Published in final edited form as:

Acta Astronaut. 2022 January ; 190: 14–23. doi:10.1016/j.actaastro.2021.09.035.

Acceptability of the Cognition Test Battery in Astronaut and Astronaut-Surrogate Populations

K. Casario[1,*], K. Howard[1,*], M. Cordoza[1], E. Hermosillo[1], L. Ibrahim[1], O. Larson[1], J. Nasrini[1], M. Basner[1]

[1]Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Abstract

Background: Sustained high levels of astronaut cognitive performance are a prerequisite for mission success. A neuropsychological battery of 10 brief cognitive tests (Cognition) covering a range of cognitive domains was specifically developed for high performing astronauts to objectively assess cognitive performance. Extended mission durations require repeated cognitive testing and thus high acceptability of the Cognition software to the astronaut population. The aim of this qualitative study was to evaluate acceptability of Cognition to astronauts and astronaut surrogate populations.

Methods: Cognition was administered repeatedly to N=87 subjects (mean age ±SD 35.1 ±8.7 years, 52.8% male) on a laptop or iPad across five individual studies on the International Space Station or in space analog environments on Earth. Following completion of each study, participants were interviewed regarding their experience using Cognition in a semi-structured debrief. Participant comments were analyzed using a qualitative conventional content analysis approach.

Results: The majority of participants' comments (86.1%) were coded as positive or neutral in valence, with most positive comments relating to software usability, engagement, and overall design. Among the 10 Cognition tests, subjects liked the Visual Object Learning Test most (28 likes, 32.2% of participants), while the Emotion Recognition Test was liked least (44 dislikes, 50.6% of participants). Some subjects (36.8%) were frustrated with the level of difficulty of some of the 10 Cognition tests, especially during early administrations, which was by design to avoid ceiling effects in repeated administrations of high-performers. Technical difficulties were rare (20.7% of participants), and most often observed in environments with restricted internet access. Most participants (82.3% of those who commented) liked the feedback provided by Cognition after each test, which includes a graph showing performance history.

Conclusion: Cognition was found to be acceptable to astronaut and astronaut-surrogate populations across a variety of settings and mission durations. Participant feedback provided was used to further improve Cognition and increase its acceptability during sustained space missions.

Corresponding Author: Mathias Basner, MD, PhD, University of Pennsylvania basner@pennmedicine.upenn.edu.

[*]Contributed equally

Keywords

neuropsychological test; cognition; cognitive test; spaceflight; astronaut; performance

1. Introduction

Successful human space exploration depends on an astronaut's ability to maintain a high level of cognitive functioning in the presence of environmental and psychological stressors related to living in an isolated, confined and extreme (ICE) environment, where small errors can have catastrophic consequences [1]. To measure the impact of these stressors on astronaut cognitive performance, we developed a brief, comprehensive, and sensitive computerized cognitive test battery for spaceflight called Cognition [2]. Cognition consists of 10 brief cognitive tests that cover a range of cognitive domains, is administered electronically (via Windows laptop or Apple iPad), allows remote quality control via a web interface, and supports repeated administration (15 unique versions) [3]. The Cognition tests are based on tests with well-validated psychometric properties that were then modified to reflect the high aptitude and motivation of the astronaut population [2].

As space exploration expands, and mission durations increase, a computerized cognitive test battery geared towards high-performing astronaut populations not only needs to be sensitive and valid, but also simple to use, free of errors, engaging, and at the right level of difficulty to avoid both boredom and frustration. Therefore, a prerequisite for use of Cognition in spaceflight is the feasibility of administering the battery on a spacecraft, and acceptability of the software to astronauts. Cognition has been successfully administered on the ISS [4] and is part of NASA's standard measures, i.e., it will be administered to every astronaut who agrees to participate in the Standard Measures project going forward. It has therefore been shown to be feasible for self- administration in spaceflight. This paper therefore focuses on the acceptability of the Cognition software.

Here, we present data collected across 5 studies in N=87 astronauts or astronaut-surrogate subjects. These participants performed the Cognition battery repeatedly and participated in a semi-structured debrief after the end of each study to inform Cognition acceptability. We used a content analysis to extract data from the interviews. The primary objectives of this analysis were to (1) assess the acceptability of the Cognition battery to astronauts and astronaut-like participants and (2) identify aspects of the battery that could improve user acceptability and be incorporated in future iterations of the software.

2. Methods

2.1 Subjects and Protocol

Structured debriefs were performed in N=87 participants (Mean age ±SD 35.1 ±8.7 years, N=41 female; Table 1) participating in 5 different research protocols as described below and in Table 1. Only the Johnson Space Center (JSC) Astronaut (N=8) and International Space Station (ISS) (N=4) studies enrolled astronauts, while the other studies included astronaut surrogates (i.e., participants that share demographic, educational, and aptitude characteristics similar to astronauts).

The **JSC Astronauts** study took place at JSC in Texas. The purpose of this study was to investigate astronaut acceptability of Cognition, and to investigate practice effects related to repeated administration. Participants completed the battery in the comfort of their home. A total of 8 astronauts and astronaut candidates performed 15 bouts of Cognition over an average period of 6.5 months.

The **JSC Mission Control** study also took place at JSC. The purpose of this study was to investigate acceptability of Cognition in an astronaut-surrogate population of 11 mission controllers. Participants completed the battery at their workplace. The mission control personnel completed 15 bouts of Cognition each, over an average period of 6.5 months.

The **Human Exploration Research Analog** (HERA) study took place at the HERA facility, which is a two-story, four-port habitat unit residing in Building 220 at JSC [5]. The building is cylindrical in shape with a vertical axis, and connects to a simulated airlock and hygiene module. The study consisted of 47 astronaut surrogates spanning 3 campaigns that varied in duration: Campaigns 1 (1 week), 2 (2 weeks), and 3 (30 days). The days in HERA were operationally structured comparable to current operations on the ISS, which included research tasks, operational tasks, and emergency simulations. Each participant completed a total of 7, 16, and 18 bouts of Cognition per respective campaign.

The **Hawaii Space Exploration Analog & Simulation** (HI-SEAS) study was located at the HI- SEAS facility, an analog for spaceflight that mimics the isolated, confined, and extreme environments inherent to long-duration spaceflight missions. The HI-SEAS facility is located in a Mars-like environment on the slopes of Mauna Loa on the Big Island of Hawaii at approximately 8,200 feet above sea level. This research study investigated crew composition, cohesion, and performance in crews selected from a pool of astronaut-like candidates [6]. The study consisted of 17 participants spanning 3 campaigns that varied in duration: campaigns 1 (4 months), 2 (8 months), and 3 (12 months). Each participant completed 12, 24, and 33 bouts of Cognition per respective campaign.

