A Multi-Session Evaluation of a Haptic Device in Normal and Critical Conditions: a Mars Analog Mission

Julie Manon[1,2,3,7*], Jean Vanderdonckt[4,5], Michael Saint-Guillain[4], Vladimir Pletser[6], Cyril Wain[7], Jean Jacobs[7], Audrey Comein[7], Sirga Drouet[7], Julien Meert[7], Ignacio Sanchez Casla[7], Olivier Cartiaux[8], Olivier Cornu[1,3]

[1] Neuromusculoskeletal Lab (NMSK), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[2] Anatomy and Morphology Lab (MORF), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[3] Cliniques universitaires Saint-Luc, Orthopedic Surgery Department, Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[4] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[5] Louvain Research Institute in Management and Organizations (LouRIM), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[6] European Space Agency (ret.), Blue Abyss (United Kingdom)
[7] Crew 227 – Mission Analog Research Simulation (M.A.R.S. UCLouvain) – Mars, Desert Research Station (MDRS) Simulation (27 March to 10 April 2022), UT (USA)
[8] Department of Health Engineering, ECAMBrussels Engineering School, Haute Ecole "ICHEC-ECAM-ISFSC", Brussels (Belgium)

* Corresponding author: julie.manon@uclouvain.be

Received 26 May 2023 | Accepted 7 February 2025 | Published 11 April 2025

Abstract

While visual interaction is typically evaluated as an instantaneous, one-shot activity that considers only a snapshot of factors, haptic interaction is more challenging to evaluate as it involves a continuous touch process evolving over time. To better understand how to evaluate haptic interaction, this paper performs a multi-session evaluation of a haptic device to be used by astronauts in future lunar and Mars missions, based on eight factors. Three groups of two members ($n = 6$) applied, either as operator or assistant, a newly developed external fixator (EZExFix) to fix a fracture of the tibial shaft. Astronauts had different levels of expertise, i.e., in anatomy, mechanical engineering, and without, and participated in eight timed runs. Among these eight matches, four sessions were conducted with different time frames and compared to a stress test, a reproduction of the experiment in very stressful conditions, and a session simulating critical conditions in an extra-vehicular activity.

Keywords

Context of Use, Haptic Device, Multi-Session, User Experience, UEQ+.

DOI: 10.9781/ijimai.2025.04.001

I. Introduction

Haptic interaction typically promotes the sense of touch as an alternate modality to visual interaction [1] when the visual channel can be occupied, overwhelmed, or simply constrained by other factors, such as in critical conditions. While the visual channel is instantaneous, as for immediate feedback, haptic interaction involves tactile sensations which are part of our somatosensory system, a system that is rather continuous and not as instantaneous as the visual channel. Using a haptic device requires physical manipulation in real time that necessarily involves collision detection and effort to compensate for it. Learning haptic interaction is a continuous process with variations, as for gestural [2] and vocal interaction [3]. For these rea-sons, evaluating the haptic interaction that people can have with a physical device is not just a one-shot action but should be continuously examined over time to capture how people progressively acquire, manipulate, and react to such a haptic device, or, in other words, its evolution over time.

Although there are several methods to quantitatively evaluate a haptic device, they are mostly device-dependent or metric-dependent, which makes them challenging to transpose to another context of use [4]. In contrast, qualitative methods are device-independent but are usually self-reported questionnaires that are limited in scope or not very specific. To better understand how to evaluate a haptic device in

Please cite this article as: J. Manon, J. Vanderdonckt, M. Saint-Guillain, V. Pletser, C. Wain, J. Jacobs, A. Comein, S. Drouet, J. Meert, I. Sanchez Casla, O. Cartiaux, O. Cornu. A Multi-Session Evaluation of a Haptic Device in Normal and Critical Conditions: a Mars Analog Mission, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 164-174, 2025, http://dx.doi.org/10.9781/ijimai.2025.04.001

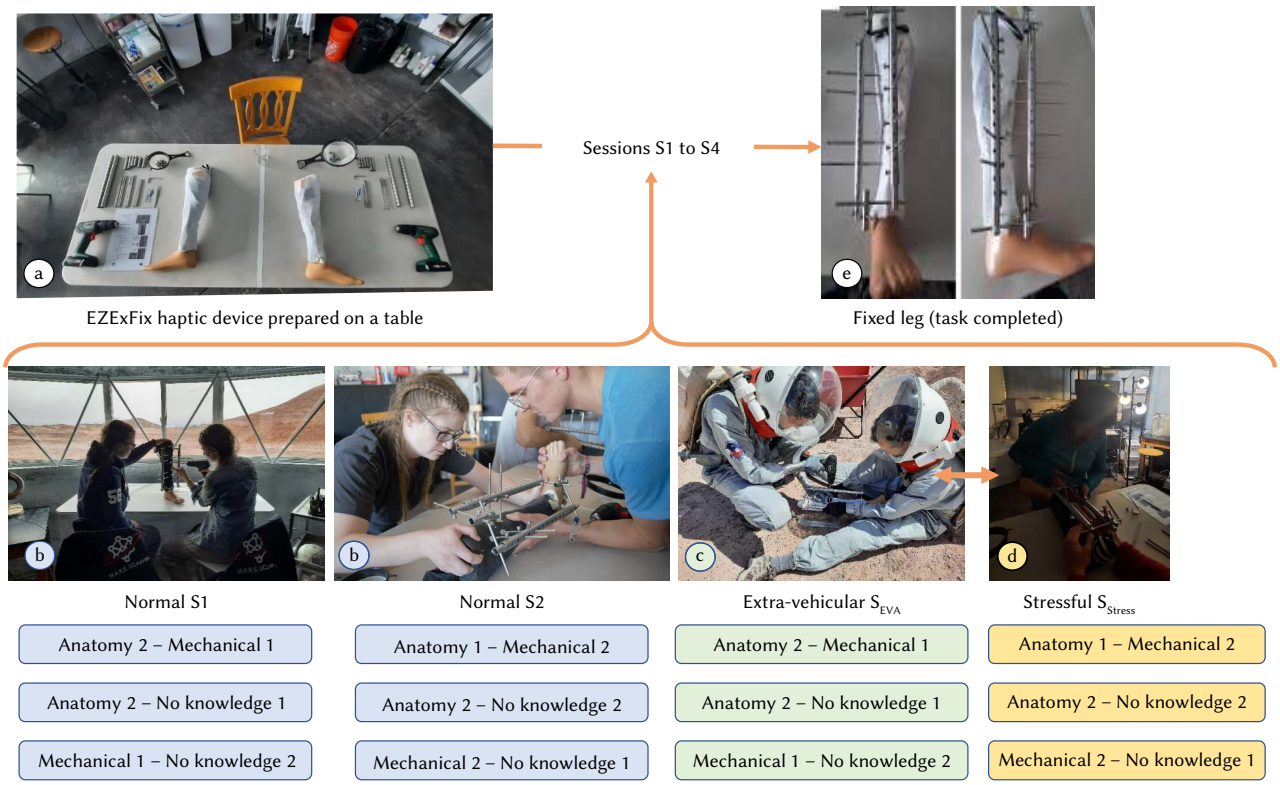


Fig. 1. Overview of the experiment: (a) initial preparation of material needed for 2 simultaneous surgeries, (b) sessions covering various conditions, including (c) extra-vehicular activity (EVA) and (d) under stress, (e) final results. Sessions S1 to S4were organized in 12 timed runs (bottom). Six analog astronautswere divided into 3 groups based on their educational backgrounds: with knowledge in anatomy ("Anatomy"), with knowledge in mechanical engineering ("Mechanical"), and without any knowledge in anatomy and engineering ("No knowledge"). Each person was identified by 1 or 2.

this context, we wanted a device that supports haptic assembly [5], as it incorporates a complete haptic effect for mounting and dismounting operations. The task of haptic assembly consists of combining the mechanical joints of a device while focusing on the guidance of objects and the activation signals of the kinematic constraints posed by the device [6].

For this purpose, we selected the EZExFix, a low-cost, fast, and easy-to-use external fixator to handle tibial shaft fractures, which are among the most common open or closed long bone fractures [7][8][9].

The EZExFix consists of a metallic device made up of pins that are inserted into the fractured bone, connecting rods outside the leg (Fig. 1-a and 1-e). The background of its creation, purpose, and validation has been previously published [10]. To compare the operation of such a device in normal and critical conditions, we applied this newly developed EZExFix in realistic operational conditions on Mars during a two-week simulation mission at the Mars Desert Research Station (MDRS, Utah, USA) [11][12].

II. Background

The evaluation of haptic interfaces has been the subject of a great deal of work [13][14] in the context of a range of haptic applications or a particular interactive application with haptic use, mainly in games, virtual reality [4], and machines [6]. For example, Hamam and Saddik [15] pro-posed a mathematical model to evaluate the quality of experience of haptic-based applications, which has been validated through a user study, showing that a Principal Component Analysis performs slightly better than other approaches. Höver et al. [16] presented a user-based evaluation of data-driven haptic rendering, emphasizing the importance of dynamic material effects for achieving realistic haptic feedback. While these studies evaluate the haptic modality in isolation, they acknowledge the need for evaluating both graphical and haptic elements. After reviewing a series of physical and psychophysical metrics used for evaluating a haptic interface [17], Samur derived a psychophysical method for evaluating a force-feedback device [14], which includes guidelines for characterizing such a device along the new dimensions. The specific functions of vibration [13] and sensitivity and friction [18] have been also addressed for a haptic device.

