evident when considering users' (i.e., healthcare employees) social interactions. Their approach used focus group interviews, individual interviews, and naturalistic patient observations over 24 months to inform the design and development of patient-oriented technologies for years to come. Other empirical studies have successfully characterized computer-supported cooperative work specifically using the conversation analysis method in many fields, including public transportation coordination [7] and telemedicine [8]. As a proposed cornerstone of this framework, voice recordings in particular [9–11] have garnered much recent attention as promising enhancements to usability evaluation. Work by Li and colleagues [12] suggests that trust in autonomous agents can be inferred from lexical and acoustic features in speech. Similarly, Magnusdottir et al. [13] coupled cardiovascular measures and speech to measure the dynamics of cognitive workload during various tasks (see also [14]). Voice recordings may serve as a non-intrusive measure of critical dynamics in human–machine interactions. In essence, empirical research underscores the value of in-

tegrated audio recordings in capturing nuanced dynamics of human–computer interactions and collaborative work.

Our team has chosen to expand the application of conversation analysis to spaceflight analogs. In these analog environments, crew members spend extended periods together in isolation, often tasked with solving problems that demand coordination. Moreover, analog crews undergo close monitoring like astronauts in space, as analogs serve both as a research platform and a simulation of spaceflight. Analog missions are equipped with various video and audio recordings that can be leveraged to collect remote observations. This facilitates the collection of data while crews engage in problem-solving tasks using novel software. Past work has utilized voice data from a spaceflight analog setting to infer team cohesion based on utterances and vocal microbehaviors [11], which underscores the feasibility of using analog audio to explore the dynamics of collaborative HCI tasks. Consequently, we advocate for conversation analysis as a valuable addition to conventional usability methods, especially when studying remote populations that pose challenges in accessibility.

2.2. Integrated Timeline Software

Our HCI team at NASA Ames Research Center has developed *Playbook*, an integrated timeline software designed for future astronauts [15] (see hci.arc.nasa.gov/work/playbook.html [accessed on 8 February 2024] for more details). *Playbook* serves as a “one-stop shop”, enabling future astronauts to schedule and execute assigned activities and tasks. Its primary goal is to better support Earth-independent operations and enhance crew autonomy, thereby reducing the crew’s reliance on ground stations to organize daily tasks during communication delays in deep space. NASA envisions future astronauts conducting self-scheduling to manage, schedule, and reschedule their timelines, enabling more autonomous crew. *Playbook* has been deployed in various analogs to support mission operations. Specifically, the team has been evaluating *Playbook* for the task of self-scheduling [16]. Analog crews utilize *Playbook* to create and manage complex mission timelines, incorporating crewmate preferences (e.g., the flight engineer’s preference for exercising after lunch) and meeting various constraints (e.g., task completion before 11:00 a.m.) [17]. Additionally, analog crew members rely on *Playbook* to complete operationally relevant tasks. In previous analog missions, the team has requested crews to perform self-scheduling using *Playbook*. We have used traditional HCI methods, post hoc asking analog crews to fill out surveys and participate in debriefs which provide valuable insights into their experiences using *Playbook*.

Self-scheduling with *Playbook* in an analog environment provides a research opportunity for conversation analysis, applying Clinkenbeard’s [5] framework to augment our previous laboratory-based approaches [18] in validating the tool for future long-duration exploration missions.

3. Methods

3.1. Analog Mission—Study Overview

Our research was conducted in the Human Exploration Research Analog (HERA) at NASA Johnson Space Center (JSC) from September 2021 to March 2023. HERA, a 650-square-foot isolation analog, simulates future long-duration exploration missions. HERA is split between two floors and a loft, designed to replicate conditions of isolation, confinement, and remote scenarios in exploration missions. NASA’s Human Research Program sets specific research goals and selects research projects to be completed for each HERA Campaign. In each of HERA Campaign 6’s (C6) four missions, four astronaut-like crew members lived in the habitat for 45 days and “embarked” on a mission to the Martian moon Phobos without physically leaving Earth. Throughout these missions, the crew exclusively engaged in virtual interactions with their family, friends, and Mission Control. While conducting tasks in isolation, these analog crews were closely monitored by research teams to gain insights into behavior, health, and human–systems integration. Despite

this, collecting usability and user evaluations in these missions remained challenging as researchers were not physically present and had to collect data remotely.