The **International Space Station** study took place on the ISS. The study consisted of 4 astronauts on 6- or 12-month ISS missions. Each astronaut completed 4 bouts of Cognition pre- flight, 11 bouts in-flight, and 3 bouts post-flight.

All studies were approved by the Institutional Review Boards of Johnson Space Center and the University of Pennsylvania. Written informed consent was obtained from all participants prior to participation.

Following each mission, semi-structured face-to-face (or videoconference) debriefs were conducted with each study participant individually by members of the research team familiar with the protocol. Responses were elicited by direct questions in the following categories: (1) location where Cognition was administered, (2) distractions encountered while completing the battery, (3) experience with each of the 10 individual tests, (4) overall Cognition experience, (5) feedback provided by Cognition after each test and suggested improvements, (6) helpfulness of Cognition in spaceflight, (7) Cognition administration frequency, and (8) open-ended question about any additional comments (full interview questions available in supplement). The debrief questions were used as a guideline by the interviewer. Although all participants were interviewed, not all participants received all questions.

2.2 Cognition Test Battery

The Cognition test battery was administered either on laptops calibrated for timing accuracy (JSC Astronauts, JSC Mission Control, and ISS studies) or on a 4th generation Apple iPad (HERA and HI-SEAS studies). The first administration served as familiarization and required participants to perform brief practice versions of 8 of the 10 tests in addition to performing the actual test. After the first administration, participants performed all 10 Cognition tests, which typically took less than 20 minutes. The 10 Cognition tests are described in detail below (modified from [3]) and were always performed in the order listed.

The **Motor Praxis Task** (MP) [7] was administered at the start of testing to ensure that participants had sufficient command of the computer interface, and immediately thereafter as a measure of sensorimotor speed. Participants were instructed to click on squares that appeared in random locations on the screen, with each successive square smaller and thus more difficult to track. Performance was assessed by the speed with which participants click each square.

The **Visual Object Learning Test** (VOLT) [8] assessed participant memory for complex figures. Participants were asked to memorize 10 sequentially displayed three-dimensional shapes. Later, they were instructed to select the objects they memorized from a set of 20 such objects, also sequentially presented, half of which were from the learning set and half new.

Subjects were given four response options to indicate whether they believed they had seen the presented object before: (1) ‘Definitely yes’, (2) ‘Probably yes’, (3) ‘Probably no’, (4) ‘Definitely no’.

The **Fractal 2-Back** (F2B) [9] is a nonverbal variant of the Letter 2-Back. N-back tasks have become standard probes of the working memory system and activate canonical working memory brain areas. The F2B consisted of the sequential presentation of a set of figures (fractals), each potentially repeated multiple times. Participants were instructed to respond when the current stimulus matched the stimulus displayed two figures ago. The current implementation used 62 consecutive stimuli.

The **Abstract Matching** (AM) test [10] is a validated measure of the abstraction and flexibility components of executive function, including an ability to discern general rules from specific instances. The test paradigm presented subjects with two pairs of objects at the bottom left and right of the screen, varied on perceptual dimensions (e.g., color and shape). Subjects were presented with a target object in the upper middle of the screen that they had to classify as belonging more with one of the two pairs, based on a set of implicit, abstract rules. The current implementation used 30 consecutive stimuli.

The **Line Orientation Test** (LOT) [11] is a measure of spatial orientation and derived from the well-validated Judgment of Line Orientation Test. The LOT format consisted of presenting two lines at a time, one stationary while the other could be rotated by clicking an arrow. Participants could rotate the movable line until they perceived it to be parallel to the stationary line. The current implementation used 12 consecutive line pairs that varied in length and orientation.

The **Emotion Recognition Task** (ERT) [12] is a measure of facial emotion recognition. It presented subjects with photographs of professional actors (adults of varying age and ethnicity) portraying emotional facial expressions of varying types and intensities (biased toward lower intensities, and with the prevalence of emotion categories balanced within each version of the test). Subjects were given a set of emotion labels ("happy"; "sad"; "angry"; "fearful"; and "no emotion") and had to select the label that correctly described the expressed emotion. The current implementation used 40 consecutive stimuli, with 8 stimuli each representing one of the above 5 emotion categories.

The **Matrix Reasoning Test** (MRT) [7] is a measure of abstract reasoning and consists of increasingly difficult pattern matching tasks. It is analogous to Raven's Progressive Matrices [13] an established measure of general intelligence. The test consisted of a series of patterns overlaid on a grid. One element from the grid was missing and the participant was asked to select the element that best fit the pattern from a set of several options. The current implementation used 12 consecutive stimuli. The difficulty level of this task gradually increases over the course of the 12 stimuli.

The **Digit-Symbol Substitution Task** (DSST) [14] is a measure of complex scanning, visual tracking, and working memory. DSST is a computerized adaptation of a paradigm used in the Wechsler Adult Intelligence Scale (WAIS-III). The DSST required the participant to refer to a displayed legend relating each of the digits one through nine to specific symbols. One of the nine symbols appeared on the screen and the participants were asked to type the corresponding number using the keyboard as quickly as possible. The test duration was fixed at 90 s, and the legend key was randomly reassigned with each administration.

The **Balloon Analog Risk Test** (BART) is a validated assessment of risk taking behavior [15]. The BART required participants to either inflate an animated balloon or stop inflating and collect a reward. Participants were rewarded in proportion to the final size of each balloon, but a balloon popped after a hidden number of pumps, which changed across stimuli [15]. The current implementation used 30 consecutive stimuli. The average tendency of balloons to pop was systematically varied between test administrations. This required subjects to adjust the level of risk based on the behavior of the balloons.

The **Psychomotor Vigilance Test** (PVT) is a validated measure of vigilant attention based on reaction time (RT) to visual stimuli that occur at random inter-stimulus intervals [18, 19]. Subjects were instructed to monitor a box on the screen and press the spacebar once a millisecond counter appeared in the box and started incrementing. The reaction time was then displayed for one second. Subjects were instructed to be as fast as possible without hitting the spacebar in the absence of a stimulus (i.e., false starts or errors of commission).

Cognition uses a validated 3-min. PVT (PVT-B) with 2–5 s inter-stimulus intervals and a 355 millisecond lapse threshold [16].

After each test, a feedback screen was presented that displayed a score between 0 (worst possible performance) and 1,000 (best possible performance), as well as the history of feedback scores of past administrations if Cognition had been performed more than once. At the end of the battery, a summary page displayed the feedback scores of all 10 tests and a summary score across all 10 tests (now ranging between 0 and 10,000).