In sum, existing methods focus mainly on the haptic modality, either in general using a model or in particular for a certain type of application in an activity domain. They do not put into perspective evaluation along several dimensions of usability or user experience in a uniform way. For these reasons, we chose a method that evaluates different dimensions in the same way to compare them with each other and across different sessions.

III. Multi-Session Evaluation

A. Participants and Sessions

Three groups of two analog astronauts ($n = 6$) were recruited from the crew 227, who participated in the Tharsis 2022 mission at the MDRS, depending on their level of expertise, established based on their respective degrees or studies: with knowledge in anatomy ("Anatomy"), with knowledge in mechanical engineering ("Mechanical"), and without any knowledge in anatomy and engineering ("No knowledge"). On the first day of the mission, these three groups first attended a short theoretical course on the indications, anatomical landmarks, and steps of EZExFix setup for 1 hour followed by a practical demonstration. Then, they had to perform the task one after another, sometimes playing the role of *operator*, who put the EZExFix on the broken leg (called for example "Anatomy 1" or

"Anatomy 2" depending on the person in the "Anatomy" group - Fig. 1, bottom), sometimes in the role of assistant, who helps to maintain the fracture reduction. Each astronaut met each other in timed runs during which they had to set up the EZExFix to repair an artificial tibial shaft fracture (Fig. 1), in the most efficient way and in the least possible time. Therefore, the number of timed runs consisted of twelve blocks that covered both operators and assistant pairs. Within these blocks, each person was given four times the role of the operator and evaluated on each trial achievement, totaling 24 trials ($N = 24$).

The different groups can also be compared in terms of skills to assess the need to have basics in anatomy or mechanics. Since a fracture could occur in space and therefore be stressful, different conditions were evaluated. The trials were scheduled at the MDRS [19] in good conditions with all the instrumentation prepared on a table (in blue in Fig. 1), during an *extravehicular activity* (EVA) with space suits (in green) or at an unexpected moment, such as at night or dinner, with nothing prepared (in yellow), both considered stressful conditions. The extensive and detailed information has previously been covered [20][21].

B. Design, Measures, and Protocol

Demographic information was collected from the participants before the beginning of the mission. Parti-cipants were instructed to complete a UEQ+ questionnaire (User Experience Questionnaire) [22], a modular extension of the UEQ evaluation method in which eight scales were selected, *i.e.*, Attractiveness, Efficiency, Perspicuity, Trust, Adaptability, Usefulness, Intuitive use, and Haptics, among the 20 scales available to evaluate the user experience of participants interacting with the haptic device. We chose these eight scales for the following reasons: the original UEQ [23] includes Attractiveness as the topmost scale covering (Fig. 2) pragmatic and hedonic qualities, in particular with Perspicuity, Efficiency; we did not consider Dependability, Stimulation and Novelty from the original UEQ as we preferred to focus on more relevant scales and avoid fatigue; we selected the other scales to preserve a balance between these qualities and incorporated explicitly Adaptability to investigate how different users could accommodate the EZExFix, Trust [24] to assess the confidence of the participants, and Haptics [25] because of the haptic nature of the device. These scales typically refer to a question, which is specifically tailored for our experiment:

- Attractiveness: do users like the EZExFix?;
- Efficiency: can users apply the EZExFix without unnecessary effort?;
- Perspicuity: is it easy to get familiar with the EZExFix?;
- Trust: are the users not harmed by the EZExFix?;
- Adaptability: can the EZExFix be adapted to various working styles?;
- Usefulness: does using the EZExFix bring benefits?;
- Intuitive use: can the EZExFix be used immediately without any training or help?;
- Haptics: what is the haptic feeling resulting from using the EZExFix?

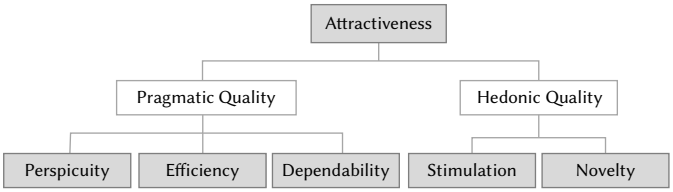


Fig. 2. Initial Scale structure of UEQ [22].

Attractiveness is an overall positive or negative impression of the product, while Perspicuity and Efficiency are hard aspects of the user experience representing the pragmatic quality of the device. Users typically perceive products with greater pragmatic quality as intuitive to use, efficient, and trustworthy.

UEQ+ was used to compare various designs of Playbook, a self-scheduling software used by astronauts [26]. Each scale is, in turn, decomposed into four subscales or items to be evaluated (*e.g.*, Attractiveness is decomposed into annoying *vs.* enjoyable, bad *vs.* good, unpleasant *vs.* pleasant, and unfriendly *vs.* friendly), each subscale being a differential scale with 7 points between items of each pair (*e.g.*, annoying ⟶ enjoyable). Each item is measured employing a 7-point Likert-type scale with answer categories "Strongly disagree" (=1) to "Strongly agree" (=7). UEQ+ is selected as an evaluation method because it is a modular and modern evaluation method where scales can be decided based on the artifact to evaluate and cover its user experience, not just its usability. UEQ+ is also straightforward and cost-effective to administer to participants. The number of participants required is still an open question [27]: recruiting (analog) astronauts is a challenging task as very few candidates are available. Furthermore, for some scales, a comparison of their values leads to an interpretation of five effect sizes [28]: bad, below average, above average, good, and excellent. Our within-subject study design has two dependent variables:

1. The Scale mean score, a real variable that measures the average score obtained on all items on each scale for each of the six sessions $S_i \in \{S_1, S_2, S_3, S_4, S_{EVA}, S_{Stress}\}$.
2. The Scale importance, a real variable that measures the average weight of importance of each scale for each of the six sessions $S_i \in \{S_1, S_2, S_3, S_4, S_{EVA}, S_{Stress}\}$.

Participants' answers are computed with the UEQ data analysis tool and interpreted as follows [22]: "it is extremely unlikely to observe values above +2 or below -2,..., the standard interpretation of the scale means is that values between -0.8 and 0.8 represent a neutral evaluation of the corresponding scale, values superior to 0.8 represent a positive evaluation, and values inferior to -0.8 represent a negative evaluation". Based on this interpretation, the results obtained for the multi-session evaluation are first discussed regarding the global results for all scales, then regarding each individual scale.

C. Inter-Scale Global Results and Discussion

1. Interrater Consistency

Table I reports Cronbach's coefficient α [29] computed to quantify the internal consistency, which expresses the extent to which the scale measurements remain consistent within a session or over subsequent sessions under identical or different conditions. A high value indicates that the answers of participants across items are consistent. When participants give a high value for one of the scale items, they are also likely to provide high values for the other items. The mean coefficient for all scales on all sessions is $\alpha = .66$, which suggests a global questionable consistency, but close to $\alpha = 0.7$, which is considered as an acceptable consistency. Trust ($\alpha = 0.91$), Intuitive use ($\alpha = 0.86$), and Usefulness ($\alpha = 0.80$) received the highest values, thereby indicating that these three scales were consistently assessed by participants.

Although other scales received reasonably good values, Haptics ($\alpha = 0.07$) received the lowest value, highlighting that participants did not assess this scale uniformly, probably because they belong to three different profiles. Depending on their knowledge, they assessed in different ways this scale, which seems to be more profile-dependent as opposed to the others. On the one hand, the diversity of these profile categories improves the representativeness of participants but, on the other hand, they tend to lower their consistency. Among all items,

TABLE I. Interrater Consistency. Cronbach's α: α ≥ 0.9 = Excellent (E), 0.9 > α ≥ 0.8 = Good (G), 0.8 > α ≥ 0.7 = Acceptable (A), 0.7> α ≥ 0.6 = Questionable (Q), 0.6>α ≥ 0.5= Poor (P), α < 0.5 = Unacceptable (U)

Scale	Cronbach's α (interpretation)						
	S_1	S_2	S_3	S_4	Mean	S_{EVA}	S_{Stress}
Attractiveness	0.46 (U)	0.79 (A)	0.87 (G)	0.10 (Q)	0.55 (P)	0.93 (E)	0.89 (G)
Efficiency	0.69 (Q)	0.71 (A)	0.38 (U)	0.88 (G)	0.66 (Q)	0.55 (P)	0.90 (E)
Perspicuity	0.37 (U)	0.78 (A)	0.88 (G)	0.69 (Q)	0.70 (A)	0.62 (Q)	0.90 (E)
Trust	0.95 (E)	0.95 (E)	0.93 (E)	0.81 (G)	0.91 (E)	0.73 (A)	0.91 (E)
Adaptability	0.64 (Q)	0.82 (G)	0.10 (U)	0.81 (G)	0.59 (P)	0.85 (G)	0.15 (U)
Usefulness	0.73 (A)	0.96 (E)	0.66 (Q)	0.83 (G)	0.80 (G)	-0.63 (U)	0.91 (E)
Intuitive Use	0.82 (G)	0.92 (E)	0.87 (G)	0.81 (G)	0.86 (G)	0.43 (U)	0.80 (G)
Haptics	0.21 (U)	0.42 (U)	0.26 (U)	-0.61 (U)	0.07 (U)	-0.01 (U)	-0.64 (U)
Mean	0.61 (Q)	0.79 (A)	0.69 (Q)	0.54 (P)	0.66 (Q)	0.43 (U)	0.60 (Q)