In HERA C6, crew autonomy was operationalized as the crews' ability to independently manage their own operational timelines. The missions underwent distinct phases:

- No Autonomy (first seven days): During this initial period, all timelines were fixed and could not be edited by the crew.
- Limited Autonomy (next eight days): The subsequent eight days were considered "limited autonomy", allowing the crews to reschedule many activities.
- High Autonomy (last thirty days): The final thirty days were called "high autonomy", where crews' self-scheduling was entirely initiated by crew without external constraints.

This phasic shift in autonomy over time aligns with expectations for long-duration space exploration missions. In our research experiment, analog astronauts were specifically asked to self-schedule four of their days, as opposed to just following the timeline provided by Mission Control. For detailed procedures and the experimental design used in the C6 scheduling experiment, refer to Marquez et al. [19].

During the limited autonomy phase, our team requested that each crew member lead a team preference meeting (TPM) and participate in a self-scheduling session (SS). Crew members were responsible for scheduling their operational timelines using Playbook (v13; see Figure 1). In the TPM, the assigned planner led an open discussion on timeline preferences. The crew was expected to discuss items such as "lots of free time in the afternoon" or "hygiene periods in the morning". In the SS, the planner independently created the operational timeline, aiming to integrate the team preferences into a feasible schedule. Participants were allocated thirty minutes for their TPM and one hour to complete their SS. Following the SS, crew members completed the NASA-Task Load Index (NASA-TLX; [20]), a commonly used workload questionnaire in usability studies. In HCI, workload is defined as the demands of a task on an operator's mental and physical resources—developers often desire lower levels of workload to avoid users' cognitive overload. Finally, the crew executed the scheduled day, and the planner provided feedback through a brief questionnaire on their experience.

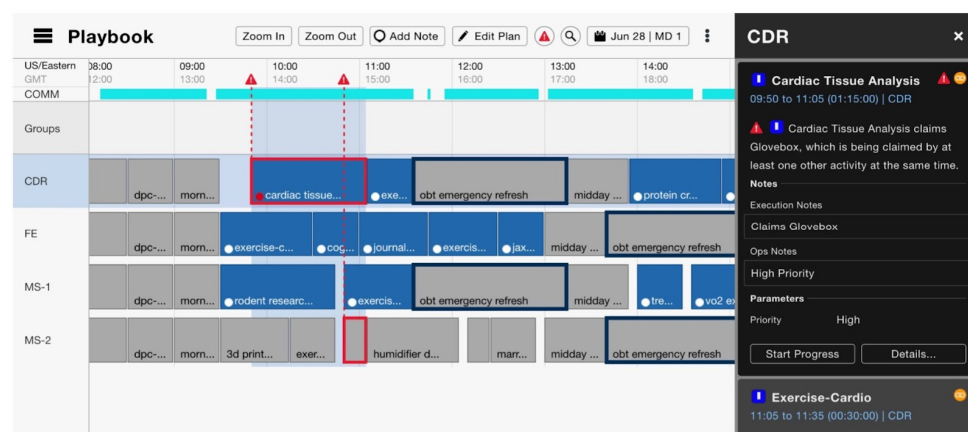


Figure 1. A screenshot of Playbook's user interface. Schedule violations are highlighted so the user can make adjustments until a timeline is feasible for the whole crew.

At the end of their mission, crews completed the User Experience Questionnaire (UEQ; [21]) to evaluate Playbook. The UEQ is a 26-item scale that breaks down UX into six distinct factors: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. Scores for each of these subscales were computed using a provided calculator (www.ueq-online.org [accessed on 8 February 2024]). The goal with the UEQ was to quantify the usability of our tool in the analog and roughly compare it with prior UEQ scores from lab-based testing [18,22]. Because usability questionnaires are commonly used in

human–systems integration research, we sought to supplement our proposed conversation analysis by presenting scores for Playbook in a simulated spaceflight environment.