2.3 Qualitative Data Analysis

Participant feedback regarding aspects of their experience using Cognition from the structured debriefs was analyzed using a qualitative conventional content analysis approach [19, 20]. A conventional approach was used to develop codes and themes directly from the content of the debrief comments [20]. Data were divided into responses for each debrief question category, creating 16 datasets (6 debrief questions plus the 10 individual Cognition tests). Groups of two reviewers were assigned to each of the 16 datasets. Given that these reviewers were not the same researchers who conducted the interviews, reviewers had little previous background on participants’ experience with Cognition. Individually, each reviewer familiarized themselves with the data by reading each comment and generating initial codes. Codes were identified from recurring words, phrases, or units of meaning. From these codes, possible themes were extracted. Each reviewer also coded the valence of each comment as either positive, neutral/mixed, or negative. Each comment could have as many codes and fit into as many subsequent themes as appropriate; however, only a single valence was assigned to each comment.

After independent coding, each reviewer dyad appraised their analysis together and discussed responses. Consensus was reached on common codes and comment valence, and any discrepancies were discussed within the group of all authors through regular team meetings. As a part of this process, codes were amalgamated or new codes were created. Finally, for every dataset, an agreed upon set of themes was identified. For each debrief question, the number of comments and frequency of codes were reported. The frequency of codes that fell under a specific theme were calculated as an indication of the prominence of that specific theme.

3. Results

In the following sections, we will present results from the content analysis for each of the topics addressed by the debrief questionnaire.

3.1 Location & Distractions

Participants were asked about the location in which they typically completed Cognition to better contextualize subsequent debrief feedback. In total, 37 participants (42.5%) were asked, “Where would you usually complete the cognition test bouts?” The most common locations were in the crew common area (n=16, 43.2% of those who commented), followed by their sleeping quarters (n=12, 32.4% of those who commented), and/or in varying locations (n=6, 16.2% of those who commented). Of those who commented, a higher percentage (35.1%) said that they sought to complete the test with others in a social setting, while 13.5% explicitly stated they found it better to take the test alone in an isolated setting. Only HERA participants were asked to take Cognition together at the same time in the common area of the facility; otherwise, participants could take Cognition in the place of their choosing.

Out of the 87 participants, 78.1% (n=68) were asked, "Did you find that you were distracted while completing the cognition measures. If so, how did you compensate for the distractions?" Only 14.7% of those who commented reported feeling regularly distracted, 10.3% reported varying levels of distraction, and the majority (73.5%) reported not feeling distracted. While some participants mentioned environmental noise due to the facilities where the studies took place, none of them attributed it to causing distractions when completing Cognition. Of those who self-reported feeling distracted, participants attributed the distractions to movement of others (34.4%), chatter (18.8%), and the first person to finish within a group setting (12.5%). To combat these distractions, participants reported that they spoke with others about being quiet (21.9%), changed locations (21.9%), and/or wore earplugs (25%).

3.2 Overall Experience

In total, 85 participants (97.8% of the total number of participants) gave feedback regarding their overall experience with Cognition. The most frequent theme was *software acceptability*.

Participants commented that the test battery had a well-constructed interface/software (n=33, 38.8% of those who commented), stating, "yours was pretty much the most professional, smooth working test we had," and "test is very user-friendly compared to other tests and well-polished." An equal number of comments described some of the Cognition tests as either interesting, fun, or enjoyable, or as confusing, frustrating, or discouraging (n=32, 37.6% for each category) with statements such as, "some of the games in the battery were actually quite enjoyable", "Cognition was a pleasure to do during the mission, it was one of the best things they had onboard", "there were some [tests] that were a bit weird and a bit frustrating", and "frustrating because it's difficult to complete a few of the tests without memorizing". Fifteen participants (17.6% of those who commented) reported Cognition was easy to use. An equal number of participants commented on looking forward to Cognition or noting that it was their favorite in some regard, and that they felt Cognition was game- and competition-like (n=11, 12.9% for each category).

Comments that described looking forward to Cognition included sentiments such as, "Overall enjoyed Cognition, looked forward to a break in the day," and "Cognition was my favorite part of HERA." Game/competition-like comments included, "Felt competitive during Cognition and wanted to get good scores," and "felt like fun games at the end of the day."

The second most common theme was *technical issues*. A total of 18 participants (21.2% of those who commented) reported connectivity and/or software issues, which were mostly attributed to either HERA (n=13, 72.2% of those who mentioned technical issues) or

HISEAS (n=4, 22.2% of those who mentioned technical issues). For both of those studies, Cognition was performed on an iPad that had restricted Internet access. Participants reported, “Every once in a while the test would freeze up and I would have to log out and log back in”, and “Cognition crashed momentarily a couple of times throughout the study. Exit and open again would solve the issue.”

The third theme identified was *study design*. There were comments related to the quantity, timing, or length of Cognition (n=10, 11.8% of those who commented). Specifically, some participants (n=4, 40% of those who mentioned study design) found it difficult to complete Cognition before their bedtime, “I didn’t like that I always had to do it right before bed when I was dead tired so it was the last thing I had to do. I found it stimulating and therefore difficult to get to sleep afterwards.” Some also thought the individual tests were tedious or would drag on, “Sometimes the tests seemed like they would drag on forever, especially the PVT and F2B,” and “fun at first, then not as fun towards the end as it got tedious.”

3.3 Individual Cognition Tests

When asked “Have you had any history of completing neurocognitive measures?”, 14.9% of the total number of participants claimed they had previous experience in one form or another prior to cognitive testing, 41.4% of the total number of participants claimed that they did not have previous experience, and the remaining percentage of participants did not answer the question.

The ratio of positive to negative comments per test were accordant with net preference of favorite and least favorite tests. Tests containing more positive comments received a higher net test preference score, whereas tests containing more negative comments typically received a lower net test preference score (Table 2).

Participants were asked “Please tell us which test you liked doing the most and why” and “Please tell us which test you liked doing the least and why.” The tests more liked than disliked were MP, VOLT, AM, LOT, MRT, and DSST, with VOLT being liked most. In contrast, F2B, ERT, BART, and PVT were more disliked than liked, with ERT being most disliked (Table 2).

Motor Praxis Task (MP)—The most common theme was *test execution*. Participants found it difficult to use the trackpad when completing this test (n=6, 31.6% of those who commented). Participants mentioned that they were “not used to the trackpad”, or that their score might be low because they “accidently used the mouse and had to switch to the trackpad.” All comments for this theme were marked as neutral or mixed valence. Another theme was *test ease*. A total of five participants (26.3% of those who commented) found the MP to be “easy” or “simple”; of these five, four were marked as positive and one was marked as neutral/mixed.

