Evaluation of User Experience of Self-scheduling Software for Astronauts: Defining a Satisfaction Baseline

Shivang Shelat[1(✉)], John A. Karasinski[2], Erin E. Flynn-Evans[2], and Jessica J. Marquez[2]

[1] San Jose State University Research Foundation, San Jose, CA 95112, USA
sshelat@ucsb.edu
[2] NASA Ames Research Center, Mountain View, CA 94043, USA

Abstract. As NASA turns its sights to deep-space exploration, a greater focus on supporting crew autonomy has led to the development of Playbook, a self-scheduling software tool. Evaluating the user satisfaction of Playbook is essential in ensuring its usability for critical spaceflight operations. Satisfaction of an interface is often quantified with attitude surveys, such as the User Experience Questionnaire (UEQ). This paper demonstrates an application of the UEQ in comparing the user experience of Playbook interface designs for displaying graphical data. We lay the foundation for future user experience comparisons by defining a satisfaction baseline, which is crucial as more features are integrated into Playbook's interface. This work extends a validated user experience framework into a spaceflight domain, allowing optimization of human-computer interaction as future operational tools are developed.

Keywords: Spaceflight software · Self-scheduling · User experience · Usability

1 Introduction

Future NASA deep-space human exploration missions will require astronauts to behave and perform more autonomously from ground flight controllers. One manner to provide additional crew autonomy is to enable astronauts to manage their own sets of activities, or schedules, during spaceflight operations. Astronauts aboard the International Space Station (ISS) follow detailed schedules designed and supported by ground-station planners, but do not currently participate in the development of these schedules. Crew self-scheduling would allow planners and astronauts to collaborate, providing astronauts the ability to reschedule or manage their own timeline [17, 18]. Playbook, a mission planning and scheduling software tool, allows users to self-schedule by organizing mission activities on a user-friendly interface. The interface has a timeline that displays all of the scheduled activities for the crewmembers. If activities do not meet their constraints, the software highlights the scheduling violations in the user interface. This prompts crewmembers to reorganize their activities until they have a plan that is feasible and

D. Harris and W.-C. Li (Eds.): HCII 2022, LNAI 13307, pp. 433–445, 2022.
https://doi.org/10.1007/978-3-031-06086-1_34

violation-free. The design of the timeline allows astronauts, who are inexperienced planners, to build comprehensive schedules without any intervention from ground personnel [18].

A valuable practice in software development is to maximize the usability of system interfaces through user-focused research methods. Unusable systems pose risks to a variety of fields, and are often associated with high error rates, high workload, user discomfort, and poor productivity [1, 7, 9, 23, 27]. A human-centered design approach that evaluates the specifics of any human-computer interaction supports the creation of a usable system, which is vital to performing NASA operations [20]. The *Spaceflight Human-System Standard* specifies the need for a defined usability acceptance criteria for each NASA program [21], and asserts that a usable crew interface must allow users to achieve their tasks efficiently, effectively, and with satisfaction. Efficiency refers to the use of resources and time involved with completing a task, effectiveness refers to accuracy and the ability to complete the task, and satisfaction refers to the comfort and attitude carried towards the interface.

Maximizing satisfaction ensures that the user has a positive experience with the interface and is comfortable using it to perform operations. The International Organization for Standardization states that for the satisfaction of a system to be acceptable, the user experience from interacting with the system must meet the user's needs and expectations [14]. To verify that the interface of operational software meets user expectations, we must evaluate user experience when any changes are made to the software design. NASA's *Human Integration Design Handbook* [20] suggests that this evaluation procedure can be done by using quantitative measures of user experience, such as standardized questionnaires. If a researcher wanted to investigate whether a particular new design of a system met an acceptable level of satisfaction, they could compare user experience scores to a defined quantitative baseline derived from that same system [13, 15]. A baseline provides a simple approach to evaluating the user experience of new interfaces and enables frequent testing to confirm that users are satisfied with the system.

