Multidimensional usability assessment in spaceflight analog missions

Shivang Shelat
San Jose State University, Moffett Field, CA 94043, USA

Katherine E. Homer
NASA Ames Research Center, Moffett Field, CA 94043, USA

John A. Karasinski
NASA Ames Research Center, Moffett Field, CA 94043, USA

Jessica J. Marquez
NASA Ames Research Center, Moffett Field, CA 94043, USA

Abstract

2012 ACM Subject Classification Human-centered computing → HCI design and evaluation methods

Keywords and phrases space usability, crew autonomy, self-scheduling software

Funding This research was funded 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.

1 Introduction

The National Aeronautics and Space Administration (NASA) has envisioned that future astronauts will operate autonomously, without relying on ground support, due to communication latencies as exploration missions venture deeper into space [15]. Thus, new technologies and standards must become a core part of space operations to enable crew autonomy and keep deep space crews resilient to varying degrees of isolation. Given the large number of novel technologies that crews will encounter, it is imperative that these technologies be highly usable; they must be efficient, effective, easy to use and learn, and generally associated with a positive user experience [2, 18].

Prior work demonstrates the feasibility of quantitative approaches to assessing the usability of tools in the aerospace and spaceflight domain (e.g., [7, 17]). Many factors contribute to product usability, but not all factors apply to all products equally. While there are many usability questionnaires available, none incorporate every possible metric [23]. Therefore, the choice of which questionnaire to use can be product- or domain-specific. NASA's recommended metric to validate usability is the System Usability Scale (SUS; [5]), a questionnaire that captures overall user perceptions in a single score. In user experience (UX) today, SUS is the most widely used usability questionnaire and is considered simple, reliable, and valid [19]. A proposed threshold for acceptable usability is a SUS score of 85 out of 100 [1]. Most recently, efforts have led to the development of a NASA-specific modified SUS (NMSUS; [4]) that has curated questions for space-relevant technologies.

While the SUS quantifies usability on a single encompassing dimension, usability has traditionally been considered a multidimensional concept. Nielsen [18] described five key attributes of usability: learnability, efficiency, memorability, errors, and satisfaction. Paralleling this, much empirical work has aimed to develop a multidimensional usability questionnaire that is brief and sensitive to variations in design. One example is the User Experience Questionnaire (UEQ; [12]), a 3–5 minute survey that breaks down usability into six subscales:

attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty [20]. Shelat and colleagues [21] demonstrated the UEQ could capture fine-grained differences in the user experience of various planning interface designs. It remains unclear how different dimensions of the UEQ relate to the SUS in space environments, although it is reasonable to predict that they will be strongly positively correlated. This paper evaluates the relationship between SUS and UEQ multidimensions regarding assessing the usability of a spaceflight scheduling interface. Given the development of the NMSUS, verifying that the SUS and UEQ converge in spaceflight analog environments is a primary aim here.

Self-scheduling is one important capability required for space crews to become autonomous from Earth-based ground support. Research efforts at NASA have culminated in developing a self-scheduling software tool, Playbook, which allows astronauts to manage their own operational timelines [15]. Akin to a calendar, users can see their planned schedule as well as the schedule for the rest of the crew. Their schedules are laid out horizontally so that each crew member's timeline can be seen concurrently. Users can add new activities or add activities from a predetermined list. More frequently, assigned crew activities can be rescheduled or reassigned by simply dragging and dropping them along or between timelines. These activities have modeled constraints (e.g., the exercise activity uses the treadmill) and Playbook automatically checks if the rescheduled activity fulfills all its constraints (e.g., two people are not trying to use the same treadmill simultaneously).