Visual Object Learning Test (VOLT)—One of the most common themes found for the VOLT was regarding *test strategy*. Participants assigned “words,” “names,” or “phrases” to shapes as a memory tool, which was communicated by seven participants (28% of those who commented). Additionally, participants mentioned using a strategy when choosing "probably" vs. "definitely" as responses (n=4, 16% of those who commented) to certainty of stimulus classification. Seven participants (28% of those who commented) mentioned that they were not confused by prior administrations of the VOLT (e.g., "Enjoyed the test. Used a word to help remember the shapes. Was not getting confused by the previous week's symbols"). Another theme addressed *personal progress*. Three comments (12% of those who commented) mentioned improving on the VOLT over time.

Fractal 2-Back (F2B)—The primary theme was *test difficulty*. This theme was identified by 17 participants (44.8% of those who commented), with comments such as, "very difficult," and "I struggled most with F2B, the fractals. I just couldn't seem to, in many cases, try to quantify what the fractal was."

Use of a *test strategy* was the next frequent theme reported (n=15, 39.5% of those who commented). Most participants described the technique of trying to come up with words to remember images (n=10, 66.7% of those who mentioned using a test strategy), or using image colors or shapes to improve recall (n=4, 26.7% of those who mentioned using a test strategy). The final theme was *test delivery* (n=5, 13.2% of those who commented). Participants either reported that the images were displayed too quickly (n=3, 60% of those who mentioned test delivery) or wanted more feedback about performance (n=2, 40% of those who mentioned test delivery).

Abstract Matching (AM)—The most common theme was *test conceptualization*. Over half of the participants experienced difficulty figuring out the test's rules (n=16, 53.3% of those who commented), stating, "It's not that I didn't like the AM, but I never mastered it. I enjoyed it but I never came up with the algorithm. It's not disliked, it was a challenge." Some participants mentioned figuring out the test's algorithm to some degree by using a *test strategy* (n=10, 33.3% of those who commented). Participants mentioned trying various strategies of focusing on shapes, colors, and even used poker analogies on how they completed this test. A third theme was *personal improvement*.

Participants commented that they began to see improvement at the very end (n=4, 13.3% of those who commented), while others improved after a few attempts (n=4, 13.3%).

Line Orientation Test (LOT)—The most common theme for the LOT was *test scoring*. Participants could not figure out how the feedback scores were generated (n=6, 25% of those who commented). This was commonly referred to by participants as "the scoring algorithm". Of the individuals who commented on the scoring algorithm, most noted that they were unsure whether their speed (based on trial completion times; n=4, 66.7% of those who mentioned test scoring) or their accuracy (based on the final orientation of the two lines; n=1, 16.7% of those who mentioned test scoring) was preferred by the feedback scoring algorithm. The second identified theme was *test delivery*.

Participants suggested improvements on the LOT's input method (n=4, 16.7% of those who commented; e.g., "would have preferred to be able to click and hold rather than keep making individual clicks"). Lastly, *test ease* was the third theme. Participants mentioned the test was "simple" or "straightforward" (n=3, 12.5% of those who commented).

Emotion Recognition Task (ERT)—The most common theme found for the ERT was regarding *test difficulty*. Over half of the participants who commented found it hard to discern the emotions of the stimuli (n=29, 65.9% of those who commented). *Test conceptualization* was the second most common theme. Participants mentioned not being sure of the test's goal and/or what information could be extracted from their performance on this specific test (n=7, 15.9% of those who commented) e.g., "Kept trying to figure out the purpose of the test. Was challenging. Thought the test was accessing the subject's mood based on response to ambiguous stimuli.".

Matrix Reasoning Test (MRT)—*Test difficulty* was the most common theme. Participants commented that the test was "hard" or "difficult," (n=7, 26.9% of those who commented). Of the 7 participants that mentioned the difficulty of the test, 4 were coded as neutral in valence (e.g., "Matrix Reasoning seemed to get easier part-way through but then got more difficult again."), suggesting that the difficulty of the test was not always perceived as negative. However, two participants specifically mentioned that the difficulty of the test was "frustrating," and both of these comments were coded as negative in valence. Participants also mentioned that they had to guess as items became more difficult (n=4, 15.4% of those who commented). The second most common theme was *test delivery*. Either participants "timed out" of the trials at least once (n=7, 26.9% of those who commented) or they wanted more time to respond during each trial (n=2, 7.7% of those who commented).

Participants were given 30 seconds to select the appropriate missing element on each trial. The third theme was *test scoring*. Participants mentioned that they were not sure if the algorithm to generate feedback scores valued speed over accuracy or vice versa (n=3, 11.5% of those who commented), Finally, *personal performance* was the fourth theme; 11.5% of participants who commented mentioned not feeling good performers at this particular test.

Digit-Symbol Substitution Task (DSST)—*Test strategy* was the only theme for this test. Participants mentioned implementing a finger placement strategy in which they placed all eight fingers (excluding thumbs) on the keyboard (n=7, 26.9% of those who commented). This resulted in the pinky fingers resting on the numbers one and nine keys. Others assumed speed was important in obtaining a good score (n=3, 11.5% of those who commented). Additionally, participants reported that they conceptualized the test as a memory test and attempted to memorize the stimuli in order to obtain a better score (n=7, 26.9% of those who commented).

Balloon Analog Risk Test (BART)—One of the most common themes was regarding *test scoring*. Participants mentioned their inability to identify the algorithm behind the feedback scores (n=15, 32.6% of those who commented). Participants described feeling "confused" or "frustrated" due to the test's feedback scoring (n= 14, 30.4% of those who commented), reporting "The scoring system was confusing and frustrating. Same strategy would lead to vastly different scores." Consequently, these participants reported using little to no strategy and inflated each balloon a random number of times. *Test strategy* was the second theme. Of those who described utilizing a strategy, participants mentioned pumping a set number of times (n=9, 19.6% of those who commented), determining an average number of pumps it would take to pop a balloon across tests (n=5, 10.9% of those who commented), and assuming a random distribution for the rate of balloons popping (n=5, 10.9% of those who commented).

Psychomotor Vigilance Test (PVT)—The most common theme was *test preference*. Participants commented that the test felt long and monotonous (n=18, 42.9% of those who commented). *Test focus* was the second theme.

Participants mentioned there was a great level of focus and attention required for this test (n=10, 23.8% of those who commented), stating, “I didn’t want to blink”. *Test delivery* was the last theme, with some participants experiencing software issues such as “freezing” (n=8, 19.1% of those who commented).