In this paper we demonstrate the use of the User Experience Questionnaire (UEQ) in comparing the user experience of different iterations of the Playbook interface. Additionally, we define a baseline satisfaction standard to assist evaluations of Playbook's user experience moving forward. The goal of the paper is to fulfill the need for a NASA program usability standard and to support a research approach that allows for the optimization of Playbook's interface as future designs are developed for operational use. By extending the UEQ's ability to evaluate user experience into the realm of operational spaceflight tools, we emphasize that a usability-oriented research approach in systems development is key to optimizing human-computer interaction.

2 The User Experience Questionnaire

The User Experience Questionnaire (UEQ) is a short, 3–5-min scale that was developed to quantify the user experience of interactive products such as business applications, web shops, and development tools [26]. Subjects respond to 26 items after interacting with a product, and computed scores can be used to evaluate how they felt and identify what facets of the user experience were positive or negative.

Each of the items provides a 7-point Likert scale between two adjectives with opposite meanings. Each item pertains to one of 6 subscales of the UEQ, and scores for each subscale can be computed on a calculator provided by the UEQ developers [24]. Scores can range from −3 to 3, with values between −0.8 and 0.8 representing a neutral evaluation, values greater than 0.8 representing a positive evaluation, and values less than −0.8 representing a negative evaluation.

Each of the six subscales of the UEQ reflect an aspect of user experience [25]. The aspects are:

- *Attractiveness*: Do users like or dislike the product?
- *Perspicuity*: Is it easy to get familiar with the product?
- *Efficiency*: Can users solve their tasks with the product without unnecessary effort?
- *Dependability*: Does the user feel in control of the interaction?
- *Stimulation*: Is it exciting and motivating to use the product?
- *Novelty*: Is the product innovative and creative?

Attractiveness is considered to be an overall positive or negative impression of the product. Perspicuity, Efficiency, and Dependability are considered to be aspects of "hard user experience" and represent the pragmatic quality of the product. Users typically perceive products with greater pragmatic quality as easy-to-learn, efficient, and secure. Stimulation and Novelty are considered to be aspects of "soft user experience" and represent the hedonic quality of the product. Users typically perceive products with greater hedonic quality as interesting and leading-edge [25].

UEQ developers have proposed specific approaches that can be applied to evaluate a product's user experience. The first is to compare a product's scores to a provided benchmark that consists of a large number of responses from a variety of entities [26]. This comparison allows researchers to evaluate if the user experience of their product meets the expectations of the general user population. The calculator compares the average UEQ scores of entered data to this benchmark. It is important to note that the dataset of the benchmark does not distinguish between different product categories, and it includes data from products that are drastically different from Playbook, such as social networks and household appliances. The second proposed approach is to quantitatively determine whether a new version has an improved user experience by comparing average scores among each subscale to an older version. As developers redesign software to include new capabilities and features, we may expect to see a change in the user experience. Administering the UEQ to a sample of users allows for a quick and easy comparison between an old and new version of that software.

Researchers can also use the UEQ to determine how design changes affect the scores of different versions of their tools. While not as comprehensive as qualitative user feedback from usability testing, UEQ scores can inform educated guesses as to what design element affects which aspect when comparing multiple interfaces [25]. The delineation of the six dimensions by the UEQ allows for independent subscale comparisons that provide more insight into the strengths and flaws of a product than a single score alone. Additionally, we are able to prioritize different aspects of the UEQ based on the purpose of the entity we are evaluating. For Playbook, we care more about the goal-oriented aspects such as Perspicuity to maximize operational usability, and we care less about

competing with other products in the market. We can apply comparative methods to reach conclusions about Playbook's user experience in order to guide design decisions and software development in the future.