As part of this ongoing research, Playbook is used to evaluate self-scheduling performance [14]. Controlled lab experiments were used to determine performance based on schedule characteristics (e.g., type of activity constraints). In an analog environment, we focused on assessing overall acceptability and usability, with and without visual aids to support self-scheduling performance. These visual aids, identified as countermeasures, were designed to facilitate scheduling and rescheduling, providing "No-Go" zones and suggested violation resolution [24, 10] (Figure 1). The primary countermeasure visualizes where an activity can be scheduled that meets the activity's constraints; any constraint violations are indicated with warning markers for the user to fix. If activity is scheduled on a "No-Go" zone, the user should expect the activity to create a violation.

The Human Exploration Research Analog (HERA) is an isolated spacecraft analog at the NASA Johnson Space Center used to simulate space exploration missions. This analog provides a unique opportunity to analyze spacecraft technologies in an isolated and confined environment over a relatively long mission (i.e., 45 days) with an "astronaut-like" crew. The usability data in this paper was collected over two campaigns, HERA Campaign 6 (C6) and Campaign 7 (C7). In the present work, we had two primary aims: 1) to demonstrate convergence between the SUS and the UEQ using data from astronaut-like participants in two HERA campaigns and 2) to test whether these conventional survey instruments can be leveraged to detect differences in usability based on software interface countermeasures in a spaceflight analog. With this, our broader goal was to show that conventional survey-based usability testing can be integrated into spaceflight analog environments to validate the deployment of tools designed for astronautic operations.

2 Methods

2.1 Materials

We administered two validated usability questionnaires post-mission: the System Usability Scale (SUS; [5]) and the User Experience Questionnaire (UEQ; [12]). The SUS is a unidimensional 10-item questionnaire that prompts participants to indicate how much they agree with

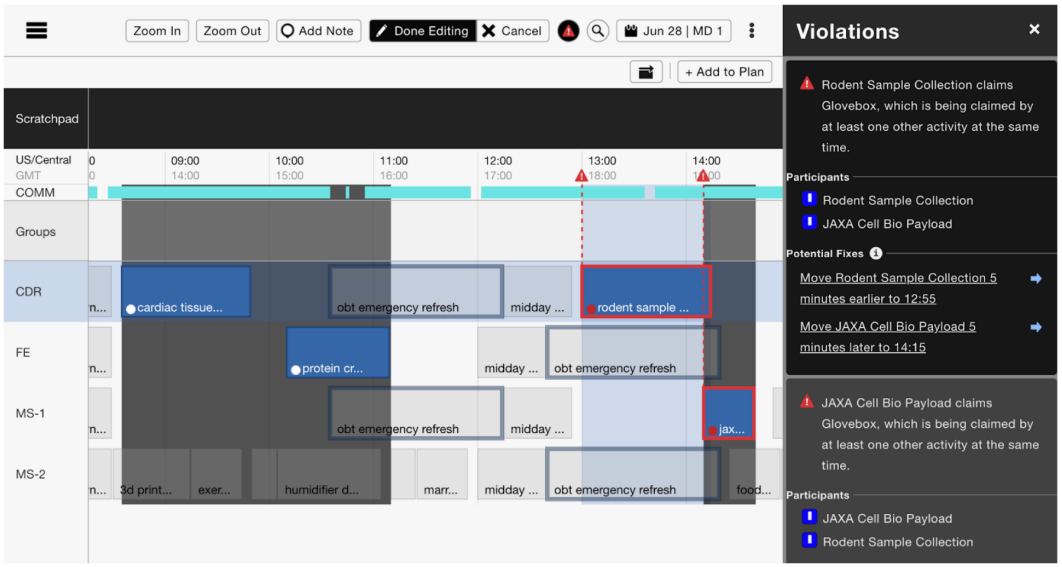

Figure 1 Visual countermeasure aids: "No-Go" zones (dark grey) and suggested violations solutions.

several statements regarding a tool or product on a 5-point Likert scale. One example item is "I found the various functions in this system were well integrated." A score is calculated by first coding responses to odd-numbered items from 0 to 4 and even-numbered items from 4 to 0. After, responses are summed and multiplied by 2.5 such that a minimum score is 0 and a maximum score is 100.