3.4 Overarching Themes

From the structured debriefs, several overarching themes emerged. The first overarching theme was *test strategy* (35% of those who commented) and involved participants’ commenting on use of a strategy to improve their performance and earn an optimal score on a test. This was reported by 28% (n=7), 39.5% (n=15), 33.3% (n=10), 26.9% (n=7), 41.4% (n=19) of participants who commented on VOLT, F2B, AM, DSST, and BART tests, respectively. The second theme regraded difficulty in *test conceptualization/understanding test scoring* (27.6% of those who commented) relative to a test’s algorithm and/or feedback scoring. This theme was reported by 53.3% (n=16), 25% (n=6), 15.9% (n=7), 11.5% (n=3), and 32.6% (n=15) of participants who commented on AM, LOT, ERT, MRT, and BART tests, respectively. The third theme, *test difficulty* (20.7% of those who commented), regarded participants’ comments of finding individual tests challenging. This was reported by 44.7% (n=17), 13.3% (n=4), 9.1% (n=4), 26.9% (n=7), and 13% (n=6) of participants who commented on the F2B, AM, ERT, MRT, and BART tests, respectively. Similarly, 26.7% (n=8), 9% (n=4), 30.4% (n=14), and 21.4% (n=9) of participants who commented on the AM, ERT, BART, and PVT tests respectively, found the aforementioned tests *frustrating* (21.6% of those who commented), resulting in a fourth overarching theme. A final overarching theme was *test enjoyment* (18.9% of those who commented). Participants found the individual tests enjoyable and fun, this was reported by 15.8% (n=3), 20.8% (n=5), 19.2% (n=5), and 19.2% (n=5) of participants who commented on MP, LOT, MRT, and DSST tests, respectively.

3.5 Improvements to Battery or Study

Participants were asked to suggest improvements to Cognition (“How could Cognition overall be improved?”) and to the overall study (“Do you have any suggestions to improve this study?”). Of the 87 participants, 33 (37.9% of the total number of participants) responded to either or both of these questions, but only 30 comments (34.5% of the total number of participants) actually specified a suggested improvement.

A proportion of participants who commented (n=14, 46.7% of those who suggested improvements) suggested making changes to the software or hardware. Half of the software/hardware comments (n=7) were in regards to the iPad. For example, two participants noted that the reflectiveness of the iPad was distracting, particularly during the PVT. Using a matte finish for iPads to reduce glare and to allow the ability to adjust the brightness was suggested. Another participant mentioned the login process was “confusing” and suggested that usernames and passwords be standardized.

Several comments (n=6, 20% of those who suggested improvements) involved suggestions regarding test instructions, training, and practice sessions. Specifically, one participant noted that they had to press the “Next” button too many times to start each session and suggested a button to skip instructions. Additionally, another participant suggested practice sessions should be more similar to the actual tests, (e.g., “Giving participants some idea what is expected in each test would make it better. Was not aware what the test was asking for a while. Future crews would benefit from more thorough training with the battery”).

The frequency Cognition was administered varied across all five studies, ranging from daily to bi- or tri-weekly. When asked “What rate of administration would be perfect for you?” a total of 83 participants commented (95.4% of the total number of participants). Most preferred to have the test administered 1–3 times per week (n=46, 55.4% of those who commented). Once every other week was the second most preferred administration frequency (n=23, 27.7% of those who commented).

3.6 Feedback & Improvements to Feedback

In total, 79 participants (90.8% of the total number of participants) were asked “How helpful did you find the feedback for the tests?” and “What could be improved about the feedback?” The majority of comments (n=43, 54.4% of those who commented) were scored as having a neutral/mixed valence, 33 (41.8% of those who commented) as positive and the remaining three comments (3.8% of those who commented) were scored as negative. A total of 65 participants (82.3% of those who commented) reported feedback to be helpful overall. Participants reported that the feedback influenced motivation and elicited a competitive performance in improving their score (n=17, 21.5% of those who commented) stating, “Feedback helped the crew remain engaged in the study and motivated.” Conversely, 9 participants (11.4% of those who commented) reported feedback as discouraging, “Good to know the population norms, but may be discouraging, option to display to astronauts may need to be a choice.”

A common theme found among comments was a request for more in-depth feedback (n=24, 30.4% of those who commented). Cognition only displays a single value that is based on both accuracy and speed, not raw scores such as reaction time. Specifically, participants requested more detailed feedback plotting historical data of the Cognition scores combined across tests, color coding scores, and various graphs (n=10, 12.7% of those who commented). Others expressed interest in receiving normative data compared to other astronauts (n=9, 11.4% of those who commented). Participants also requested to learn the percent of responses they answered correctly, both in real time and in the summary at the end of each test (n=4, 5.1% of those who commented) stating, “Would prefer to be told how many I was getting correctly.”

3.7 Helpful Tool

The question "Do you think Cognition could be a helpful tool for spaceflight operations? If yes, why? If not, why not?" was posed to 25 participants (28.7% of the total number of participants). The majority of comments (n=22, 88% of those who commented) were coded as having a neutral valence, and the rest of the comments (n=3, 12% of those who commented) were coded as positive. There were no negative comments. Comments were coded as positive for participants who used words/phrases like "absolutely," "enjoyable," and "fun." Twenty-four participants (96% of those who commented) said they did think Cognition would be a helpful tool for spaceflight operations. Over half of participants (n=14, 56% of those who commented) found Cognition to be a direct way to measure cognitive ability and mental health at the moment of administration. For example, one participant stated, "The test was useful to judge one's own level of alertness and say, 'Maybe I'm not at my best'. Also good for judging one's ability to perform." Participants also mentioned that Cognition provides the ability to see patterns in participant performance over time (n=5, 20% of those who commented).

3.8 Other Comments:

Lastly, 31 participants (35.6% of the total number of participants) were asked if they had any other comments regarding Cognition. The most frequent theme among all "other comments" was *software enjoyment* (n=11, 35.5% of those who commented). Other themes regarded *technical delivery* or *user interface* of Cognition (n=8, 25.8% of those who commented). For example, "the iPad interface was very nice," and "it was a really solid software", or that "there was one time it froze up in the battery so I quit and it started right back up where it was when it froze, so it kept tabs on where you were." Additional common themes regarded *timing* (n=6, 19.4%) or *length/frequency* of Cognition (n=7, 22.6% of those who commented). Timing comments included "would have preferred not to do it daily", "it was hard to do before bed", and that "it became part of sleep hygiene or a ritual that prepared one to go to sleep". The majority of these comments were neutral in valence (n=27, 87.1% of those who commented). Three responses were coded as positive, and 1 comment mentioning significant sleeping difficulties exacerbated by Cognition was coded as negative.