3 User Interface Evaluations in HERA Campaign 4

3.1 Study Overview

In 2018, NASA conducted the fourth campaign in the Human Exploration Research Analog (HERA) to simulate deep-space missions. There were a multitude of research objectives during the campaign, including evaluating the effectiveness of different biomathematical sleep models in predicting crewmember fatigue [10]. In addition to having numerous physiological measures evaluated during the 45-day missions, a subset of HERA crewmembers was administered four Fatigue Interface Testing (FIT) sessions in which they used different designs of Playbook to solve scheduling problems. Each design involved using predictions from fatigue models to inform the user of how their performance would be expected to change based on their prior sleep and circadian rhythm phase. Model predictions were generated offline and integrated into Playbook's interface [11]. This integration is a prototyping technique called "Wizard of Oz" (WOZ), in which the user thinks that the software has certain capabilities, but a person has actually simulated these capabilities behind the scenes [3].

The four interfaces each displayed model predictions, but the presentation of information had different designs. Specifically, the design varied among the types of predictions of performance on a standard five-minute reaction time test (lapses [count of reaction times $>$ 500 ms], mean response time in milliseconds, and mean speed [1000/reaction times] vs. only lapses), the number of models used (one model output vs. three model outputs), and the visual representations of the predictions. Crewmembers filled out the UEQ after each FIT session to give us a comprehensive impression of their user experience.

Participants were recruited through a variety of methods, from advertisements at NASA to appeals to the general public. Those who were selected were "astronaut-like." Individuals were divided to populate five separate missions of four members each [11]. In this investigation, we consider the eight HERA crewmembers of Missions 3 and 4, as they were the only crewmembers that completed the task with all four of the designs.

- *FIT Session 1*: Sleep model outputs (lapses, mean response time, and mean speed) for each scheduling scenario were presented in graph format in the context of their day's schedule (see Fig. 1).
- *FIT Session 2*: Sleep model output (lapses) for each scheduling scenario were presented in graph format alongside a red-green-yellow "heatmap", in the context of their day's schedule (see Fig. 2).
- *FIT Session 3*: Sleep model output (lapses) for each scheduling scenario were presented in graph format alongside a red-green-yellow "heatmap", in the context of their day's schedule. It also contained a legend for the graphs and data (see Fig. 3).

- *FIT Session 4*: Sleep model output (lapses) for three different models for each scheduling scenario were presented in graph format in the context of their day's schedule. Each graph was colored red-green-yellow accordingly (see Fig. 4).

Fig. 1. FIT session 1 interface.

Fig. 2. FIT session 2 interface.

Fig. 3. FIT session 3 interface.

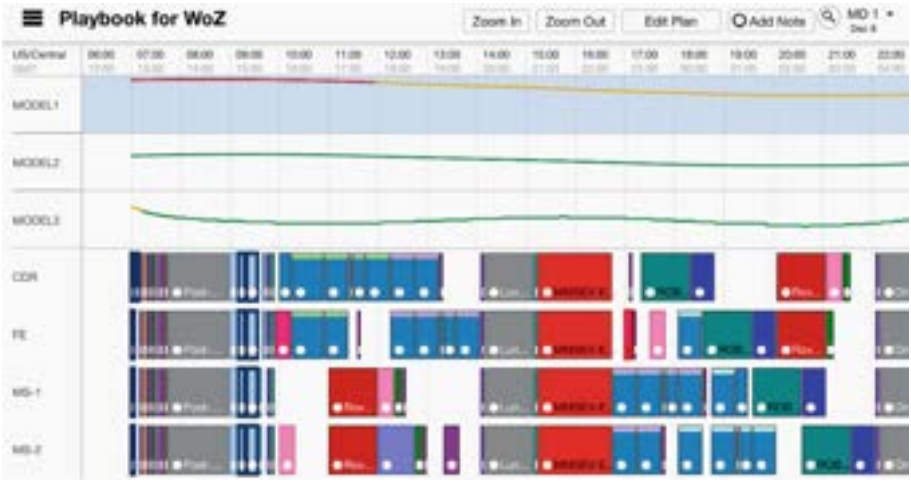

Fig. 4. FIT session 4 interface.