TABLE II. Interrater Reliability. Kendall's W: $W \leq 0.2$ = Poor (P), $0.21 \leq W \leq 0.4$ = Fair (F), $0.41 \leq W \leq 0.6$ = Moderate (M), $0.61 \leq W \leq 0.8$ = Good (G), and $0.81 \leq W \leq 1$=Very Good (V)

Scale	Kendall's W (p-value, interpretation)						
	S_1	S_2	S_3	S_4	Mean	S_{EVA}	S_{Stress}
Attractiveness	0.35 (0.096, F)	0.11 (0.54, P)	0.10 (0.59, P)	0.18 (0.35, P)	0.185 (P)	0.31 (0.12, F)	0.23 (0.24, F)
Efficiency	0.075 (0.71, P)	0.31 (0.12, F)	0.086 (0.67 P)	0.067 (0.75, P)	0.134 (P)	0.042 (0.82, P)	0.13 (0.49, P)
Perspicuity	0.51 (0.025, M)	0.33 (0.11, F)	0.36 (0.084, F)	0.31 (0.13, F)	0.38 (F)	0.44 (0.047, M)	0.33 (0.10, F)
Trust	0.15 (0.41, P)	0.15 (0.41, P)	0.15 (0.44, P)	0.15 (0.44, P)	0.15 (P)	0.19 (0.32, P)	0.061 (0.77, P)
Adaptability	0.16 (0.38, P)	0.21 (0.26, F)	0.046 (0.82, P)	0.053 (0.81, P)	0.18 (P)	0.22 (0.25, F)	0.078 (0.70, P)
Usefulness	0.30 (0.14, F)	0.15 (0.44, P)	0.27 (0.17, F)	0.11 (0.54, P)	0.21 (F)	0.21 (0.26, F)	0.13 (0.49, P)
Intuitive Use	0.42 (0.053, M)	0.067 (0.75, P)	0.50 (0.029, M)	0.25 (0.20, F)	0.31 (F)	0.37 (0.08, F)	0.19 (0.32, P)
Haptics	0.28 (0.16, F)	0.29 (0.15, F)	0.41 (0.059, M)	0.45 (0.043, M)	0.36 (F)	0.51 (0.026, M)	0.30 (0.14, F)
Mean	0.28 (F)	0.21 (F)	0.24 (F)	0.21 (F)	0.24 (F)	0.28 (F)	0.18 (P)

Item1 "Stable-Unstable" ($M = 1.83$), Item3 "Rough-Smooth" ($M = 0.83$), and Item2 "Unpleasant to touch-Pleasant to touch" ($M = 0.50$) were rather positively assessed while Item4 "Slippery-Smooth" ($M = 0.0$) was rated as null. Perhaps the label "Smooth" shared by two bipolar scales confused participants. Three inter-item correlations of this scale were negative, Corr(Item3, Item4) = −0.27, Corr(Item1, Item2) = −0.10, and more surprisingly Corr(Item1, Item4) = −0.80, thereby suggesting that participants did not understand the items in the same way as the low values for some items counterbalanced the high values of other items, which creates a null effect. We therefore re-computed Cronbach's coefficient with missing items to obtain: $\alpha_{Item1} = 0.38$, $\alpha_{Item2} = 0.71$, $\alpha_{Item3} = 0.92$, and $\alpha_{Item4} = 0.66$. Remo-ving Item4 has a positive impact in our case.

In general, S_1 started with a questionable mean value ($\alpha = 0.61$), then increased to acceptable in S_2 ($\alpha = 0.79$) to return to a questionable one in S_3 ($\alpha = 0.69$) to sum up finally with an almost acceptable mean value ($\alpha = 0.66$). Haptics also drags the average consistency down from an acceptable global value ($\alpha = 0.72$ without Haptics) to a questionable one ($\alpha = 0.66$ with Haptics). However, the S_{Stress} situation ($\alpha = 0.60$), although fairly close, is interpreted as questionable.

2. Interrater Reliability

Since performance assessment in essential to experiments, interrater consistency and reliability are two indices that are commonly used to ensure such scoring consistency [30]. Therefore, after computing and reported the interrater consistency, we compute and evaluate the interrater reliability. Table II reports Kendall's coefficient of concordance W [31], a measure of agreement among participants which is equal to 0 when there is no agreement among them and 1 when a total agreement exists. The scales receiving the highest values are Perspicuity ($W = 0.38$), Haptics ($W = 0.36$), Intuitive Use ($W = 0.31$), and Usefulness ($W = 0.21$), all interpreted as a fair agreement. All other scales received a limited agreement when $W \leq 0.2$. Similarly, all sessions benefit from a fair agreement ($W \geq 0.21$), including the mean overall coefficient. Since Kendall's coefficient is very strict and demanding in its value, a fair value was not considered a disadvantage in our case, given the heterogeneity of the participants' profiles.

3. Scale Mean Scores and Importance

Fig. 3 shows the mean scores for the eight scales evalu-ated, each time across the four sessions (see Appendix for the detailed histograms for the four sessions). As a reference, S_{EVA} and S_{Stress} are represented each time as a horizontal bar calibrated on the corresponding mean scores. The mean scores and scale importance are positive for all scales for all sessions in this figure, even the most critical ones, which is very encouraging, thus explaining why the vertical left axis is reduced to [−1 ...3] instead of [−3... +3]. Only six out of 32 mean scores are below the 0.8 threshold, thus locating them in the neutral zone while all others are located in the positive one. Although below this threshold, the mean scores are not very far away: *e.g.*, Efficiency received 0.75 and 0.79 for S_1 and S_2, respectively, Adaptability received 0.67 and 0.79 for S_1 and S_2, respectively, with the exception that S_2 received the lowest mean score 0.33 for Haptics of all scales on all sessions.

Furthermore, a Wilcoxon signed rank test was calculated for a single sample to test significant differences above the median for each subscale for each session to discover that all the means of each subscale for each session were above their respective medians (*e.g.*, Attractiveness for S_1 gave a highly significant difference, $score = 3.10$, $p = 0.0007$*** with a large effect size $r = 0.63$) with only one exception (*i.e.*, Attractiveness for S_2 gave $z\text{-}score = 1.35$, $p = 0.09$, n.s.).

Shapiro-Wilk and d'Agostino-Pearson tests of normality were computed to determine whether the scale and sub-scale data are normally distributed. These results advised the rejection of the null hypothesis for all data and concluded that the data were not normally

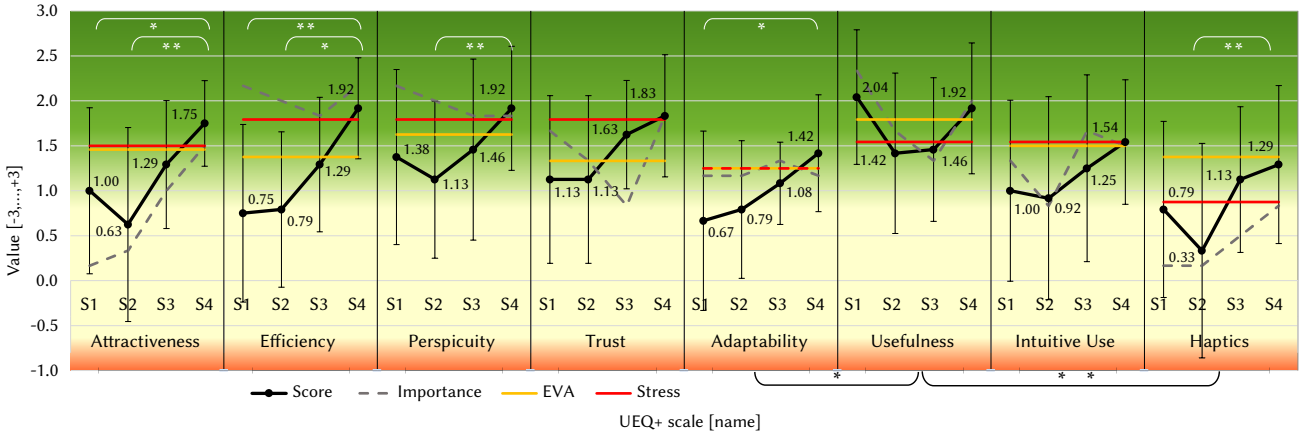


Fig. 3. Panel chart of the eight scales evaluated: mean scores (black bold straight lines) and scale importance (grey dotted lines) for sessions S_1 to S_4. Yellow lines show mean scores for SEVA and red lines show mean scores for SStress. Error bars show a confidence interval of 95% over mean scores. *: $p < 0.05$, **: $p < 0.01$.

distributed, since at least one test failed. Consequently, in the remainder of this paper, we will compute a non-parametric Friedman test of differences among repeated measures for all scales for all sessions with Dunn's test for multiple comparisons.

The first series of tests were carried out for the eight scales in all sessions and produced a Friedman statistic value of 22.61, which was significant ($p = 0.002$**): Usefulness is scored higher than Adaptability ($R = -56$, $p = 0.0271$*) and than Haptics ($R = 66$, $p = 0.028$**, see the bottom part of Fig. 3).