4. Discussion

The primary objectives of this analysis were to assess the acceptability of the Cognition test battery to astronauts and astronaut-like participants, and to identify aspects of the battery that could be improved in future iterations of the software. A total of 87 participants performed Cognition repeatedly in five different research settings, including astronauts on the ISS; astronauts, astronaut candidates and mission controllers in a ground study at JSC; and astronaut- like participants that were confined for durations between 1 week and 1 year in two space analog environments (HERA and HI-SEAS). Following study completion, participants were debriefed individually about several aspects of repeatedly performing Cognition in the research setting.

Overwhelmingly, participants across all 5 research settings found Cognition to be acceptable, with an overall positive or neutral valence in 86.1% of all comments. In evaluating comments made by astronauts compared to astronaut surrogates, there were no relevant differences in acceptability.

On long-duration space missions, astronauts will have to perform cognitive testing repeatedly to continuously evaluate neurobehavioral performance. It is therefore important that the cognitive test software not only works flawlessly but is also engaging. In this study, the feedback relative to software design and usability was overwhelmingly positive. More than one third of participants (37.9% of the total) said the software was well constructed, while 36.8% of participants found Cognition to be fun, interesting, or entertaining; suggesting that the Cognition software was acceptable to the target population, especially in the setting of repeated administrations.

Cognitive tests should be geared towards the population of interest. Tests that are too easy are boring, and performance easily hits a ceiling, while tests that are too hard can be frustrating. Cognition was specifically designed for high-performing astronauts, so it was expected that some participants were frustrated with the difficulty level of some of the tests, especially during early administrations. Several negative valence comments reflected the level of test difficulty related to scoring, not being able to figure out algorithms behind tests, and simply being challenged by the tests' stimuli. During familiarization, we distinctly emphasize that some tests are difficult by design, and that participants would get better with repeated administration.

Technical issues that did arise were primarily found during HERA and HI-SEAS campaigns where participants completed Cognition on Apple iPads. Restricted Internet access in these two analogs sometimes prevented the software from connecting to the Cognition server and caused spontaneous test abortions or prevented a test from starting. These connectivity issues, which are prevalent in spaceflight and space-analog environments (e.g., Antarctic research stations), were incorporated into a design change for version 3 of the Cognition laptop software. Version 3 can be operated in offline mode, where the software does not exchange data with the central server. A database file can be downloaded from the server and used to install or update the client software on a laptop without Internet connection. The data can be exported locally and then later uploaded to the server database. Operating Cognition in offline mode can also be beneficial in research settings where the transfer of data to a central server can be problematic from a data privacy perspective. The connectivity issues on the iPad have also been fixed.

Whether or not to present performance feedback to participants, and in what form, was a fiercely discussed topic in the development of Cognition. Feedback can affect participant motivation and, in repeated administration settings, future performance [21]. However, there was agreement that not presenting any kind of feedback would probably dampen participant motivation considerably. Also, the decision was made to provide a history of past performance scores together with the current score. Crew autonomy will be much higher on exploration-class missions, and a comparison to past performance (i.e., self-assessment) can quickly verify relevant performance deficits in the presence of communication delays. In addition, comparisons to past performance can further boost astronaut motivation. Comments received by the participants are largely in line with these design decisions.

The majority of participants who were asked the question regarding feedback reported that they liked receiving feedback on their performance (82.3% of those who commented), some even requested more detailed feedback (30.4% of those who commented), and 21.5% of participants who commented claimed that the feedback was useful in driving them to perform better.

The feedback scores implemented across the 5 studies were generally based on both accuracy and speed, but they were informed by data from only a handful of participants that participated in early trials of Cognition. Some of these feedback scores did not perform well, which may explain some of the negative comments expressed, (e.g., “I hated BART, it’s not transparent how you score the test,” and “ERT was frustrating. I thought my performance was good but was discouraged by the scores”). For version 3 of Cognition, the feedback scores were overhauled.

They are still based on accuracy and speed, but are now informed by more than 3,500 administrations of 588 participants. Also, they were generated in a way that they can now be interpreted as percentile ranks relative to this norm population (details on the generation of these feedback scores and the make-up of the norm population will be reported elsewhere). More detailed feedback on accuracy and speed can be counter-productive, as participants may start tweaking their response strategy from trial to trial. Practice versions sometimes included feedback to individual stimuli while the actual version did not, which may have prompted some participants to ask for more detailed feedback on the stimulus level. Finally, an option to turn feedback on and off would be desirable for future versions of Cognition, but it would also complicate comparing results across studies that used or did not use the feedback option.

While most Cognition tests were liked by some participants and disliked by others, the ERT stood out as the most disliked test (44 participants reported it as one of their least favorite tests, Table 2). Two reasons likely accounted for the ERT’s low popularity. First, an Item Response Theory (IRT) analysis of individual stimuli showed that many of the stimuli did not work well in the sense that they were not able to differentiate high from low performers. Second, the feedback score did not reflect actual emotion recognition performance well, partly due to the “bad” stimuli discussed above. In addition to changing the feedback score (see above), we developed a revised version of the ERT for version 3 of Cognition that only uses 20 instead of 40 stimuli, but 20 stimuli with favorable properties as established via IRT (details on the revised version of the ERT will be reported elsewhere).

4.1 Strengths and limitations

This study had several strengths and limitations. Strengths were that a relatively large number of astronauts and astronaut-like participants were investigated in a range of spaceflight-relevant environments, and that reliability of data collection was high due to backups both on the local client and on a web server. We also systematically evaluated participant feedback using a valid qualitative approach, and analytic rigor was addressed in several ways. Each reviewer engaged with assigned datasets for extended periods of time [22], reading each comment by each participant several times individually and then with a second reviewer. Specific codes, potential themes, and comment valence were first independently appraised by two reviewers prior to deciding common themes for each data set. Confirmability of results was established via detailed discussion among each reviewer dyad, and through consistent team meetings in which discrepancies within the reviewer dyad were further discussed to reach group consensus.

Limitations include that while debriefs were based on a structured questionnaire, this was not always strictly followed, thus not every participant was asked the exact same questions. This restricts interpretability of the number of participants who made comments regarding a specific question. To address this limitation, researchers calculated the percentages from the total number of people who answered each question, not out of the total sample size. Also, we did not have full transcripts of the debrief for analysis. Participant comments were often summarized and not always recorded verbatim by the interviewer. As such, we may not have captured the complete nature of the feedback. However, much of the feedback was consistent across studies. Thus we likely captured the most significant themes related to the acceptability of Cognition.