3.2 Results

HERA Campaign 4's exploration of user experience (see Fig. 5) showed steep differences in the Perspicuity aspect, particularly between FIT Sessions 1 and 3 and FIT Sessions 3 and 4. Because of these differences, we wanted to test an approach in comparing the user experience of different interfaces with the UEQ.

By attributing differences in the designs to the observed variation in the Perspicuity scores, we hypothesized that the crew preferred FIT 3 over FIT 1 because of the red-green-yellow lines, corresponding heatmap, and the legend for graphs/data. Additionally, we hypothesized that they did not prefer FIT 4 over FIT 3 because of the confusing (large) amount of model suggestions.

We chose to use a linear mixed-effect model to test the hypotheses because of the repeated measures experimental design. Each participant underwent each FIT Session,

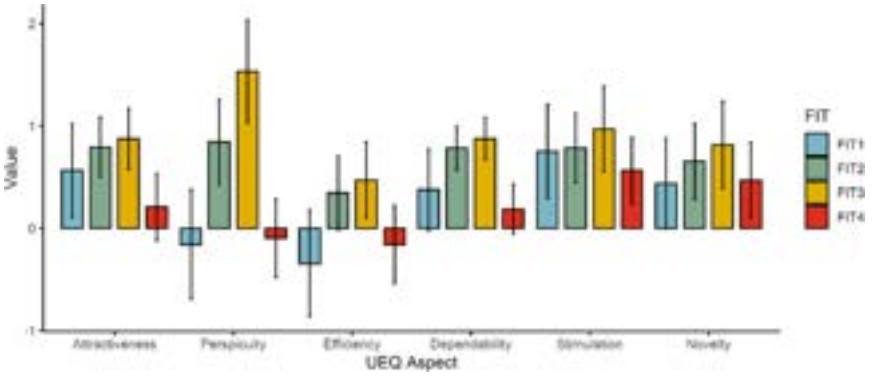

Fig. 5. A comparison of different scores on each UEQ subscale among the four FIT Session designs. Error bars represent the standard error.

resulting in dependency among the data. A linear mixed-effect model allows us to control for dependency by treating the repeated measurements of each subject as a random effect.

We ran the linear mixed-effect model analysis in R (version 4.0) [22] using the GAMLj package [12]. We included UEQ Value as the dependent variable and added fixed effects of UEQ Aspect and FIT Design, as well as the interaction between UEQ Aspect and FIT Design. We included Participant as a random effect. The model specification was as follows:

$$\text{Value} \sim \text{FIT\_Design} + \text{UEQ\_Aspect} + \text{FIT\_Design} : \text{UEQ\_Aspect} + (1 \mid \text{Participant})$$

The linear mixed-effect model detected statistically significant differences among UEQ aspects between different FIT designs, $F(3, 161) = 8.21$, $p < .001$. Post-hoc analysis with a Bonferroni adjustment showed that the Perspicuity of FIT 3 ($M = 1.53$, $SD = 1.44$) was significantly greater than the Perspicuity of FIT 1 ($M = –0.16$, $SD = 1.51$), $t(161) = –4.03$, $SE = 0.42$, $p < .05$. Additionally, the Perspicuity of FIT 4 ($M = –0.09$, $SD = 1.09$) was significantly lesser than the Perspicuity of FIT 3 ($M = 1.53$, $SD = 1.44$), $t(161) = 3.88$, $SE = 0.42$, $p < .05$. There were no other significant differences among matched UEQ aspects between the different conditions.

The HERA Campaign 4 analysis showed that the UEQ can be used to evaluate the user experience of different design iterations within Playbook. Similar research approaches like that of Campaign 4 can be used to detect significant differences in the user experience of an interface. To assist future user experience comparisons, we define a satisfaction baseline that consists of a more recent set of UEQ scores.