It is crucial to note that HERA’s controlled setting and simulation demands necessitate a highly selective and small participant pool to accurately emulate space missions. Consequently, the limited sample size ($n = 16$; 4 females; $M_{\text{age}} = 35.93$, $SD_{\text{age}} = 5.64$) in this study serves as an important caveat to all reported findings.

3.2. Processing Voice Recordings

All crew members wore Philips Audio Recorders (DVT4010) throughout each mission day, providing a substantial bank of naturalistic audio data. These data serve as a valuable resource for characterizing scheduling-related collaboration as a crucial step in confirming Playbook’s usability and inferring crew behavior. To post-process the data, we first identified timeframes in which TPM and SS took place and extracted the relevant audio recordings. We then combined Whisper, a speech recognition system developed by OpenAI [23], with manual transcription. We used Whisper to automate the initial transcription and corrected errors through manual review. Transcripts were formatted to represent different speakers and verbalizations on each line.

While reviewing the transcripts, we synthesized our research question in a bottom-up fashion as recommended by Clinkenbeard [5]: what is the focus and frequency of different types of verbal collaboration during TPM and SS? After a comprehensive examination of the topics discussed by the crews, we identified four categories of interest (Table 1): collaboration regarding the timeline, collaboration regarding the task, collaboration regarding Playbook, or off-topic conversation. Each of these categories could provide insight into group behavior and attitudes; for example, collaboration regarding the nature of the self-scheduling task might reflect confusion about the instructions. Similarly, lots of off-topic chatter might represent low task burden or unengaging demands.

Table 1. The four observed conversation types and respective examples.

Label	Category	Description	Example
CRTimeline	Collaboration Regarding Timeline	Discussion on timeline content and preferences	“I personally like the questionnaires stacked together. . . knock ‘em all out.”
CRTask	Collaboration Regarding Task	Discussion on the nature of the assigned task at hand	“We can talk about what our preferences are.”
CRPlaybook	Collaboration Regarding Playbook	Discussion on how to use the tool or navigate the interface	“But where do you see the tasks?” “They’re right there on add-to-plan.”
OT	Off-Topic	Jokes, tangents, or unrelated topics	“Have you watched the latest season of that show?”

Next, we appointed two independent raters to categorize the crews’ conversations into one of the four categories. These raters were provided with the definitions of the collaboration categories from Table 1 and were instructed to tally the instances of each category for all TPM and SS. Given our primary interest in the topical focus of collaboration, raters were specifically instructed to closely monitor content shifts in blocks of verbalizations. To avoid potential confusion, we provided an example of a same-category content shift: if the discussion shifted from scheduling preferences for hygiene to preferences for when to do surveys, it would be counted as two instances of CRTimeline. To address the possibility of low interrater reliability, we predetermined that a third independent mediator would meet with the raters to resolve transcripts with significant count discrepancies.

We calculated an intraclass coefficient (ICC) to assess interrater reliability and found it to be excellent ($ICC = 0.91$, 95% $CI = [0.83, 0.95]$) according to guidelines set forth by Koo and Li [24], removing the need for any follow-up mediation. We then averaged the tallies from our independent raters to produce a single score for each category per transcript. Our primary variable of interest for analysis is the total counts for a discussion category during

a TPM or SS. We also considered the total duration of the TPM or SS as extracted from the length of our trimmed audio files.

4. Results

4.1. Team Preference Meetings (TPMs)

Figure 2 illustrates the overall time in minutes dedicated to team preference meetings by each crew and the breakdown of verbal collaboration counts to explore shifts over the course of the mission. Discussions consumed less time in later meetings ($M_{MD7} = 11.51$, $SD_{MD7} = 9.17$; $M_{MD8} = 12.07$, $SD_{MD8} = 4.70$; $M_{MD11} = 4.45$, $SD_{MD11} = 4.72$; $M_{MD12} = 3.73$, $SD_{MD12} = 2.57$), with counts of collaboration categories remaining relatively consistent over the MDs. Notably, two crews in C6 chose not to conduct TPMs at all (M2’s third meeting and M3’s fourth meeting), and these meetings are excluded from subsequent TPM plots. In Figure 3, the collaboration breakdown during TPM and SS sessions reveals that crews generally remained on task, primarily focusing on timeline preferences.