The UEQ is a 26-item survey that breaks usability into six discrete dimensions. Each item is an adjective that participants rate on a 7-point Likert scale (-3 to +3) as they pertain to the user experience of a tool or product. Here, we define the six UEQ subscales with an accompanying example item:

- Attractiveness: Do users like or dislike the tool? (enjoyable)
- Perspicuity: Do users find it easy to learn how to use the tool? (understandable)
- Efficiency: Can users solve their tasks without unnecessary effort? (practical)
- Dependability: Does the user feel in control of the interaction? (supportive)
- Stimulation: Do users find the tool exciting and motivating to use? (interesting)
- Novelty: Do users find the tool design creative and leading edge? (inventive)

We calculated scores on each of these dimensions using a data analysis tool (v12) provided by developers of the UEQ (`https://www.ueq-online.org/`). Attractiveness is generally considered an overall user experience metric, whereas perspicuity, efficiency, and dependability are more pragmatic qualities. Similarly, stimulation and novelty are considered hedonic qualities. The multidimensional structure of the UEQ enables researchers to isolate precise shifts in usability, which has proven useful for evaluating spaceflight tools (e.g., [21]).

2.2 Procedure

We analyzed data from 32 astronaut-like crew members who participated in an isolated 45-day simulated space mission to Phobos. This simulated spaceflight mission occurred over two HERA campaigns. Each campaign consisted of four missions, and each mission

was assigned a crew of four participants. During the mission, participants faced varying communication latencies and were instructed to follow operational timelines using Playbook; they could also schedule and reschedule part of their own operational timelines on Playbook. In Campaign 6, each participant was required to entirely self-schedule one operational day to be executed by the entire crew. Campaign 6's Missions 1 and 2 did not have Playbook countermeasures, whereas 3 and 4 did [16]. In Campaign 7, each participant completed this self-scheduling task twice, resulting in 8 self-scheduled mission days. We counterbalanced countermeasure exposure in Campaign 7, where Missions 1 and 2 did have countermeasures, whereas 3 and 4 did not. Participants could also use Playbook to spontaneously self-schedule flexible tasks on their timelines [3]. At the end of each mission, participants completed the SUS and UEQ surveys regarding their experience with Playbook. This study was approved by the NASA Institutional Review Board under expedited review (STUDY00000187) and determined to be no greater than minimal risk.

3 Results

First, we assessed the degree of coupling between SUS scores and different dimensions of the UEQ. We merged UEQ and SUS data ($n = 32$) from Campaigns 6 and 7 before performing Pearson correlation tests. The results are presented in Figure 2. All dimensions of the UEQ were significantly positively correlated with SUS, although it is worth noting that the dependability dimension was most tightly interlinked with SUS scores.