Overall, the mean scale scores start at S_1 with a rather positive value, then decrease or remain at the same level at S_2 to progressively increase again at S_3 and even more at S_4. For example, Attractiveness progresses as follows: it starts at $M_{S_1} = 1.00$, then decreases to $M_{S_2} = 0.63$, then increases to $M_{S_3} = 1.29$ to end at $M_{S_4} = 1.75$. All scale curves are globally increasing curves, *i.e.*, $\forall S_i \in \{S_1, S_2, S_3, S_4\}$: $M_{S_1} < M_{S_4}$ except for Usefulness but with very close values, *i.e.*, $M_{S_1} = 2.04 \geq M_{S_4} = 1.92$.

Contrary to an S-shaped performance curve that progressively increases until reaching a plateau or to an adoption curve that could decrease after the plateau, the eight factors that were continuously evaluated through a multi-session tend to follow a hype cycle curve. This type of curve starts with figures expressing a high expectancy in the device, then progressively decreases as the device is more frequently used in difficult and various conditions, to end up with a final increase to converge to a plateau expressing the final assessment of the device. For example, the important Haptics scale starts with a moderate mean score, then decreases and increases to end up with a more positive score. The importance is rated similarly.

At S_1, scales are sorted in decreasing order of their mean scores as follows: Usefulness ($M = 2.04$), Perspicuity ($M = 1.38$), Trust ($M = 1.13$), Attractiveness ($M = 1.00$), Intuitive Use ($M = 1.00$), Haptics ($M = 0.79$), Efficiency ($M = 0.75$), and Adaptability ($M = 0.67$). At S_4, this order remains mostly the same: Usefulness ($M = 1.92$), Perspicuity ($M = 1.92$), Efficiency ($M = 1.92$), Trust ($M = 1.83$), Attractiveness ($M = 1.75$), Intuitive Use ($M = 1.54$), Adaptability ($M = 1.42$), and Haptics ($M = 1.29$). Only the last two scales swapped their order, with Haptics slightly decreased but the Efficiency climbed up to the 3rd position. This result suggests that while the value of scale mean scores increased over sessions, participants tend to rate them in the same order except for the Efficiency that scales up throughout the training.

The mean scores of the scale for the stressful condition S_{Stress} (red lines in Fig. 3) are most of the time above the corresponding scores for the extra-vehicular condition S_{EVA}, except for Usefulness and Haptics. They even coincide for Attractiveness, Adaptability, and Intuitive Use, thus suggesting the real conditions in space may affect the user experience with respect to in-lab conditions, even with some stress imposed. Mean importance (grey dotted lines in Fig. 3) seems to follow the same hype cycle as scale mean scores, except for Usefulness.

4. Benchmarking of Scales

Schrepp *et al.* [28] mention more precise intervals to interpret some scales based on benchmarking performed on a set of evaluations. Each scale is decomposed into five intervals based on this benchmarking: bad, below average, above average, good, and excellent. Fig. 4 shows the interval in which the three concerned scales, *i.e.*, Attractiveness, Perspicuity, and Efficiency, are falling over the four sessions, from S_1 to S_4.

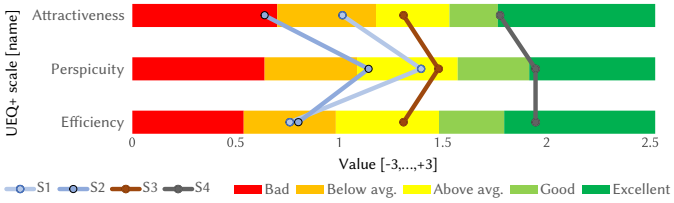


Fig. 4. Benchmarking of scales for all sessions S_i.

5. Key Performance Indicators

Another recent development of UEQ+ is the construction of the Key Performance Indicators (KPI) extension [32]. The KPI combines the subjectively perceived importance of user experience factors and the results of the UEQ+ into one figure. Fig. 5 shows the KPI for all sessions S_i. All mean KPIs are above the positive threshold of 1, except S_2 with a close value ($M = 0.94$, $SD = 0.75$). Indeed, S_1 ($M = 1.11$, $SD = 0.37$) initiates the multi-session that ends up with almost the same value in S_4 ($M = 1.11$, $SD = 0.36$). Interestingly, the KPI for the EVA ($M = 1.47$, $SD = 0.23$) and for the stress conditions ($M = 1.54$, $SD = 0.54$) are above the final value, thereby suggesting that participants were particularly attentive in expressing a higher performance in those critical conditions as opposed to normal ones. However, a non-parametric Kruskal-Wallis test revealed that there are no significant differences ($H(5) = 5.79$, $\alpha = 0.05$, $p = 0.33$, *n.s.*) between the six KPIs.

D. Intra-Scale Results and Discussion

This section gathers all results for one scale at a time and consolidates them into a dedicated discussion considering scale mean scores (Fig. 3), and their mean importance (Fig. 6). Furthermore, this section concludes the individual discussion of each scale with the results of

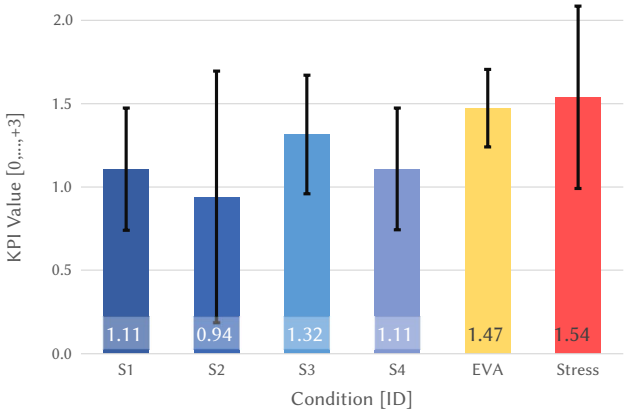


Fig. 5. Key Performance Indicators (KPI) for all sessions S_i. Error bars show the standard deviation.

their Importance-Performance Analysis (IPA) [33]. This analysis aims to assign every scale to four different quadrants determined by two methods: (1) a differentiation by the coordinate origin at (0, 0), which is represented by a solid green line in Fig. 7; (2) a differentiation by the coordinate origin in the mean value of all scale values, which is represented by green dotted lines. Since we were interested in assessing the evolution of user experience of the EZExFix device over time, Fig. 7 shows the transition from the initial session S_1 to the final S_4: the X axis shows the performance computed as the scale mean score for the related session (S_1 and S_4, respectively) while the Y axis shows the mean importance for each scale. The dotted lines represent the average of all scale mean scores on X and the average of all mean importance on Y . The blue arrows show the transition from S_1 to S_4 for each single scale. Thus, each quadrant provides a recommendation for action for the respective scales, depending on its positioning. The plot is therefore divided into four quadrants [33]: Q1 = "Keep Up the Good Work" (top right quadrant when both the performance and the importance are above the corresponding mean value), Q2 = "Possible Overkill" (bottom right quadrant when the performance is above the corresponding mean value and the importance is below), Q3 = "Low Priority" (bottom left when both the performance and the importance are below the corresponding mean value), and Q4 = "Concentrate Here" (top left when the performance is below the mean value but the importance is above the mean value). These results are summarized in Table III. Note that all scales are located in Q1 with respect to the center of the scale at (0, 0) as they were all positive in terms of the mean values of the scale (Fig. 3) and the mean importance (Fig. 6).

Attractiveness. This scale is probably the most important among all scales assessed since it is supposed to capture "a user's general impression", one of the three dimensions of user experience. The

TABLE III. Assignment Scales to IPAquadrants According to the Two Methods. Each Quadrant Provides a Recommendation for Action: Q1="Keep Up the Good Work", Q2="Possible Overkill", Q3="Low Priority", Q4="Concentrate Here"

Scale	Scale Center		
	(0, 0)	Mean *S*1	Mean *S*4
Attractiveness	Q1	Q3	Q2
Efficiency	Q1	Q4	Q1
Perspicuity	Q1	Q1	Q1
Trust	Q1	Q1	Q1
Adaptability	Q1	Q3	Q3
Usefulness	Q1	Q1	Q1
Intuitive Use	Q1	Q3	Q3
Haptics	Q1	Q3	Q3

perceived attractiveness of an artifact is considered to be the result of an averaging process of the perceived quality of the software with respect to the relevant aspects in a given usage scenario [23]. The mean score ($\pm SD$) of Attractiveness was 1.00 ($\pm$1.15) at baseline S_1, 0.63 ($\pm$1.35) at S_2, 1.29 ($\pm$0.89) at S_3 and 1.75 ($\pm$0.60) at S_4. All figures are above their respective importance values. The S_4 mean value is similarly above the critical conditions M_{EVA} and M_{Stress}, thus exceeding the expectations (Fig. 3). The hype cycle is even more revealing for this scale: a Friedman test ($F = 19.10$, $n = 4$) shows a significant difference ($p = 0.0003$***) between sessions. Post hoc analysis with Dunn's multiple comparison tests was performed with a Bonferroni correction applied, resulting in a significance level set at $p < 0.05$. A significant increase was observed between S_1 and S_4 ($R = -24.00$, $p = 0.0437$ *) and between S_2 and S_4 ($R = -28.50$, $p = 0.086$**). Furthermore, the standard deviation is progressively reduced as sessions progress: from $SD_{S1} = 1.15$ to $SD_{S4} = 0.60$. The overall good assessment is reinforced by a final 'excellent' position in benchmarking (Fig. 4) and a "Q2=Possible overkill" position (Table III). This should be mitigated by varying interrater consistency (limited on average, but good in critical conditions) and reliability (again limited on average, but fair in critical conditions), probably due to the small number of people having heterogeneous profiles.