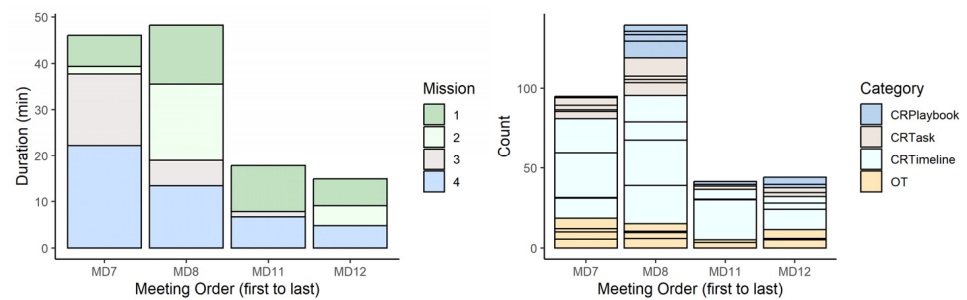


Figure 2. The duration (minutes) and count of category breakdown of TPMs. MD indicates the mission day the TPM took place.

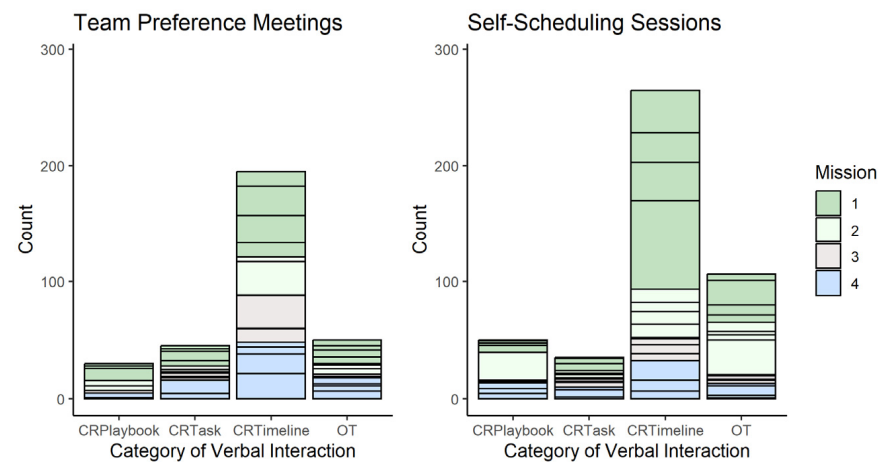


Figure 3. Counts during TPM and SS dedicated to different topics of discussion. Two TPM were skipped and are not included in this plot ($n = 14$). All SS are included ($n = 16$).

4.2. Self-Scheduling Sessions (SS)

During SS, crew planners were instructed to work independently to complete the self-scheduling activity. Figure 3 shows that they ignored this instruction, however, and collaborated often, predominantly discussing timeline content and preferences. In fact, 15 out of 16 planners engaged in collaboration during the SS process. The extent of this behavior greatly varied across individuals and crews, as illustrated by the variability in the stacked CRTimeline bar (see also Table 2).

Table 2. Descriptive statistics for counts of each category and overall durations across TPM and SS. The two skipped TPMs were not included in our calculations. Statistics are formatted as mean (standard deviation).

	Duration in Minutes	CRTimeline	CRTask	CRPlaybook	OT
Team Preference Meetings	9.08 (6.20)	13.93 (10.11)	3.21 (3.17)	2.14 (2.92)	3.57 (2.12)
Self-Scheduling Sessions	28.91 (19.25)	16.53 (19.11)	2.19 (1.98)	3.13 (5.80)	6.66 (7.86)

Exploratory correlation tests were conducted to investigate the potential relationship between SS interactions and workload, which was measured directly after SS using the NASA-TLX. Kendall's rank correlations were used due to the nonparametric form of the data. A negative correlation was found between counts of off-topic (OT) conversation and workload ($\tau_b = -0.42, p = 0.026$), while CRPlaybook, CRTimeline, and CRTask showed no significant correlation with workload ($\tau_b = -0.04, p = 0.85$; $\tau_b = -0.21, p = 0.26$; $\tau_b = -0.02, p = 0.93$).