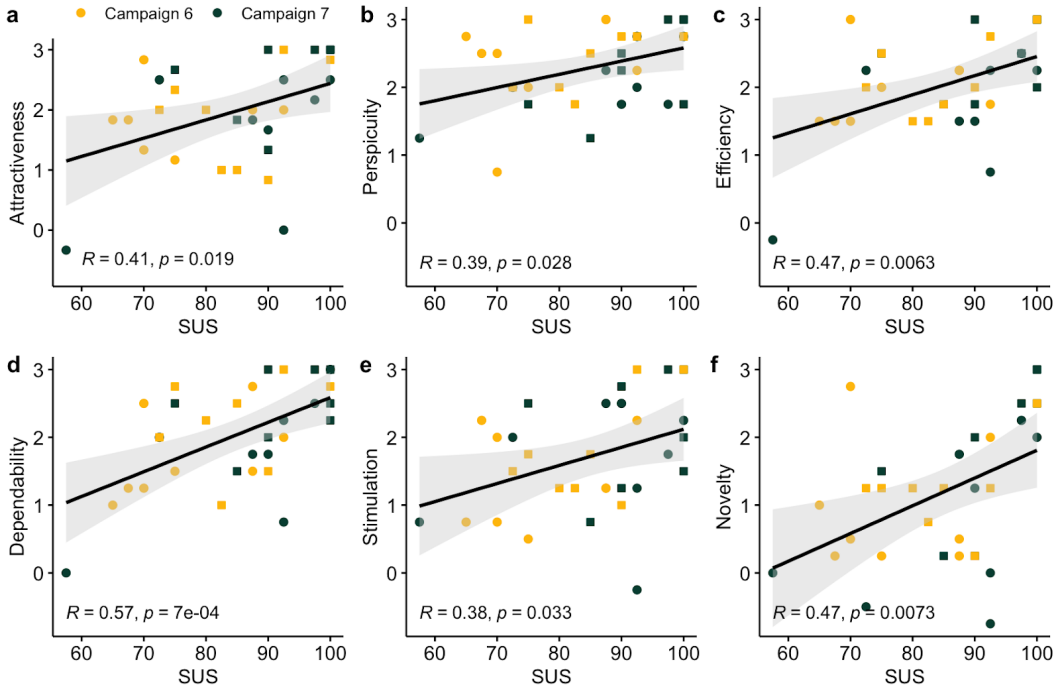

Figure 2 Pearson correlations between overall scores on the SUS and the a) Attractiveness, b) Perspicuity, c) Efficiency, d) Dependability, e) Stimulation, and f) Novelty dimensions of the UEQ. Gray shading represents the 95% confidence interval. Gold and dark green points represent HERA Campaigns 6 and 7 data, respectively. Squares and dots represent participants with and without countermeasures, respectively.

Next, we aimed to test whether either the SUS or the UEQ are sensitive enough to detect Playbook usability differences in spaceflight analogs based on the presence of scheduling interface countermeasures. We first split our sample into two groups: those with countermeasures ($n = 16$) and those without ($n = 16$). Next, we performed Welch's independent samples t-tests, assuming unequal variances in each UEQ dimension and overall SUS score. Due to the small sample size in each group and concerns about non-normality, we augmented this analysis with a permutation-based approach (e.g., [6]). We shuffled the "with" vs. "without" countermeasures group labels and recalculated the t-statistic 10,000 times for each test. This process creates a null distribution of t-statistics that does not rely on any parametric assumption. A p-value can then be extracted by taking the proportion of permuted t-statistics that exceed the original observed t-statistic. We present these permuted p-values along with other statistical information in Table 1. The group comparison is visualized in Figure 3.

Table 1 Welch's independent samples t-tests on usability dimensions between HERA crewmembers with vs. without Playbook interface countermeasures.

	Mean (standard deviation)					
Variable	**With**	**Without**	**t**	**df**	**p**	**d**
Attractiveness	2.24 (0.82)	1.73 (0.86)	1.71	29.94	0.092	0.61
Perspicuity	2.38 (0.56)	2.20 (0.63)	0.82	29.65	0.405	0.29
Efficiency	2.28 (0.56)	1.78 (0.76)	2.11	27.57	0.033*	0.75
Dependability	2.34 (0.61)	1.73 (0.79)	2.44	28.26	0.018*	0.86
Stimulation	1.95 (0.80)	1.48 (0.82)	1.63	29.98	0.106	0.58
Novelty	1.55 (0.94)	0.84 (1.04)	2.00	29.68	0.059	0.71
SUS	88.44 (9.61)	81.56 (13.10)	1.69	27.52	0.090	0.60