5. Conclusions

The Cognition test battery for spaceflight was highly acceptable to astronauts and astronaut-like participants across a range of settings and mission durations, with overwhelmingly positive comments in terms of software usability, engagement, and other design choices. At the same time, several areas of improvement were identified, and some suggestions (revised feedback scores, revised version of the ERT, possibility to operate Cognition in offline mode) have already been implemented in a revised version of Cognition. As Cognition was recently selected as NASA's standard cognitive test battery for research, both for ground-based and spaceflight studies, future comparisons and evaluations of acceptability across different research settings will be achievable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

The research was supported by the National Space Biomedical Research Institute (NSBRI) through NASA NCC 9-58 and by NASA through grants NNX14AM81G, NNX14AH27G, NNX14AH98G, NNX16AI53G, 80NSSC18K0765 and 80NSSC17K0644. The authors declare no conflicts of interest related to the work presented in this manuscript.

References

[1]. Strangman GE, Sipes W, and Beven G. Human Cognitive Performance in Spaceflight and Analogue Environments. Aviat Space Environ Med 85, no. 10 (10 2014): 1033–48. [PubMed: 25245904]

[2]. Basner M, Savitt A, Moore TM, Port AM, McGuire S, Ecker AJ, et al. Development and Validation of the Cognition Test Battery for Spaceflight. Aerospace medicine and human performance. 2015;86(11):942–52. [PubMed: 26564759]

[3]. Basner M, Hermosillo E, Nasrini J, Saxena S, Dinges DF, Moore TM, et al. Cognition test battery: Adjusting for practice and stimulus set effects for varying administration intervals in high

performing individuals. Journal of Clinical and Experimental Neuropsychology. 2020;42(5):516–29. [PubMed: 32539487]

[4]. Garrett-Bakelman FE, Darshi M, Green SJ, Gur RC, Lin L, Macias BR, McKenna MJ, et al. The Nasa Twins Study: A Multidimensional Analysis of a Year-Long Human Spaceflight. Science 364, no. 6436 (4 12 2019).

[5]. Nasrini J, Hermosillo E, Dinges DF, Moore TM, Gur RC, Basner M. Cognitive Performance During Confinement and Sleep Restriction in NASA's Human Exploration Research Analog (HERA). Frontiers in physiology. 2020;11:394. [PubMed: 32411017]

[6]. Goemaere S, Van Caelenberg T, Beyers W, Binsted K, Vansteenkiste M. Life on mars from a Self-Determination Theory perspective: How astronauts' needs for autonomy, competence and relatedness go hand in hand with crew health and mission success - Results from HI-SEAS IV. Acta Astronautica. 2019;159:273–85.

[7]. Gur RC, Ragland JD, Moberg PJ, Turner TH, Bilker WB, Kohler C, et al. Computerized Neurocognitive Scanning: I. Methodology and Validation in Healthy People. Neuropsychopharmacology. 2001;25(5):766–76. [PubMed: 11682260]

[8]. Glahn DC, Gur RC, Ragland JD, Censits DM, Gur RE. Reliability, performance characteristics, construct validity, and an initial clinical application of a Visual Object Learning Test (VOLT). Neuropsychology. 1997;11(4):602–12. [PubMed: 9345704]

[9]. Ragland JD, Turetsky BI, Gur RC, Gunning-Dixon F, Turner T, Schroeder L, et al. Working memory for complex figures: An fMRI comparison of letter and fractal n-back tasks. Neuropsychology. 2002;16(3):370–9. [PubMed: 12146684]

[10]. Glahn DC, Cannon TD, Gur RE, Ragland JD, Gur RC. Working memory constrains abstraction in schizophrenia. Biological Psychiatry. 2000;47(1):34–42. [PubMed: 10650447]

[11]. Benton AL, Varney NR, Hamsher KD. Visuospatial Judgment: A Clinical Test. Archives of Neurology. 1978;35(6):364–7. [PubMed: 655909]

[12]. Gur RC, Richard J, Hughett P, Calkins ME, Macy L, Bilker WB, et al. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. Journal of Neuroscience Methods. 2010;187(2):254–62. [PubMed: 19945485]

[13]. Raven J The Raven's Progressive Matrices: Change and stability over culture and time. Cognitive Psychology. 2000;4(1):1–48.

[14]. Usui N, Haji T, Maruyama M, Katsuyama N, Uchida S, Hozawa A, et al. Cortical areas related to performance of WAIS Digit Symbol Test: A functional imaging study. Neuroscience Letters. 2009;463(1):1–5. [PubMed: 19631255]

[15]. Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, et al. Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). Journal of Experimental Psychology: Applied. 2002;8(2):75–84. [PubMed: 12075692]

[16]. Basner M, Dinges DF. Maximizing Sensitivity of the Psychomotor Vigilance Test (PVT) to Sleep Loss. Sleep. 2011;34(5):581–91. [PubMed: 21532951]

[17]. Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, et al. Cumulative Sleepiness, Mood Disturbance, and Psychomotor Vigilance Performance Decrements During a Week of Sleep Restricted to 4–5 Hours per Night. Sleep. 1997;20(4):267–77. [PubMed: 9231952]

[18]. Barger LK, Flynn-Evans EE, Kubey A, Walsh L, Ronda JM, Wang W, et al. Prevalence of sleep deficiency and use of hypnotic drugs in astronauts before, during, and after spaceflight: an observational study. The Lancet Neurology. 2014;13(9):904–12. [PubMed: 25127232]

[19]. Elo S, Kyngäs H. The qualitative content analysis process. Journal of Advanced Nursing. 2008;62(1):107–15. [PubMed: 18352969]

[20]. Hsieh H-F, Shannon SE. Three Approaches to Qualitative Content Analysis. Qualitative Health Research. 2005;15(9):1277–88. [PubMed: 16204405]

[21]. Gjedrem WG. Relative performance feedback: Effective or dismaying? Journal of Behavioral and Experimental Economics. 2018;74:1–16.