4.3. Playbook Usability Scoring

Having established that self-scheduling in HERA is inherently collaborative, we aimed to quantify Playbook's usability in a spaceflight analog environment. Post-mission UEQ scores across crews and missions were amalgamated. These scores were then compared to a benchmark (v12) derived from thousands of UEQ ratings of various products, provided by Schrepp et al. [25]. Additionally, we compared these scores to past scores from a recent, larger in-lab sample ($n = 30$; [18]) which serve as a satisfaction baseline specifically for Playbook. This comparative analysis enables an assessment of whether Playbook performs in the upper percentiles of software UX, even when in a complex, collaborative environment.

Figure 4 shows that, in C6, Playbook scored highly in all aspects. It received excellent ratings (top 10% of products in the benchmark dataset) for attractiveness ($M = 1.87, SD = 0.67, 95\% CI = [1.54, 2.19]$), perspicuity ($M = 2.39, SD = 0.59, 95\% CI = [2.10, 2.68]$), efficiency ($M = 2.05, SD = 0.53, 95\% CI = [1.79, 2.31]$), and dependability ($M = 1.97, SD = 0.69, 95\% CI = [1.63, 2.31]$). Its stimulation ($M = 1.59, SD = 0.75, 95\% CI = [1.23, 1.92]$) and novelty ($M = 1.08, SD = 0.79, 95\% CI = [0.69, 1.47]$) were categorized as good (10% better, 75% worse) and above average (25% better, 50% worse), respectively.

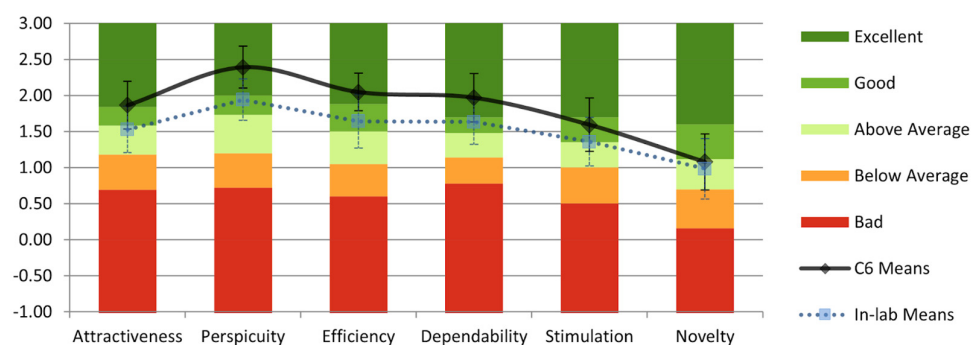


Figure 4. Average UEQ scores for Playbook from in-lab testing ($n = 30$) and HERA C6 ($n = 16$) mapped onto benchmark categories. Error bars represent the 95% confidence intervals.

5. Discussion

Our method, rooted in conversation analysis, aimed to comprehensively characterize collaboration among spaceflight analog crew members during timeline self-scheduling. Leveraging voice logs for unobtrusive monitoring, we remotely studied crew members'

verbal social interactions as they participated in team preference meetings (TPMs) and self-scheduling sessions (SS).

We observed a general downward trend in the duration of TPMs over time. Crews appeared to dedicate progressively less time to discussing preferences throughout their missions, with two crews forgoing some TPMs in the latter half of their mission. No crew used the full duration allocated for any TPMs, and all latter TPMs had durations of only a few minutes. After reviewing the transcripts, we were unable to find explicit content in discussions that explained this trend; however, we inferred that crews became more efficient at discussing their preferences or that these preferences remained generally static after the first couple of discussions. This insight will inform future spaceflight analog missions, and we plan to allocate less time for these meetings before the self-scheduling sessions. It may be adequate to allocate a single longer block for the first meeting and shorter meetings to update and re-discuss preferences every week afterward.

We also found that crew collaboration predominantly stayed on task during TPMs, focusing on time preferences rather than delving into Playbook features, the nature of the task at hand, or unrelated topics. This suggests that task instructions for these TPMs were clear, there were no critical Playbook usability issues, and assigned planners utilized the time to understand everyone's preferences before creating the team's schedule.