Efficiency. This scale is considered to be a pragmatic quality of user experience and is "goal-directed" (its assessment is based on tasks that can be performed with the device) [28]. The mean score ($\pm SD$) of Efficiency was 0.75 (1.23) at baseline S_1, 0.79 (1.08) at S_2, 1.29 (0.93) at S_3 and 1.92 (0.70) at S_4. All figures are below their respective importance values. Typically, expectations are met when the mean values are equal to or greater than their importance. However, in this case, all values are highly positive, the difference between both values at S_4 is small and the S_4 mean value is still above the critical

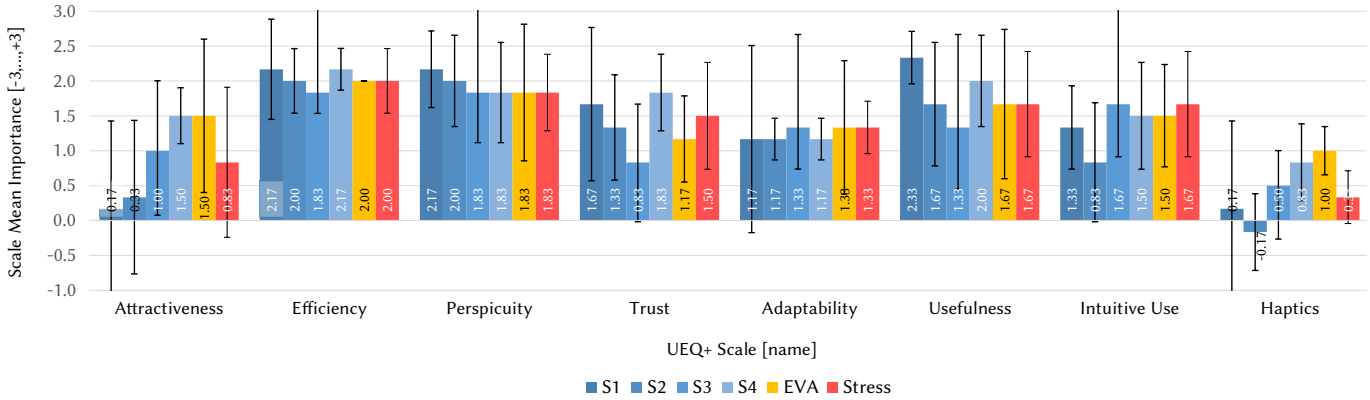


Fig. 6. Scale importance for sessions S_1 to S_4, S_{EVA}, and S_{Stress}. Error bars show a confidence interval of 95%.

conditions M_{EVA} and M_{Stress}, thus being anyway interpreted as a good assessment (Fig. 3). Similarly to Attractiveness, this scale also knows a significant increase between the initial session S_1 and the final session S_4 ($R = -32.00$, $p = 0.0021$**, *z-score* = 3.578) and between S_2 and S_4 ($R = -26.50$, $p = 0.0183$*, *z-score* = 2.963). Efficiency ends with the highest mean score of all scales ($M_{S_4} = 1.92$) and is interpreted as 'excellent' in the benchmarking (Fig. 4). Although this scale was located in the "Q4=Concentrate Here" quadrant during S_1, its evolution reaches the best quadrant "Q1=Keep up the Good Work" during the final session S_4 with the best position of all scales (Fig. 7). The interrater consistency is better than that of Attractiveness, but with similar reliability for the same reasons.

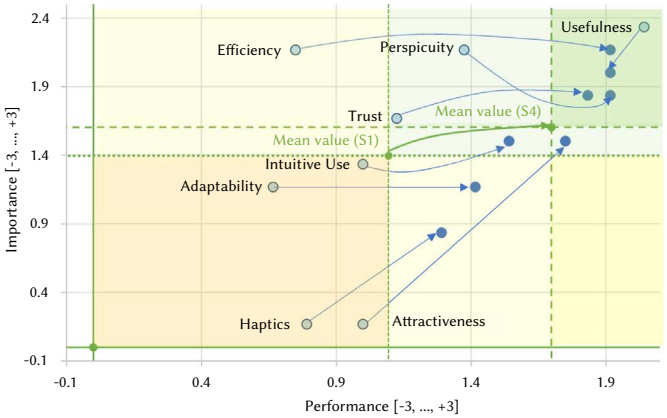


Fig. 7. Results of the IPA analysis: transition from S_1 to S_4.

Perspicuity. This scale expresses to what extent participants considered it easy to get familiar with the EZExFix device and to learn how to use it. Therefore, it is also considered a pragmatic quality that is an important part of the user experience and is "goal-directed" [28]. The mean score ($\pm SD$) of Perspicuity was 1.38 (1.22) at baseline S_1, 1.13 (1.09) at S_2, 1.46 (1.26) at S_3 and 1.92 (0.86) at S_4. The S_4 mean value is above its importance and critical conditions (both M_{EVA} and M_{Stress}) (Fig. 3). Several signs concur to conclude that this scale is very positively and rigorously assessed: its importance remains consistently estimated across sessions (Fig. 6), it knows a significant increase only between S_2 and S_4 ($R = -29.50$, $p = 0.0058$**, *z-score* = 3.298) (Fig. 3), it is interpreted as 'excellent' in the benchmarking (Fig. 4), it consistently stays in the quadrant Q1="Keep Up the Good Work" during all sessions considered, and it is the third of all scales in this quadrant (Fig. 7).

Trust. This scale expresses the extent to which participants are confident in the use of the device, in its correct functioning, and, above all, that it will not harm them, which is crucial in the case of a limb fracture. The mean score ($\pm SD$) of Trust was 1.13 (1.17) at baseline S_1 and at S_2, 1.63 (0.75) at S_3 and 1.83 (0.85) at S_4. The S_4 mean value is nearly equal to its importance and M_{Stress}, and above M_{EVA} (Fig. 3). There were no significant differences among matched UEQ+ items of this scale between the different sessions (Fig. 3). It stays consistently in quadrant Q1 = "Keep Up the Good Work" during all sessions considered ending in the fourth place of all scales (Fig. 7), thereby suggesting that it was positively recognised, especially with excellent average consistency (the best among all scales).

Adaptability. This scale expresses the extent to which participants felt that the device could be adapted to a range of personal parameters, such as their own physical configuration, preference, or individual way of working. Although we did not take into account a factor of the participants' physical morphology, which could be considered in this scale, the participants did not express a negative opinion in this respect. The mean score ($\pm SD$) of Adaptability was 0.67 (1.25) at baseline S_1, 0.79 (0.96) at S_2, 1.08 (0.57) at S_3 and 1.42 (0.81) at S_4 (Fig. 3). This shows that Adaptability received the lowest, yet neutral, scores during S_1, but that slightly increases over sessions until a positive value ($M_{S_4} = 1.42$) is obtained. This scale knows a significant increase only between S_1 and S_4 ($R = -25.00$, $p = 0.03118$*, *z-score* = 2.795) (Fig. 3). This scale received low mean importance, and participants agreed to rate this scale below the corresponding means of all scales. Adaptability remains in the Q3="Low Priority" during S_1 to S_4 (Fig. 7), thereby suggesting that any form of adaptation is not that important for the participants.

Usefulness. This scale expresses how useful the participants felt the device was in fixing a broken leg, which is already a critical situation, even though it was assessed under normal, stressful conditions without any participant actually having a limb with a fracture. The mean score ($\pm SD$) of Usefulness was 2.04 (0.93) at baseline S_1, 1.42 (1.11) at S_2, 1.46 (1.00) at S_3 and 1.92 (0.91) at S_4. All the mean values are close to their corresponding scale importance and end up above the critical conditions (both M_{EVA} and M_{Stress}) (Fig. 3). Among all scales, Usefulness received the highest mean scores both in S_1 and S_4, which makes this scale the most positively assessed. This is particularly important since the main goal of the device lies in its usefulness first, and in its user experience second. There were no significant differences among matched UEQ+ items of this scale between the different sessions. This scale remains uniformly located in Q1="Keep Up the Good Work" both during S_1 and S_4 sessions (Fig. 7).

Intuitive Use. This scale expresses the extent to which participants were able to manipulate the device in the task assigned to them with minimal use of any form of assistance or guidance. The mean score ($\pm SD$) of Intuitive Use was 1.00 (1.26) at baseline S_1, 0.92 (1.41) at S_2, 1.25 (1.30) at S_3 and 1.54 (0.87) at S_4 (Fig. 3). The values during S_1 and S_4 were below, respectively similar with their corresponding importance and under critical conditions (both M_{EVA} and M_{Stress}). Thus, the mean score of this scale is only aligned with that of M_{EVA} and M_{Stress} when the last session S_4 was reached. There were no significant differences between the UEQ+ items matched on this scale between the different sessions ($F = 5.432$, $p = 0.1427$, *n.s.*)(Fig. 3). This scale remains in Q3="Low Priority" during S1 and S4, thus suggesting that the participants really needed the familiarisation to properly operate the device and that some effort should be devoted to improving this aspect.