4 Developing a Playbook Satisfaction Standard

Our research study in HERA Campaign 4 only focused on the user experience of a specific Playbook feature for analog crewmembers and not about the Playbook tool in general. In order to consider future features that support and enable self-scheduling, we collected user experience data on Playbook as a whole. This allows us to create a standard that

future design iterations of Playbook can be compared to so we can adequately compare integrated features and evaluate their effects on user experience.

4.1 Study Overview

In a study conducted remotely due to the COVID-19 pandemic, participants solved a number of scheduling and rescheduling problems using Playbook. Participants were split into two groups of subjects. 15 participants were instructed to schedule their task timeline and 16 participants were instructed to reschedule a predetermined task timeline. This experiment was designed to evaluate the differences in performance between the scheduling conditions (scheduling vs. rescheduling), type of task constraint (time range vs. requires vs. claim vs. ordering), and number of task constraints (33% vs. 66%) [8]. Each participant solved one baseline condition in which there were no scheduling constraints and 8 other conditions in which the type of constraint and the number of constraints varied. At the end of all the trials, they filled out the UEQ to evaluate their overall experience with Playbook. In this investigation, we only consider 30 participants because 1 subject in the rescheduling condition experienced a technology issue that interfered with their ability to respond to the UEQ.

4.2 Results

The results showed only minor differences between the conditions among each UEQ subscale (see Fig. 6), suggesting that there were no significant differences in user experience between using Playbook for scheduling and rescheduling.

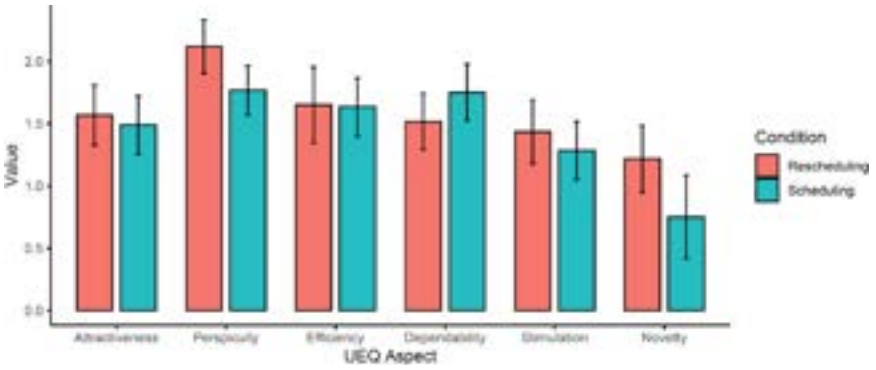

Fig. 6. Average UEQ scores for the scheduling and rescheduling conditions. Error bars represent the standard error.

To test for statistically significant differences between the scheduling and rescheduling conditions, we used a two-sample t-test assuming unequal variances. The results are shown in Table 1 and indicate that there were no significant differences between the two conditions. To create a larger, singular dataset that can serve as a satisfaction baseline for

Table 1. Results of a 2-sample t-test assuming unequal variances. There are no significant differences between the scheduling and rescheduling conditions. SD = standard deviation.

Aspect	Scheduling		Rescheduling		*t*	*df*	*p*
	Mean	SD	Mean	SD			
Attractiveness	1.49	0.91	1.57	0.93	0.23	27.99	0.82
Perspicuity	1.77	0.76	2.12	0.83	1.20	27.82	0.24
Efficiency	1.63	0.91	1.65	1.19	0.04	26.23	0.97
Dependability	1.75	0.88	1.52	0.87	-0.73	28.00	0.47
Stimulation	1.28	0.91	1.43	1.00	0.43	27.74	0.67
Novelty	0.75	1.29	1.22	1.04	1.09	26.79	0.28

Table 2. The descriptive statistics and UEQ benchmark comparison of the Playbook satisfaction baseline. An interpretation of the comparison is provided by developers of the UEQ [26]. For each aspect, N = 30. SD = standard deviation; CI = confidence interval; LL = lower limit; UL = upper limit.