Surprisingly, our conversation analysis revealed an unexpected aspect of SS. The assigned crew scheduler engaged in impromptu conversations with the rest of the crew to integrate team preferences into the timeline. Without the use of conversational analysis, we never would have discovered that crews were collaborating during SS. This collaborative behavior was unexpected since the SS task was designed for a single scheduler to independently create a feasible timeline for the whole team. We designed the experiment expecting the scheduler to have discussed these preferences before (during TPM) and then self-schedule alone. As mentioned earlier, this tendency for sporadic collaboration varied greatly across planners and crews, but almost all SS contained some type of timeline-related collaboration. For the first time, we have quantitative data that show that self-scheduling is a collaborative task.

Similarly to the TPMs, the discussions during SS often stayed focused on timeline preferences. If much of the discussion revolved around maneuvering within Playbook's user interface, we would have interpreted that as a signal of some design flaw that required intra-crew collaboration to solve. If much of the discussion centered on the nature of the task itself or irrelevant topics, we would have interpreted that as confusing or unengaging task demands. Instead, it appears that timeline-related discussion between crew members was often a necessary ingredient for completing the self-scheduling task. These interactions consisted of crew members seeking and offering recommendations, which is a keystone behavior of team coordination. NASA has acknowledged positive team dynamics as vital for mission success [26].

We found a negative correlation between off-topic conversations and workload. There are three possible interpretations of this finding. First, crews may engage in more casual conversation during self-scheduling when the task incurs low workload. Second, the workload questionnaire may be measuring general workload within the analog rather than the specifics of the self-scheduling task, explaining the crew's chatting behavior. The third interpretation is that the NASA-TLX accurately reflected workload during the task, and that off-topic chatting did not increase workload. The first interpretation suggests that tracking conversations could serve as a viable unobtrusive proxy for workload in remote environments, which is particularly valuable in a spaceflight analog where participant survey compliance may be challenging. It aligns with literature that suggests that low workload indicates spare cognitive resources [20], which can then be directed to task-unrelated behaviors like mind-wandering (e.g., [27]) or, as presently observed, chatting. These preliminary results require exploration in subsequent analogs with larger astronaut-like samples.

After using the voice logs to establish that HERA is a collaborative, social environment during self-scheduling, we analyzed the descriptive statistics of post-mission usability survey responses to infer UX in a spaceflight analog. Our findings indicate that Playbook in C6 scored extremely well relative to a benchmark provided by the developers of the UEQ [25] and demonstrated comparable scores to those from prior testing in our controlled in-lab experiments [18]. While it remains uncertain to what extent general improvements and new features in Playbook [17] contributed to shifts in UX between lab and analog testing, our results suggest that Playbook is a usable system for actual autonomous crews in space missions.

Conversation analysis proved to be a useful technique for characterizing the nuances of self-scheduling with Playbook. We found that using this method to unobtrusively study group problem-solving was invaluable for exploring how our collaborative software was used in a realistic setting. The present work aligns with a recent trend in usability science that encourages scholars to address the broader context in which computer systems are used [5]. By using voice data to enrich our understanding of user interactions with spaceflight technology, we have paved the way for subsequent context-sensitive approaches in usability science as the technologies of the future are designed, developed, and deployed.

6. Conclusions and Future Work

Recorded conversations provided, for the first time, an unobtrusive glimpse into behaviors during self-scheduling in a spaceflight analog environment. The conversation analysis did not identify any usability issues but surprisingly showed that self-scheduling is more of a collaborative task than previously thought. Looking forward, voice recordings may serve as a proxy for cognitive usability metrics during collaborative work in remote settings like spaceflight analogs or even space. Furthermore, our findings suggest that the conversation analysis approach may be generalized to other settings that are isolated and confined, such as deep-sea exploration. Future research can track voice recordings in real-time, collect data from a variety of analog environments, and leverage greater representative samples to verify our proposed value of voice logs in computer systems development.