Haptics. This scale expresses the extent to which participants felt their touch when handling the device, which is probably the most important factor as the device is supposed to provide the user with a sense of physical touch that best fits the task, i.e.,setting a fracture. This is largely covered by the pins of the device, but also by their configuration and handling. The mean score ($\pm SD$) of Haptics was 0.79 (1.22) at baseline S_1, 0.33 (1.49) at S_2, 1.13 (1.01) at S_3 and 1.29 (1.10) at S_4 (Fig. 3). None of the participants had any previous experience with such a device, nor did they have any experience in treating a fracture in a mission as perilous as that which one might imagine in space, on the moon, or on another planet such as Mars. This scale knows a significant increase only between S_2 and S_4 ($R = -31.00$, $p = 0.0032$**, *z-score* = 3.466) (Fig. 3). Of all the scales studied, this one had the lowest start with the highest rise to finish with a respectable value, but lower than the other scales. This scale continuously remains in the third quadrant Q3="Low Priority" during S_1 and S_4 (Fig. 7).

E. Scale Correlation Analysis

Laugwitz *et al.* [23] reported that the UEQ scales were statistically independent of each other, apart for Attractiveness, thus assuming that conclusions related to Q4="Concentrate here" and Q1="Keep up the Good Work" will generate the highest impact. To confirm or to disconfirm that scales are indeed independent of each other, we computed Pearson's ρ correlation coefficient between the eight scales

TABLE IV. Inter-Correlations of the UEQ+ Scales: Pearson's ρ Coefficient

	Attractiveness	Efficiency	Perspicuity	Trust	Adaptability	Usefulness	Intuitive Use	Haptics
Attractiveness	–	0.48	0.38	0.13	0.47	0.27	0.38	0.20
Efficiency	0.93	–	0.34	0.14	0.22	0.07	0.52	0.04
Perspicuity	0.98	0.93	–	0.14	0.43	0.11	0.30	0.13
Trust	0.93	0.97	0.84	–	0.16	0.03	0.32	0.14
Adaptability	0.89	0.99	0.87	0.97	–	0.16	0.11	0.31
Usefulness	0.43	0.18	0.52	0.07	0.05	–	0.25	0.31
Intuitive Use	0.98	0.99	0.96	0.97	0.96	0.28	–	0.04
Haptics	0.97	0.84	0.90	0.90	0.79	0.38	0.91	–

Note: The upper-right half shows the correlation based on raw data ($N = 196$), the lower-left half those of the means across the four sessions ($N = 4$). For overall correlation, Pearson coefficients greater than $\rho = 0.10$ are significant, for correlations across means, coefficients greater than $\rho = 0.30$ are significant.

selected in our study, once across the whole sample (6 participants × 8 scales × 4 items = 196) and once based on the mean scores of the four sessions S_1 to S_4 ($N = 4$). The results are shown in Table IV. Following the guidelines recommended by Cohen [34], who proposed to interpret correlations of $\rho = 0.10$ as small, $\rho = 0.30$ as medium, and $\rho = 0.50$ as large, and consistently with Schankin *et al.* [35], we only interpreted correlations of $\rho > 0.30$ as being practically significant. For correlations across scale mean scores, correlations of $\rho > 0.30$ were statistically significant. As noted by Laugwitz *et al.* [23], Attractiveness is correlated with all other scales (all $\rho > 0.43$ in the lower-left part of Table IV. Although the scales are supposed to be independent of each other [23], we observed significant correlations between some of them: between Efficiency and Trust ($\rho = 0.97$), Adaptability, Intuitive Use ($\rho = 0.99$); between Perspicuity and Intuitive Use ($\rho = 0.96$); between Trust and Adaptability, Intuitive Use ($\rho =_{0.97}$). That is, scales measuring pragmatic aspects of EZExFix were correlated as well as those scales measuring non-pragmatic aspects.

IV. Conclusion

This paper presented and discussed the results of a multi-session evaluation of the EZExFix, a haptic device to be used by astronauts to fix a tibial shaft fracture in future lunar and Mars missions based on eight factors assessed by participants through corresponding items, scales, and importance ratings. The eight factors were continuously and uniformly evaluated through a multi-session, suggesting a hype cycle curve. The shape of this curve justifies the need for evaluating the scales over multiple sessions to reach a representative value and positioning in the quadrants. In the end, four of eight scales are located in the first ideal quadrant, while three are estimated to have low priority *i.e.*, Intuitive Use, Adaptability and Haptics. While Attractiveness is located in Q2 in the final session, it is so close to Q1 that we consider it encouraging.

These encouraging results should be moderated by the limitations of the study. Only 6 analogue astronauts were involved in the study because only one crew was evaluated. Having this kind of experiment is quite challenging because these facilities are not easily accessible to the general research community. They had three different backgrounds, which improves their diversity but also reduces their interrater consistency and reliability. Cautions should be taken in interpreting Usefulness, which might be somewhat biased in the sense that not all raters are equally knowledgeable in human physiology and fracture repair.

However, the high positivity of all scales allows us to be confident about the potential transposition to astronauts in real conditions. While orthopedic surgery is always based on objective learning curves, it is necessary to take into consideration the subjective learning curve, especially when it comes to putting a surgical device in the hands of astronauts without advanced medical training. The hype cycle curve attests the positive progress of the subjective perception.

Indeed, as the EZExFix is a device to treat injured astronauts in space conditions, to enhance survival and mission success, its Usefulness is considered of primary importance and meets the expectations. The Trust in the EZExFix remained constant during all sessions which is also very important in a surgical learning curve, applying the principle "*primum non nocere*" for both the patient and the operator. The best evolution along the different sessions was the Efficiency (Q4→Q1) which is task and/or goal-oriented and refers to the ability to do something with the minimum amount of time, effort, cost, or resources required to achieve the desired result. In other words, this is exactly what is sought to solve problems under extreme conditions, whether in space, in developing countries or in war medicine on Earth.

Also according to the astronauts themselves, a crew medical officer could be an essential member of a Mars mission [36] as the long-duration spaceflight and the harsh Martian environment pose various medical challenges that require expert knowledge, skills, and ability (KSAOs concept [37]) to manage. Astronauts would have more confidence in a physician with 4-6 years of clinical experience, which may increase the difficulty of selection [36]. However, our study showed that only four surgeries in a fortnight were enough to significantly upgrade the subjective learning curve of astronauts. By confronting it with the objective learning curve, the EZExFix could be a veritable tool to increase the autonomy and self-confidence of non-medical astronauts during long-duration exploration missions as well as the cost-effectiveness of meeting KSAOs requirements.

Finally, the multi-session assessment of the UEQ+ questionnaire could be of great interest for the evaluation of the subjective learning curve of surgical residents due to its ability to capture a broad range of emotional and cognitive responses during the learning process and provide valuable insights for improving surgical training programs on Earth.

Acknowledgment

The authors would like to deeply thank the Mars Society, Dr. Shannon Rupert and the Mission Support staff to welcome the crew in the MDRS (Utah desert) for the opportunity to execute this research simulation in a Mars analogue environment. We also want to thank Dr. Kouamé Jean-Eric Kouassi for his PhD thesis about the creation of the new EZExFix, Pr. Benoît Lengelé and Pr. Catherine Behets for their day-to-day support in the realization of our multiple projects, and Lies Fievé and Christine de Ville de Goyet for their help to create artificial broken legs.

This work was supported by a F.S.R. Fund («Fonds Spéciaux de Recherche», Belgium, Ref. ADi/16568.2021) and a «Student Angel Fund» (Ref. 304907648 val-31.05.), both granted by M.A.R.S. UCLouvain crew 2022, as well as a F.N.R.S. Aspirant Fund («Fonds National de la Recherche Scientifique», Belgium, Ref. ID 40004991) and a second F.S.R. Fund (Ref. ADi/IC/13936.2020), both granted by Dr. Julie Manon.