Aspect	Mean	SD	95% CI		Benchmark Comparison
			LL	UL	
Attractiveness	1.53	0.90	1.20	1.85	Above average
Perspicuity	1.94	0.80	1.65	2.23	Good
Efficiency	1.64	1.04	1.27	2.01	Good
Dependability	1.63	0.87	1.32	1.94	Good
Stimulation	1.36	0.94	1.02	1.69	Good
Novelty	0.98	1.17	0.56	1.40	Above average

future design iterations of Playbook, we merged the UEQ responses from the scheduling and rescheduling participants.

Treating the UEQ data from this experiment as a singular dataset, we evaluated the user experience of this version of Playbook using methods outlined by the developers of the UEQ. Figure 7 is a visualization of each subscale and its performance relative to a provided benchmark which consists of a large bank of data from many entities. The results indicate that Playbook has been evaluated positively (>0.8) on each aspect of user experience. The calculator provided by the UEQ creators provides a benchmark comparison, which can be found in Table 2. "Above average" means that 25% of the benchmark products score higher and 50% score lower. "Good" means that 10% of the benchmark products score higher and 75% score lower [26]. The comparison shows that Playbook meets the user experience expectations of the general user population and scores highly among the products that make up the benchmark.

All aspects of the user experience of Playbook have been evaluated positively, with the highest score being Perspicuity and the lowest being Novelty. Relative to the provided UEQ benchmark, Playbook scores either above average or good in every aspect. As Playbook's goal is to maximize the user's operational ability, we are more focused on

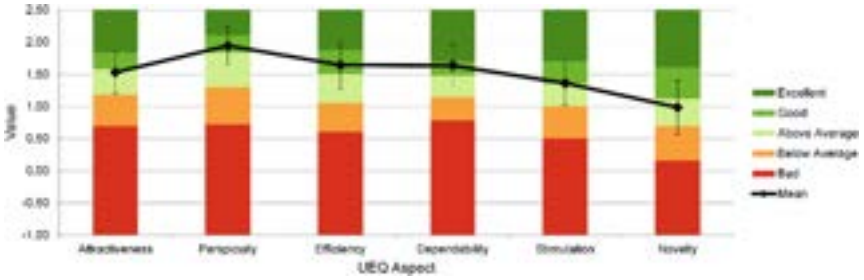

Fig. 7. Average scores for Playbook vs. the UEQ Benchmark. > 0.8 is a positive evaluation, between -0.8 and 0.8 is neutral, and < -0.8 is negative. Error bars represent a 95% confidence interval.

the pragmatic subscales of Perspicuity, Efficiency, and Dependability, which all indicate a positive user experience. After applying the methods outlined by the developers of the UEQ, we can predict that users will have a positive attitude towards Playbook's interface and be comfortable using it to complete tasks. The dataset from this experiment can be used as a satisfaction baseline that future Playbook designs can be compared to in order to preserve the positive user experience.

5 Discussion

User satisfaction is an important aspect of the overall usability of a system. When completing tasks with an unsatisfying system, user needs and expectations are not met, causing negative attitudes towards the interface [14, 20]. To ensure an acceptable level of satisfaction, NASA's *Spaceflight Human-System Standard: Volume 2* states the need for each program to define a usability standard that includes a metric of satisfaction [21]. A satisfaction baseline allows for frequent testing to maintain a system's positive user experience as developers create new versions. It is therefore critical to develop and define a user experience standard that future Playbook versions and self-scheduling features can be compared to. This paper demonstrates the use of the User Experience Questionnaire (UEQ) in comparing the user experience of different Playbook interfaces and lays the foundation in evaluating Playbook satisfaction by defining a baseline. We have demonstrated the ability to measure and compare user satisfaction with the UEQ and detect differences in Playbook feature designs.

When comparing different biomathematical model design integrations of Playbook, we found that certain features significantly improved the Perspicuity of the interface. For example, the design iteration that only presented one modeled output (lapses) instead of three (lapses, mean response times, and mean speed) scored significantly higher on the Perspicuity aspect. This indicates that crewmembers preferred a single suggestion instead of multiple, which made interpreting the user interface confusing. Our investigation of the HERA Campaign 4 UEQ data showed that Playbook interfaces can be effectively evaluated by the UEQ and compared to one another.