For research conducted in high-fidelity spaceflight simulations such as HERA, statistical power is constrained by low sample size due to selective participant criteria. Because of this, we focus on effect sizes (Cohen's d of 0.2 as small, 0.5 as medium, and 0.8 as large; [8]) in addition to significance thresholds for each t-test ($\alpha = 0.05$). The results showed that countermeasures conferred small to large benefits to usability on all dimensions. Countermeasures moderately enhanced attractiveness (p = 0.092, d = 0.61), efficiency (p = 0.033*, d = 0.75), stimulation (p = 0.106, d = 0.58), novelty (p = 0.059, d = 0.71), and SUS (p = 0.090, d = 0.60), although only the effect on efficiency was statistically significant. There was a small positive but nonsignificant effect on perspicuity (p = 0.405, d = 0.29) and a large, significant effect on dependability (p = 0.018*, d = 0.86). The results suggest that countermeasures improve Playbook's usability in spaceflight analogs. This effect is particularly pronounced for the usability dimensions of efficiency and dependability.

4 Discussion

Developing new, usable technologies for space crew autonomy is essential as long-duration exploration missions venture deeper into the unknown. Here, we analyzed crew perceptions of a self-scheduling software tool, Playbook, which is designed to enable crew autonomy. Over two 45-day campaigns in simulated spaceflight, astronaut-like crew members used Playbook to build operational timelines. Post-mission, they filled out a battery of questionnaires to characterize Playbook's usability on several dimensions. We found that scores on the SUS (a single-dimensional survey on overall usability) and on different facets of the UEQ (a six-subscale survey) were highly correlated. We also showed that countermeasures integ-