Detailed Histograms

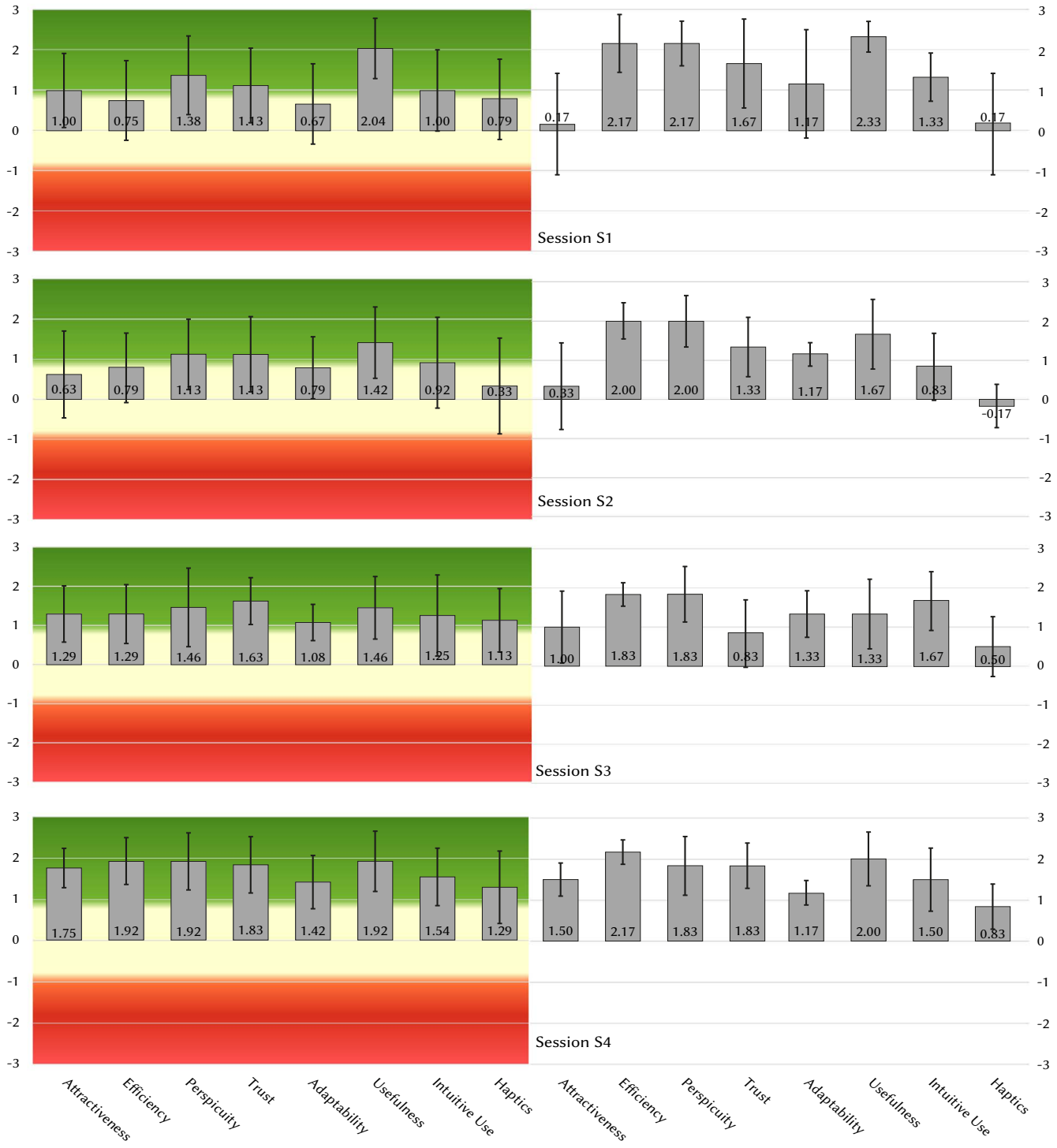


Fig. 8. Mean scores and scale importance of the four sessions. Error bars show a confidence interval of 95%.

References

[1] S. Brewster, "Haptic human-computer interaction," in *Proceedings of the 4th Annual Conference of the ACM Special Interest Group on Computer-Human Interaction*, CHINZ '03, New York, NY, USA, 2003, p. 3–4, Association for Computing Machinery.

[2] S. Villarreal-Narvaez, A. Sluÿters, J. Vanderdonckt, E. Mbaki Luzayisu, "Theoretically-defined vs. user- defined squeeze gestures," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 73–102, nov 2022, doi: 10.1145/3567805.

[3] V. Parthiban, P. Maes, Q. Sellier, A. Sluÿters, J. Vanderdonckt, "Gestural-vocal coordinated interaction on large displays," in *Companion Proceedings of the ACM Symposium on Engineering Interactive Computing Systems*, EICS '22 Companion, New York, NY, USA, 2022, p. 26–32, Association for Computing Machinery.

[4] Y. Fang, Y. Qiao, F. Zeng, K. Zhang, T. Zhao, "A human-in-the-loop haptic interaction with subjective evaluation," *Frontiers in Virtual Reality*, vol. 3, 2022, doi: https://doi.org/10.3389/frvir.2022.949324.

[5] P. Xia, A. M. Lopes, M. T. Restivo, "Virtual reality and haptics for product assembly," *International Journal of Online and Biomedical Engineering*, vol. 8, no. S1, pp. 12– 14, 2012, doi: 10.3991/ijoe.v8is1.1894.

[6] M. Haruna, M. Ogino, T. Koike-Akino, "Proposal and evaluation of visual haptics for manipulation of remote machine system," *Frontiers in Robotics and AI, Section Smart Sensor Networks and Autonomy*, vol. 7, 2020, doi: 10.3389/frobt.2020.529040.

[7] J. Manon, C. Detrembleur, S. Van de Veyver, K. Tribak, O. Cornu, D. Putineanu, "Predictors of mechanical complications after intramedullary nailing of tibial fractures," *Orthopaedics Traumatology: Surgery Research*, vol. 105, no. 3, pp. 523–527, 2019, doi: https://doi.org/10.1016/j.otsr.2019.01.015.

[8] J. Manon, C. Detrembleur, S. Van De Veyver, K. Tribak, O. Cornu, D. Putineanu, "Can infection be predicted after intramedullary nailing of tibial shaft fractures?," *Acta Orthopædica Belgica*, vol. 86, pp. 313–319, 2020.

[9] J. Manon, C. Detrembleur, S. Van de Veyver, K. Tribak, O. Cornu, D. Putineanu, "Quels sont les facteurs prédictifs d'une complication mécanique après enclouage centromédullaire d'une fracture diaphysaire du tibia?," *Revue de Chirurgie Orthopédique et Traumatologique*, vol. 105, no. 3, pp. 353–357, 2019, doi: 10.1016/j.rcot.2019.02.029.

[10] K. J.-E. Kouassi, J. Manon, L. Fonkoue, C. Detrembleur, O. Cornu, "Treatment of open tibia fractures in sub- saharan african countries: a systematic review," *Acta Orthopaedica Belgica*, vol. 87, no. 1, pp. 85–92, 2021, doi: 10.52628/87.1.11.

[11] A. Terhorst, J. A. Dowling, "Terrestrial analogue research to support human performance on mars: A review and bibliographic analysis," *Space: Science & Technology*, vol. 2022, 2022, doi: 10.34133/2022/9841785.

[12] J. Vanderdonckt, R. Vatavu, J. Manon, M. Saint- Guillain, P. Lefèvre, J. J. Márquez, "Might as well be on mars: Insights on the extraterrestrial applicability of interaction design frameworks from earth," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, 2024, pp. 239:1–239:8, ACM.

[13] C. Salisbury, R. Gillespie, H. Z. Tan, F. Barbagli, J. Salisbury, "What you can't feel won't hurt you: Evaluating haptic hardware using a haptic contrast sensitivity function," *IEEE Transactions on Haptics*, vol. 4, pp. 134–146, apr 2011, doi: 10.1109/TOH.2011.5.

[14] E. Samur, *Performance Metrics for Haptic Interfaces*. Springer Series on Touch and Haptic Systems, Springer, 2012.

[15] A. Hamam, A. E. Saddik, "Toward a mathematical model for quality of experience evaluation of haptic applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 12, pp. 3315–3322, 2013, doi: 10.1109/TIM.2013.2272859.

[16] R. Höver, M. D. Luca, M. Harders, "User-based evaluation of data-driven haptic rendering," *ACM Transactions on Applied Perception*, vol. 8, nov 2010, doi: 10.1145/1857893.1857900.

[17] E. Samur, "Systematic evaluation methodology and performance metrics for haptic interfaces," in *Proceedings of the IEEE World Haptics Conference*, WHC '11, 2011, pp. 1–1.

[18] A. Ahmad, K. Andersson, U. Sellgren, M. Boegli, "Evaluation of friction models for haptic devices," in *Proceedings of the Dynamic Systems and Control Conference*, vol. 2 of *Dynamic Systems and Control Conference*, 10 2013, p. V002T26A005.

[19] M. Saint-Guillain, J. Vanderdonckt, N. Burny, V. Pletser, T. Vaquero, S. Chien, A. Karl, J. Marquez, C. Wain, A. Comein, I. S. Casla, J. Jacobs, J. Meert, C. Chamart, S. Drouet, J. Manon, "Enabling astronaut self- scheduling using a robust advanced modelling and scheduling system: An assessment during a mars analogue mission," *Advances in Space Research*, vol. 72, no. 4, pp. 1378–1398, 2023, doi: https://doi.org/10.1016/j.asr.2023.03.045.

[20] J. Manon, V. Pletser, M. Saint-Guillain, J. Vanderdonckt, C. Wain, J. Jacobs, A. Comein, S. Drouet, J. Meert, I. J. Sanchez Casla, O. Cartiaux, O. Cornu, "An easy-to-use external fixator for all hostile environments, from space to war medicine: Is it meant for everyonersquo;s hands?," *Journal of Clinical Medicine*, vol. 12, no. 14, 2023, doi: 10.3390/jcm12144764.

[21] J. Manon, M. Saint-Guillain, V. Pletser, D. M. Buckland, L. Vico, W. Dobney, S. Baatout, C. Wain, J. Jacobs, A. Comein, S. Drouet, J. Meert, I. S. Casla, C. Chamart, J. Vanderdonckt, O. Cartiaux, O. Cornu, "Adequacy of in-mission training to treat tibial shaft fractures in mars analog testing," *Scientific Reports*, vol. 13, 2023, doi: https://doi.org/10.1038/s41598-023-43878-1.