The scheduling and rescheduling experiment gave us a novel UEQ dataset that measures the user experience of a general version of Playbook. After establishing that there

were no significant differences in UEQ scores between the two experimental conditions, we merged the scores from all subjects into a baseline dataset that other Playbook versions can be compared to. Using methods outlined by UEQ developers, we established that Playbook has a positive user experience among each subscale. We now have the precedent and the means to make future user experience comparisons and ensure that Playbook provides a satisfying interface to its users.

Administering the UEQ to a pool of subjects after they interact with a new version of Playbook provides data that is easy to statistically compare against the satisfaction standard. We can then evaluate if an integrated feature significantly improves or worsens the user experience. If a feature increases human performance without significantly harming the user experience, we may be able to treat it as operationally usable.

5.1 Future Research

Future research should seek to further validate the ability of the UEQ to conclude that a system is sufficiently usable. To do this, UEQ scores can be compared to scores from a different questionnaire also designed to measure user experience, such as the System Usability Scale (SUS) [16]. The SUS is a brief, 10-item attitude questionnaire that evaluates the perceived usability of a system. Each item is a statement that describes how a user may have felt about the system that they interacted with, followed by a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The final score can range from 1 to 100. NASA's *Human Integration Design Handbook* [20] explicitly states that usable systems receive a score of 85. Bangor, Kortum, & Miller have defined specific SUS score ranges that relate to acceptability (< 50 = unacceptable, 50–70 = marginal, > 70 = acceptable) [2, 4].

While the SUS has been used to evaluate the user experience of other NASA programs [5, 19] and is explicitly recommended by the *Human Integration Design Handbook* [20], the UEQ offers more insight into what distinct aspects of user experience have been affected by changes to an interface. While both measures have been used together in prior studies [6, 28], they have not been conjunctively applied in a spaceflight context. A comparison between Playbook's SUS and UEQ scores would shine light on the UEQ's ability to conclude that Playbook does indeed have an acceptable interface and positive user experience. If Playbook is evaluated positively by the SUS, it supports the conclusion that our UEQ analysis has yielded. It is also of interest to note whether the positive, negative, and neutral UEQ scores correlate with acceptability ranges of SUS scores as defined by Bangor, Kortum, & Miller [2].

To achieve this goal, future experiments should consider collecting both UEQ and SUS responses after administering experimental trials in Playbook. This would provide data to draw comparisons between the methods of evaluating Playbook's user experience and deriving meaningful conclusions in the future. By integrating two user experience frameworks in evaluating satisfaction, we contribute to research approaches in human-computer interaction supporting usable system development.