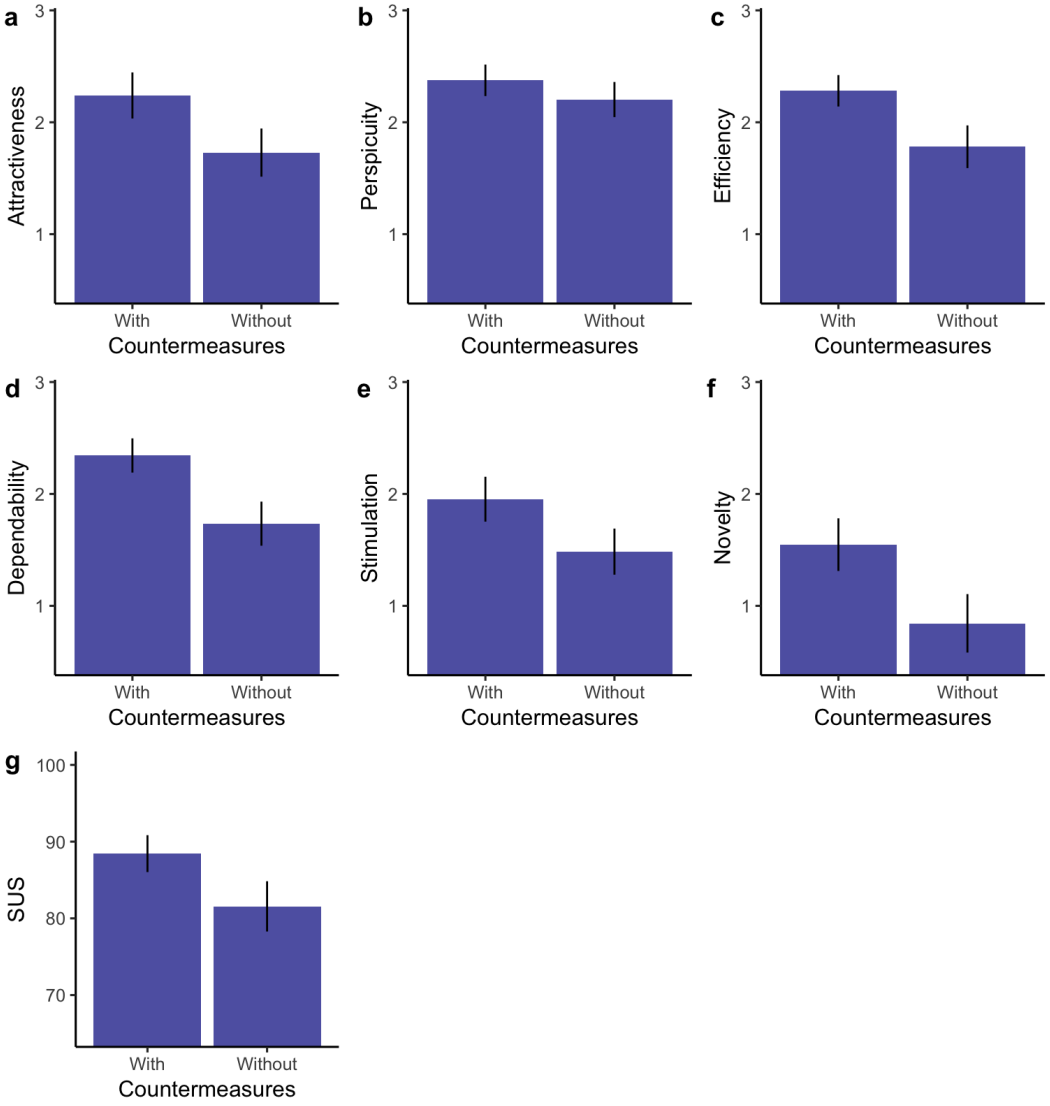

Figure 3 Group comparisons between HERA crew members with vs. without countermeasures on a) Attractiveness, b) Perspicuity, c) Efficiency, d) Dependability, e) Stimulation, f) Novelty, and g) SUS total. Error bars represent standard error.

rated into Playbook's user interface enhanced usability on multiple dimensions, with effect sizes ranging from small to large. These countermeasures elicited the strongest significant increase in perceived dependability and efficiency. Altogether, these results demonstrate convergence between a conventional usability survey recommended by NASA [1] and a newer, multidimensional survey only recently deployed in spaceflight contexts [21].

The findings speak to the utility of breaking down usability into discrete, quantifiable subdimensions to capture narrower shifts in usability due to interface changes. UEQ+, a modular tool that allows researchers to add more factors to the base UEQ, could further improve the flexibility of the questionnaire and its effectiveness in space environments. For example, UEQ+ has been validated to evaluate Trust and Voice User Interfaces, two examples likely relevant to future spaceflight UI [9, 11]. Having established that SUS is correlated with UEQ, future work with the new customizable UEQ+ has the potential to show even more precise conclusions.

The results affirm Playbook as a usable tool for spaceflight operations. Prior work has put forward score ranges to determine the rough acceptability of a tool based on overall SUS scores [13]: A+ (84.1 - 100), A (80.8 - 84.0), A- (78.9 - 80.7), B+ (77.2 - 78.8), B (74.1 - 77.1), B- (72.6 - 74.0), C+ (71.1 - 72.5), C (65.0 - 71.0), C- (62.7 - 64.9), D (51.7 - 62.6), and F (0 - 51.6). It is worth noting that without countermeasures (M = 81.56, SD = 13.10), Playbook receives an A; with countermeasures (M = 88.44, SD = 9.61), this is pushed to an A+. Similarly, NASA's *Human Integration Design Handbook* [1] states that usable systems usually get a SUS score of 85. Again, Playbook exceeds this threshold with countermeasures, but not without. A multipronged consideration of user attitudes (as we have done here), objective scheduling performance metrics [14, 10], and even conversational indices of collaborative scheduling [22] is necessary to paint the full picture of human-software dynamics for crew autonomy.

Furthermore, our objective results are consistent with our subjective user feedback. After each mission, crews are asked to name three things they liked about Playbook. Crews commented on the user interface's attractiveness and ease of use, and many appreciated the ability to conduct self-scheduling during the HERA missions. For example, one crew member mentioned how they liked "Freedom to make my schedule. One user who had the "No-Go" zone countermeasure mentioned "Ability to see restrictions/issues" among their top three Playbook qualities.