[22]. Morse JM. Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. Qualitative Health Research. 2015;25(9):1212–22. [PubMed: 26184336]

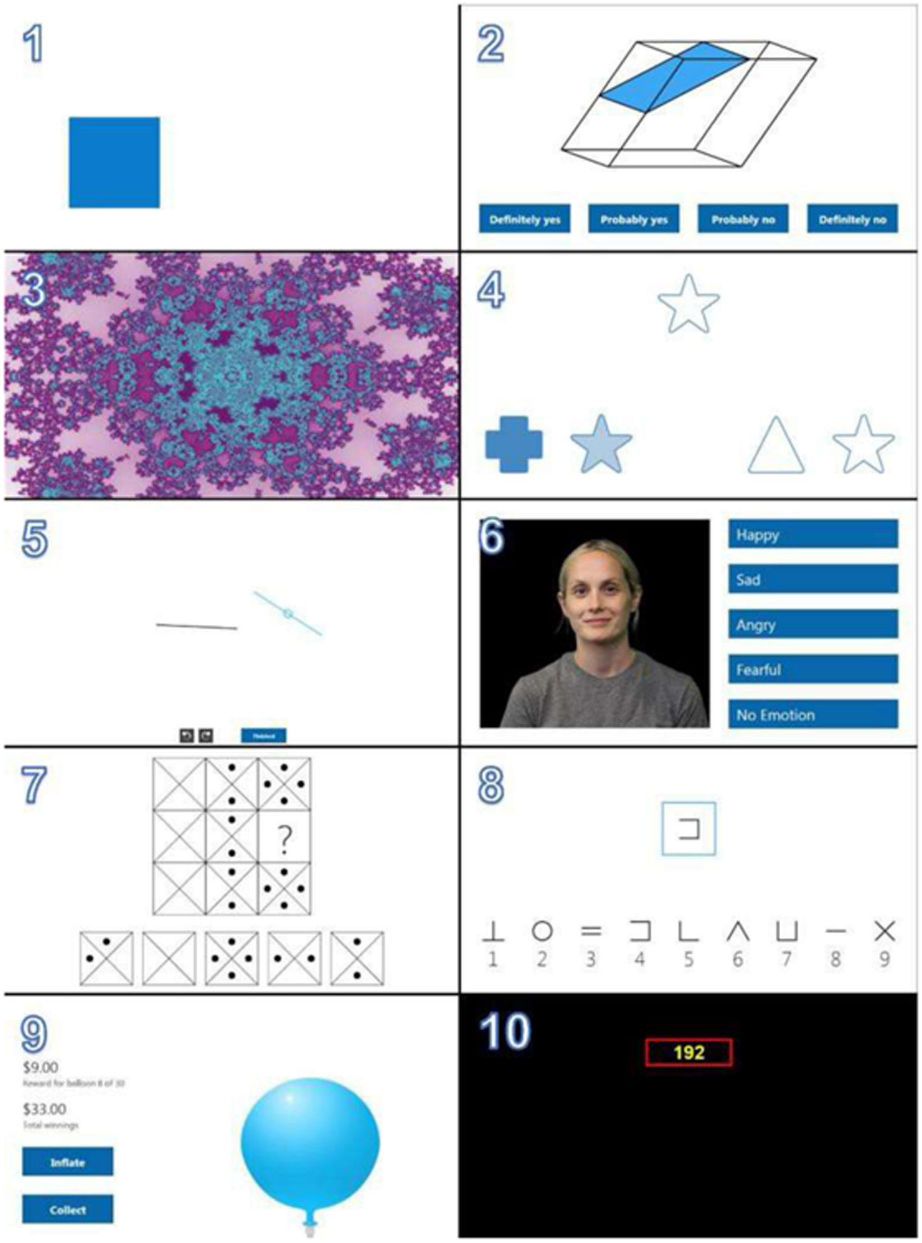

Figure 1:

Cognition Test Battery *1 = Motor Praxis (MP); 2 = Visual Object Learning Test (VOLT); 3 = Fractal 2-Back (F2B); 4 = Abstract Matching (AM); 5 = Line Orientation Test (LOT); 6 = Emotion Recognition Task (ERT); 7 = Matrix Reasoning Test (MRT); 8 = Digit Symbol Substitution Task (DSST); 9 = Balloon Analog Risk Test (BART); 10 = Psychomotor Vigilance Test (PVT)*

Table 1:

Study characteristics and demographics of study participants, N = 87. Study characteristics include number of participants per campaign, study duration, and number of Cognition test bouts completed. Participants in HERA and HI-SEAS performed the Cognition battery on Apple iPads, while the remaining participants completed the battery on laptops.

				Age		Sex	Education
Study	**N**	**Study Duration**	**# of Test Bouts per Subject**	**Mean**	**± SD**	**% Male**	**% Masters or Higher**
JSC Astronauts[×]	8	6.5 months	15	45.0	7.3	62.5	87.5
JSC Mission Control[◊]	11	6.5 months	15	30.0	6.1	45.5	45.4
HERA							
Campaign #1	16	1 week	7	38.3	7.7	37.5	75.0
Campaign #2	15	2 weeks	16	31.3	10.7	60.0	93.3
Campaign #3	16	30 days	18	36.6	7.2	56.2	81.3
HI-SEAS							
Campaign #1	5	4 months	12	29.0	2.7	40.0	40.0
Campaign #2	6	8 months	24	32.0	5.3	50.0	83.3
Campaign #3	6	12 months	33	30.8	4.2	50.0	100.0
ISS[0]	4	6 – 12 months	18	45.3	3.8	100.0	100.0
Overall	87			35.1	8.7	52.8	75.9

SD: standard deviation

[×] JSC Astronauts study duration ranged between 4.3 months to 8.2 months.

[◊] JSC MC study duration ranged between 5.5 months to 7.8 months.

[0] ISS study duration was 6 months (n=3) or 12 months (n=1). n = Number of comments.

Table 2:

Researchers coded participants' comments on their experience with the 10 individual tests as “positive”, “neutral or mixed”, or “negative” valence. Participants reported their most and least favorite tests in terms of test preference. Net score is calculated by subtracting frequency of tests reported under “Liked Least” from frequency of tests reported under “Liked Most”.

Test	Comments					Test Preference		
	n	% of N	Positive	Neutral/Mixed	Negative	Liked Most	Liked Least	Net
Motor Praxis (MP)	19	21.8	9	10	0	12	1	11
Visual Object Learning (VOLT)	25	28.7	13	12	0	28	2	26
Fractal 2-Back (F2B)	38	47.7	6	22	10	18	24	−6
Abstract Matching (AM)	30	34.5	8	11	11	21	15	6
Line Orientation (LOT)	24	27.6	5	16	3	9	4	5
Emotion Recognition (ERT)	44	50.6	1	32	11	2	44	−42
Matrix Reasoning (MRT)	26	29.9	5	15	6	11	2	9
Digit Symbol Substitution (DSST)	26	29.9	6	19	1	10	1	9
Balloon Analog Risk (BART)	46	52.9	8	26	12	17	18	−1
Psychomotor Vigilance (PVT)	42	48.3	9	17	16	13	22	−9
Total	320		70	180	70	141	133	8

n = Number of comments.

N = 87 participants.