Author Contributions: Conceptualization, S.S. and J.J.M.; methodology, S.S.; formal analysis, S.S. and J.Z.; writing—original draft preparation, S.S., J.J.M. and J.A.K.; writing—review and editing, S.S., J.J.M., J.A.K. and J.Z.; visualization, S.S.; supervision, J.J.M.; funding acquisition, J.J.M., J.A.K. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was performed under a US Govt. Contract in the Human–Systems Integration Division at NASA. This research was funded in part by the NASA Human Research Program’s Human Factors and Behavior Performance Element (NASA Program Announcement number 80JSC017N0001-BPBA) Human Capabilities Assessment for Autonomous Missions (HCAAM) Virtual NASA Specialized Center of Research (VNSCOR) effort (NASA grant number 80NSSC19K0657).

Institutional Review Board Statement: The studies were conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of NASA.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the studies.

Data Availability Statement: The data and analysis scripts from this study may be made available on request to the corresponding author. The data are not publicly available due to privacy concerns.

Acknowledgments: The authors would like to thank Jessica Mar and Joshua Mosonyi for their valuable contributions to this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nielsen, J. *Usability Engineering*; AP Professional: Cambridge, MA, USA, 1993.
2. Nielsen, L.; Madsen, S. The Usability Expert’s Fear of Agility: An Empirical Study of Global Trends and Emerging Practices. In Proceedings of the 7th Nordic Conference on Human–Computer Interaction: Making Sense through Design, Copenhagen, Denmark, 14–17 October 2012; ACM: Copenhagen Denmark, 2012; pp. 261–264. [[CrossRef](#)]