Our conclusions are not without limitations. First, without a direct comparison between Earth and real space conditions, it is possible that certain aspects of extreme environments that are not present in HERA (e.g., microgravity) shape usability. This remains an open research question in studying human spaceflight operations. Second, usability needs and priorities are likely different during space missions when compared to terrestrial operations. However, NASA currently requires terrestrial usability evaluations, which is why SUS and UEQ were selected as measures [1]. Since the SUS and UEQ were originally developed for Earth tools, this raises the possibility that some usability dimensions may be more important than others in space contexts. Mission context, complexity, and duration may have effects on usability. For the SUS, some have started evaluating these differences [4], but not yet for the UEQ. Space HCI would benefit from the consideration of these gaps moving forward.

We have shown here that such a multidimensional usability approach can be used to fully characterize usability in simulated spaceflight missions. If users are given time to fill out a 5-minute survey, the data enables researchers to identify where a tool may be lacking and then improve that area. This sets the stage for the deployment of tools such as Playbook to enable crew autonomy during actual operations. A valuable future step would be to

assess longitudinal usability assessments to see how these perceptions evolve; this approach is especially feasible for participants in controlled, long-duration analog environments.

References

1 *NASA Human Integration Design Handbook (HIDH)*. 2010.

2 ISO 9241-210:2019, 2019. URL: https://www.iso.org/standard/77520.html.

3 Renee Abbott, John Karasinski, and Jessica J. Marquez. Characterizing Spontaneous Self-Scheduling in NASA's Human Exploration Research Analog Campaign 6. In *IEEE Aerospace Conference*, 2025.

4 Kritina Anderson and Ian Robertson. Development and validation of the NASA modified System Usability Scale (NMSUS): a brief summary. November 2022. URL: https://ntrs.nasa.gov/citations/20220017395.

5 John Brooke. SUS — a quick and dirty usability scale. *Usability evaluation in industry*, 189:4–7, 1996.

6 Tom Bullock, James C. Elliott, John T. Serences, and Barry Giesbrecht. Acute exercise modulates feature-selective responses in human cortex. *Journal of Cognitive Neuroscience*, 29(4):605–618, April 2017. URL: https://direct.mit.edu/jocn/article/29/4/605/28609/Acute-Exercise-Modulates-Feature-selective, doi:10.1162/jocn_a_01082.

7 Kelly A. Burke, David J. Wing, and Mark Haynes. Flight test assessments of pilot workload, system usability, and situation awareness of tasar. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1):61–65, September 2016. URL: https://journals.sagepub.com/doi/10.1177/1541931213601014, doi:10.1177/1541931213601014.

8 Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Taylor and Francis, Hoboken, 2nd ed edition, 2013.

9 Andreas Hinderks, Martin Schrepp, Maria Rauschenberger, and Jörg Thomaschewski. Reconstruction and validation of the UX factor trust for the User Experience Questionnaire Plus (UEQ+):. In *Proceedings of the 19th International Conference on Web Information Systems and Technologies*, pages 319–329, Rome, Italy, 2023. SCITEPRESS - Science and Technology Publications. URL: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0012186700003584, doi:10.5220/0012186700003584.

10 John A. Karasinski, Shivang Shelat, and Jessica J. Marquez. Validation of self-scheduling countermeasures in NASA's HERA Campaign 6. In *AIAA SciTech Forum and Exposition*, January 2025. URL: https://ntrs.nasa.gov/citations/20240015124.

11 Andreas M. Klein, Jessica Kollmorgen, Andreas Hinderks, Martin Schrepp, Maria Rauschenberger, and Maria-Jose Escalona. Validation of the UEQ+ scales for voice quality. *Computer Standards & Interfaces*, 93:103971, April 2025. URL: https://linkinghub.elsevier.com/retrieve/pii/S0920548924001405, doi:10.1016/j.csi.2024.103971.