[22] M. Schrepp, J. Thomaschewski, "Design and validation of a framework for the creation of user experience questionnaires," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 88–95, 2019, doi: 10.9781/IJIMAI.2019.06.006.

[23] B. Laugwitz, T. Held, M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, HCI and Usability for Education and Work, USAB 2008*, vol. 5298 of *Lecture Notes in Computer Science*, 2008, pp. 63–76, Springer.

[24] A. Hinderks, M. Schrepp, M. Rauschenberger, J. 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*, WEBIST 2023, 2023, pp. 319–329, SCITEPRESS.

[25] B. Boos, H. Brau, "Erweiterung des UEQ um die dimensionen akustik und haptik," in *Proceedings of Usability Professionals*, UP 2017, 2017, Gesellschaft für Informatik e.V. / German UPA e.V.

[26] S. Shelat, J. A. Karasinski, E. E. Flynn-Evans, J. J. Marquez, "Evaluation of user experience of self- scheduling software for astronauts: Defining a satisfaction baseline," in *Engineering Psychology and Cognitive Ergonomics*, Cham, 2022, pp. 433–445, Springer International Publishing.

[27] E. Schön, J. Hellmers, J. Thomaschewski, "Usability evaluation methods for special interest internet information services," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 6, pp. 26–32, 2014, doi: 10.9781/IJIMAI.2014.263.

[28] M. Schrepp, A. Hinderks, J. Thomaschewski, "Construction of a benchmark for the user experience questionnaire (UEQ)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, pp. 40–44, 2017, doi: 10.9781/IJIMAI.2017.445.

[29] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, p. 297–334, 1951, doi: https://doi.org/10.1007/BF02310555.

[30] S.-C. Liao, E. Hunt, W. Chen, "Comparison between inter-rater reliability and inter-rater agreement in performance assessment," *Annals of the Academy of Medicine, Singapore*, vol. 39, pp. 613–8, 08 2010, doi: 10.47102/annals-acadmedsg.V39N8p613.

[31] P. Legendre, "Species associations: the Kendall coefficient of concordance revisited," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 10, no. 226, 2005, doi: https://doi.org/10.1198/108571105X46642.

[32] A. Hinderks, M. Schrepp, F. J. D. Mayo, M. J. Escalona, J. Thomaschewski, "Developing a UX KPI based on the user experience questionnaire," *Computers Standards & Interfaces*, vol. 65, pp. 38–44, 2019, doi: 10.1016/j.csi.2019.01.007.

[33] A. Hinderks, A. -L. Meiners, F. Mayo, J. Thomaschewski, "Interpreting the results from the user experience questionnaire (ueq) using importance-performance analysis (ipa)," in *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, WEBIST 2019, Setubal, PRT, 2019, p. 388–395, SCITEPRESS, Science and Technology Publications, Lda.

[34] J. Cohen, *Statistical power analysis for the behavioral sciences*. New York, NY, USA: Routledge, 7 1988.

[35] A. Schankin, M. Budde, T. Riedel, M. Beigl, "Psychometric properties of the user experience questionnaire (ueq)," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022, Association for Computing Machinery.

[36] I. S. Saluja, D. R. Williams, D. Woodard, J. Kaczorowski, B. Douglas, P. J. Scarpa, J.- M. Comtois, "Survey of astronaut opinions on medical

crewmembers for a mission to mars," *Acta Astronautica*, vol. 63, no. 5, pp. 586–593, 2008, doi: https://doi.org/10.1016/j.actaastro.2008.05.002.

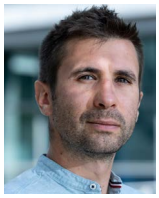
[37] L. B. Landon, C. Rokholt, K. J. Slack, Y. Pecena, "Selecting astronauts for long-duration exploration missions: Considerations for team performance and functioning," *REACH*, vol. 5, pp. 33–56, 2017, doi: https://doi.org/10.1016/j.reach.2017.03.002.

Julie Manon

She graduated physiotherapist and Medical Doctor from the Université catholique de Louvain (UCLouvain, Belgium) in 2013 and 2018, respectively. She holds a master's degree in Orthopedic and Traumatology surgery and completed various university certificates (taping, basic and advanced trauma life support, master of animal experiments, physical and biological radiation protection, microsurgery, and statistics/data science). She obtained a Ph.D. dedicated to bone reconstruction (2024). Her research interests include understanding of massive bone grafts healing in a critical-size bone defect from fundamental mechanisms to preclinical studies. She is also interested in the fracture risk of astronauts and fixation possibilities to promote healing in a hostile spatial environment. For this purpose, she enrolled as an analog astronaut (health and safety officer) in a Mars simulation mission conducted at the Mars Desert Research Station (UT, USA) in 2022.

Jean Vanderdonckt

He received a master's in mathematics, a master's degree in computer science, and a Ph.D. in Sciences from the University of Namur, Belgium, in 1987, 1989, and 1997, respectively. He is a Full Professor at Louvain School of Management, UCLouvain, Belgium. His research interests include information systems, human-computer interaction (HCI), engineering interactive computing systems (EICS), intelligent user interfaces (IUI). He is Associate Editor of ACM Trans. on Interactive Intelligent Systems (TiiS), co-editor-in-chief of the Springer Series of Human- Computer Interaction and the Springer Briefs in Human-Computer Interaction. He is ACM Distinguished Scientist and IFIP Fellow.



Michael Saint-Guillain

He received a master's degree in computer science from UClouvain in 2013 and a Ph.D. in engineering science (UCLouvain, Belgium) and computer science (INSA-Lyon, France) in 2019, while studying artificial intelligence and combinatorial optimization under uncertainty. At that time, his research interests included logistics, operations management, and decision under uncertainty, initially applied to space exploration. Since 2019, he is CEO of Rombio, a university spinoff project, helping biotechnology and pharmaceutical manufacturing companies optimize their production, decisions, and assets. Furthermore, side research interests and contributions now include planning and scheduling in space, human-computer interface, and optimization techniques applied to medical particle physics.

Vladimir Pletser

He (Ph.D., MSc, MEng) is the Director of Space Training Operations at Blue Abyss, specializing in astronaut training. Previously, he was a senior Physicist-Engineer at European Space Agency (ESA) (1985–2016), managing ISS microgravity payloads and parabolic flight programs, logging a Guinness world record of 7,350 parabolas. He has trained astronauts, participated in Mars simulation missions, and served as a Visiting Professor at 25 universities worldwide. With over 650 publications, including books and journal articles, he is a member of several prestigious astronautical and scientific organizations.

Cyril Wain

He holds a master's degree in electrical engineering from UCLouvain, with a specialization in cryptography and telecommunication systems, as well as a master's degree in management sciences from Solvay (Belgium). Initially serving as a crew astronomer, he later became the commander of the Tharsis mission. He is currently a Belgian national trainee at ESA.

Jean Jacobs

He obtained a master's degree in sciences, focusing on energy and environmental management (UCLouvain and Glasgow Caledonian University). He is a Ph.D. Candidate at the de Duve Institute (Belgium) and was the executive officer in the Tharsis mission.

Audrey Comein

She studied biological and biomedical sciences (Namur University, Belgium) and obtained her Ph.D. grade in 2025. She was enrolled twice for a manned mission (2020, 2021) and was the scientist in the Tharsis mission.

Sirga Drouet

She obtained a master's degree in biology (UCLouvain) and was the journalist in the Tharsis mission.

Julien Meert

He is a medical doctor at Cliniques Universitaires Saint-Luc, Université catholique de Louvain (UCLouvain, Belgium) where he is currently a "clinical assistant physician specialist candidate" (MACCS) in psychiatry and is interested in human psychological health and sleep. He played the role of the engineer in the Tharsis mission.

Ignacio Sanchez Casla

He holds a master's degree in mechanical engineering, from Ecole Polytechnique de Louvain (EPL), Université catholique de Louvain (UCLouvain, Belgium) and was enrolled as an astronomer in the Tharsis mission. He is currently a structural engineer at Societe Nationale de Construction Aerospatiale ("National Aerospace Construction Company").

Olivier Cartiaux

He holds a master's degree in electromechanical engineering (UCLouvain) with a specialization in mechatronics (2005). He further obtained a Ph.D. program focusing on computer and robotic assistance devices helping in orthopedic surgery (2010). After completing several Post-doctoral research, he is now the head of master's degree in health engineering (ECAM Brussels).

Olivier Cornu

He is head of the Orthopaedic and Trauma Surgery Department at the Cliniques Universitaires Saint-Luc UCL in Brussels and Professor of Anatomy and Physiology at UCLouvain. He has been deploying his expertise in the fields of musculoskeletal infections and tissue transplantation since 1996. His clinical practice is devoted to the management of infectious pathology of the musculoskeletal sector, to the reconstruction of large bone defects and revision joint replacement surgery. His research focuses on the study of the mechanical and biological properties of bone allografts, bone reconstruction, and implant-related infections, with an interest in treatments against bacterial biofilm. He also pursues numerous works oriented towards orthopedic and trauma care management in sub-Saharan Africa. He is an active member of several national and international scientific associations. He is also a member of the Royal Belgian Academy of Medicine and is Editor of the "Acta Orthopedica Belgica" Journal.