3. Li, A.C.; Kannry, J.L.; Kushniruk, A.; Chrimes, D.; McGinn, T.G.; Edonyabo, D.; Mann, D.M. Integrating Usability Testing and Think-Aloud Protocol Analysis with “near-Live” Clinical Simulations in Evaluating Clinical Decision Support. *Int. J. Med. Inform.* **2012**, *81*, 761–772. [[CrossRef](#)] [[PubMed](#)]
4. Morgan, T.L.; Pletch, J.; Faught, E.; Fortier, M.S.; Gazendam, M.K.; Howse, K.; Jain, R.; Lane, K.N.; Maclaren, K.; McFadden, T.; et al. Developing and Testing the Usability, Acceptability, and Future Implementation of the Whole Day Matters Tool and User Guide for Primary Care Providers Using Think-Aloud, near-Live, and Interview Procedures. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 57. [[CrossRef](#)] [[PubMed](#)]
5. Clinkenbeard, M. Multimodal Conversation Analysis and Usability Studies: Exploring Human-Technology Interactions in Multiparty Contexts. *Commun. Des. Q. Rev* **2018**, *6*, 103–113. [[CrossRef](#)]
6. Nicolini, D. The Work to Make Telemedicine Work: A Social and Articulative View. *Soc. Sci. Med.* **2006**, *62*, 2754–2767. [[CrossRef](#)] [[PubMed](#)]
7. Heath, C.; Luff, P. Collaborative Activity and Technological Design: Task Coordination in London Underground Control Rooms. In Proceedings of the Second European Conference on Computer-Supported Cooperative Work ECSCW '91, Amsterdam, The Netherlands, 24–27 September 1991; Bannon, L., Robinson, M., Schmidt, K., Eds.; Springer: Dordrecht, The Netherlands, 1991; pp. 65–80. [[CrossRef](#)]
8. Pappas, Y.; Seale, C. The Physical Examination in Telecardiology and Televascular Consultations: A Study Using Conversation Analysis. *Patient Educ. Couns.* **2010**, *81*, 113–118. [[CrossRef](#)]
9. Fan, M.; Lin, J.; Chung, C.; Truong, K.N. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Trans. Comput.-Hum. Interact.* **2019**, *26*, 1–35. [[CrossRef](#)]
10. Carrón, J.; Campos-Roca, Y.; Madruga, M.; Pérez, C.J. A Mobile-Assisted Voice Condition Analysis System for Parkinson’s Disease: Assessment of Usability Conditions. *BioMedical Eng. OnLine* **2021**, *20*, 114. [[CrossRef](#)] [[PubMed](#)]
11. Paromita, P.; Khader, A.; Begerowski, S.; Bell, S.T.; Chaspari, T. Linguistic and Vocal Markers of Microbehaviors Between Team Members During Analog Space Exploration Missions. *IEEE Pervasive Comput.* **2023**, *22*, 7–18. [[CrossRef](#)]
12. Li, M.; Erickson, I.M.; Cross, E.V.; Lee, J.D. It’s Not Only What You Say, but Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation. *Hum. Factors* **2023**, *ahead of print*. [[CrossRef](#)]
13. Magnúsdóttir, E.H.; Johannsdóttir, K.R.; Majumdar, A.; Gudnason, J. Assessing Cognitive Workload Using Cardiovascular Measures and Voice. *Sensors* **2022**, *22*, 6894. [[CrossRef](#)] [[PubMed](#)]
14. Yin, B.; Chen, F.; Ruiz, N.; Ambikairajah, E. Speech-Based Cognitive Load Monitoring System. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008. [[CrossRef](#)]
15. Marquez, J.J.; Pyrzak, G.; Hashemi, S.; McMillin, K.; Medwid, J. Supporting Real-Time Operations and Execution through Timeline and Scheduling Aids. In *43rd International Conference on Environmental Systems*; American Institute of Aeronautics and Astronautics: Vail, CO, USA, 2013. [[CrossRef](#)]
16. Marquez, J.J.; Edwards, T.; Karasinski, J.A.; Lee, C.N.; Shyr, M.C.; Miller, C.L.; Brandt, S.L. Human Performance of Novice Schedulers for Complex Spaceflight Operations Timelines. *Hum. Factors* **2023**, *65*, 1183–1198. [[CrossRef](#)] [[PubMed](#)]
17. Zheng, J.; Shelat, S.M.; Marquez, J.J. Facilitating Crew-Computer Collaboration During Mixed-Initiative Space Mission Planning. In Proceedings of the SpaceCHI 3.0 Conference for Human-Computer Interaction for Space Exploration, Boston, MA, USA, 22 June 2023.
18. Shelat, S.; Karasinski, J.A.; Flynn-Evans, E.E.; Marquez, J.J. Evaluation of User Experience of Self-Scheduling Software for Astronauts: Defining a Satisfaction Baseline. In *Engineering Psychology and Cognitive Ergonomics*; Harris, D., Li, W.-C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; Volume 13307, pp. 433–445. [[CrossRef](#)]
19. Marquez, J.J.; Shelat, S.; Karasinski, J.A. *Promoting Crew Autonomy in a Human Spaceflight Earth Analog Mission through Self-Scheduling*; ASCEND: Las Vegas, NV, USA, 2022; p. 4263.
20. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Hancock, P.A., Meshkati, N., Eds.; Human Mental Workload; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183. [[CrossRef](#)]
21. Schrepp, M.; Hinderks, A.; Thomaschewski, J. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*; Marcus, A., Ed.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2014; pp. 383–392. [[CrossRef](#)]
22. Saint-Guillain, M.; Vanderdonckt, J.; Burny, N.; Pletser, V.; Vaquero, T.; Chien, S.; Karl, A.; Marquez, J.; Wain, C.; Comein, A.; et al. Enabling Astronaut Self-Scheduling Using a Robust Advanced Modelling and Scheduling System: An Assessment during a Mars Analogue Mission. *Adv. Space Res.* **2023**, *72*, 1378–1398. [[CrossRef](#)]
23. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518. [[CrossRef](#)]
24. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [[CrossRef](#)] [[PubMed](#)]
25. Schrepp, M.; Hinderks, A.; Thomaschewski, J. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *IJIMAI* **2017**, *4*, 40. [[CrossRef](#)]

26. Landon, L.B.; Slack, K.J.; Barrett, J.D. Teamwork and Collaboration in Long-Duration Space Missions: Going to Extremes. *Am. Psychol.* **2018**, *73*, 563–575. [[CrossRef](#)] [[PubMed](#)]
27. Casner, S.M.; Schooler, J.W. Thoughts in Flight: Automation Use and Pilots' Task-Related and Task-Unrelated Thought. *Hum. Factors* **2014**, *56*, 433–442. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.