12 Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In Andreas Holzinger, editor, *HCI and Usability for Education and Work*, volume 5298, pages 63–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. URL: http://link.springer.com/10.1007/978-3-540-89350-9_6, doi:10.1007/978-3-540-89350-9_6.

13 James R. Lewis and Jeff Sauro. Item benchmarks for the system usability scale - jux, May 2018. URL: https://uxpajournal.org/item-benchmarks-system-usability-scale-sus/.

14 Jessica J. Marquez, Tamsyn Edwards, John A. Karasinski, Candice N. Lee, Megan C. Shyr, Casey L. Miller, and Summer L. Brandt. Human performance of novice schedulers for complex spaceflight operations timelines. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 65(6):1183–1198, September 2023. URL: https://journals.sagepub.com/doi/10.1177/00187208211058913, doi:10.1177/00187208211058913.

15 Jessica J. Marquez, Steven Hillenius, Bob Kanefsky, Jimin Zheng, Ivonne Deliz, and Marcum Reagan. Increasing crew autonomy for long duration exploration missions: Self-scheduling. In

2017 IEEE Aerospace Conference, pages 1–10, Big Sky, MT, USA, March 2017. IEEE. URL: http://ieeexplore.ieee.org/document/7943838/, doi:10.1109/AERO.2017.7943838.

16 Jessica J. Marquez, Shivang Shelat, and John A. Karasinski. Promoting crew autonomy in a human spaceflight earth analog mission through self-scheduling. In *Accelerating Space Commerce, Exploration, and New Discovery (ASCEND) 2022*, October 2022. URL: https://ntrs.nasa.gov/citations/20220013438.

17 David Meza and Sarah Berndt. Usability/sentiment for the enterprise and enterprise. NASA, 2014. URL: https://ntrs.nasa.gov/api/citations/20140007413/downloads/20140007413.pdf.

18 Jakob Nielsen. *Usability engineering*. AP Professional, Cambridge, Mass, 1993.

19 S. Camille Peres, Tri Pham, and Ronald Phillips. Validation of the System Usability Scale (SUS): SUS in the wild. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1):192–196, September 2013. URL: https://journals.sagepub.com/doi/10.1177/1541931213571043, doi:10.1177/1541931213571043.

20 Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In Aaron Marcus, editor, *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*, volume 8517, pages 383–392. Springer International Publishing, Cham, 2014. URL: http://link.springer.com/10.1007/978-3-319-07668-3_37, doi:10.1007/978-3-319-07668-3_37.

21 Shivang Shelat, John A. Karasinski, Erin E. Flynn-Evans, and Jessica J. Marquez. Evaluation of user experience of self-scheduling software for astronauts: defining a satisfaction baseline. In Don Harris and Wen-Chin Li, editors, *Engineering Psychology and Cognitive Ergonomics*, volume 13307, pages 433–445. Springer International Publishing, Cham, 2022. URL: https://link.springer.com/10.1007/978-3-031-06086-1_34, doi:10.1007/978-3-031-06086-1_34.

22 Shivang Shelat, Jessica J. Marquez, Jimin Zheng, and John A. Karasinski. Collaborative system usability in spaceflight analog environments through remote observations. *Applied Sciences*, 14(5):2005, February 2024. URL: https://www.mdpi.com/2076-3417/14/5/2005, doi:10.3390/app14052005.

23 Dominique Winter, Andreas Hinderks, and Jörg Thomaschewski. Applicability of user experience and usability questionnaires. *Journal of Universal Computer Science*, January 2020.

24 Jimin Zheng, Shivang M. Shelat, and Jessica J. Marquez. Facilitating crew-computer collaboration during mixed-initiative space mission planning. In *SpaceCHI 3.0 A Conference for Human-Computer Interaction for Space Exploration*, June 2023. URL: https://ntrs.nasa.gov/citations/20230008619.