References

1. Ahlstrom, U., Arend, L.: Color usability on air traffic control displays. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **49**, 93–97 (2005). https://doi.org/10.1177/154193120504900121
2. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. **4**, 114–123 (2009)
3. Bernsen, N.O., Dybkjær, H., Dybkjær, L.: Wizard of oz prototyping: how and when. In: CCI Working Papers in Cognitive Science and HCI., Roskilde, Denmark (1994)
4. Brooke, J.: SUS: a retrospective. J. Usability Stud. **8**, 29–40 (2013)
5. Burke, K.A., Wing, D.J., Haynes, M.: Flight test assessments of pilot workload, system usability, and situation awareness of tasar. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **60**, 61–65 (2016). https://doi.org/10.1177/1541931213601014
6. Devy, N.P.I.R., Wibirama, S., Santosa, P.I.: Evaluating user experience of english learning interface using user experience questionnaire and system usability scale. In: 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), pp. 101–106. IEEE, Semarang (2017)
7. Donahue, G.M.: Usability and the bottom line. IEEE Softw. **18**, 31–37 (2001). https://doi.org/10.1109/52.903161
8. Edwards, T., Brandt, S.L., Marquez, J.J.: Towards a measure of situation awareness for space mission schedulers. In: Ayaz, H., Asgher, U., Paletta, L. (eds.) Advances in Neuroergonomics and Cognitive Engineering, pp. 39–45. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-80285-1_5
9. Fairbanks, R.J., Caplan, S.: Poor interface design and lack of usability testing facilitate medical error. Joint Comm. J. Qual. Saf. **30**, 579–584 (2004). https://doi.org/10.1016/S1549-3741(04)30068-7
10. Flynn-Evans, E.E., et al.: Evaluation of the validity, acceptability and usability of biomathematical models to predict fatigue in an operational environment (2018)
11. Flynn-Evans, E.E., et al.: Changes in performance and bio-mathematical model performance predictions during 45 days of sleep restriction in a simulated space mission. Sci. Rep. **10**, 15594 (2020). https://doi.org/10.1038/s41598-020-71929-4
12. Gallucci, M.: GAMLj: general analyses for linear models. [jamovi module] (2019). https://gamlj.github.io/
13. Holm, J.E.W., du Plessis, E.: Usability - a full life cycle perspective. In: 2019 IEEE AFRICON, pp. 1–7. IEEE, Accra, Ghana (2019)
14. International Organization for Standardization: ISO Standard No. 9241–210:2019 (2019). https://www.iso.org/standard/77520.html
15. Jordan, P.W.: An Introduction to Usability. CRC Press, London (2020)
16. Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B. (eds.): Usability Evaluation in Industry. CRC Press, London (1996)
17. Lee, C., Marquez, J., Edwards, T.: Crew autonomy through self-scheduling: scheduling performance pilot study. In: AIAA Scitech 2021 Forum. American Institute of Aeronautics and Astronautics, Virtual Event (2021)
18. Marquez, J.J., Hillenius, S., Kanefsky, B., Zheng, J., Deliz, I., Reagan, M.: Increasing crew autonomy for long duration exploration missions: self-scheduling. In: 2017 IEEE Aerospace Conference, pp. 1–10. IEEE, Big Sky (2017)
19. Meza, D., Berndt, S.: Usability/sentiment for the enterprise and enterprise. NASA (2014)
20. NASA: Human Integration Design Handbook (HIDH) (2010)
21. NASA: Space Flight Human-System Standard Volume 2: Human Factors, Habitability, and Environmental Health (Vol. 2). NASA-STD-3001 (2011)

22. R Core Team: R: A language and environment for statistical computing. [Computer Software] (2021). https://cran.r-project.org
23. Ratwani, R.M., Reider, J., Singh, H.: A decade of health information technology usability challenges and the path forward. JAMA **321**, 743 (2019). https://doi.org/10.1001/jama.2019.0161
24. Schrepp, M., Hinderks, A., Thomaschewski, J.: Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In: Hutchison, D., et al. (eds.) Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience, pp. 383–392. Springer International Publishing, Cham (2014)
25. Schrepp, M., Hinderks, A., Thomaschewski, J.: Construction of a benchmark for the user experience questionnaire (UEQ). Int. J. Interact. Multimedia Artif. Intell. **4**, 40 (2017). https://doi.org/10.9781/ijimai.2017.445
26. Schrepp, M., Hinderks, A., Thomaschewski, J.: User experience questionnaire (UEQ), https://www.ueq-online.org/
27. Viitanen, J., Hyppönen, H., Lääveri, T., Vänskä, J., Reponen, J., Winblad, I.: National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int. J. Med. Inform. **80**, 708–725 (2011). https://doi.org/10.1016/j.ijmedinf.2011.06.010
28. Yulianto, D., Hartanto, R., Santosa, P.I.: Evaluation on augmented-reality-based interactive book using system usability scale and user experience questionnaire. J. RESTI (Rekayasa Sistem dan Teknologi Informasi) **4**, 482–488 (2020). https://doi.org/10.29207/resti.v4i3.1870