Published in final edited form as:

Aerosp Med Hum Perform. 2015 November ; 86(11): 942–952. doi:10.3357/AMHP.4343.2015.

Development and Validation of the *Cognition* Test Battery for Spaceflight

Mathias Basner, MD, PhD[1,*], **Adam Savitt**[2,*], **Tyler M. Moore, PhD**[2], **Allison M. Port**[2], **Sarah McGuire, PhD**[1], **Adrian J. Ecker**[1], **Jad Nasrini**[1], **Daniel J. Mollicone, PhD**[3], **Christopher M. Mott**[3], **Thom McCann**[4], **David F. Dinges, PhD**[1], and **Ruben C. Gur, PhD**[2]

[1]Division of Sleep and Chronobiology, Department of Psychiatry, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

[2]Brain Behavior Laboratory, Department of Psychiatry, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

[3]Pulsar Informatics, Inc., Philadelphia, PA

[4]Joggle Research, Seattle, WA

Abstract

Background—Sustained high-level cognitive performance is of paramount importance for the success of space missions, which involve environmental, physiological and psychological stressors that may affect brain functions. Despite subjective symptom reports of cognitive fluctuations in spaceflight, the nature of neurobehavioral functioning in space has not been clarified.

Methods—We developed a computerized cognitive test battery (*Cognition*) that has sensitivity to multiple cognitive domains and was specifically designed for the high-performing astronaut population. *Cognition* consists of 15 unique forms of 10 neuropsychological tests that cover a range of cognitive domains including emotion processing, spatial orientation, and risk decision making. *Cognition* is based on tests known to engage specific brain regions as evidenced by functional neuroimaging. Here we describe the first normative and acute total sleep deprivation data on the *Cognition* test battery as well as several efforts underway to establish the validity, sensitivity, feasibility, and acceptability of *Cognition*.

Results—Practice effects and test-retest variability differed substantially between the 10 *Cognition* tests, illustrating the importance of normative data that both reflect practice effects and differences in stimulus set difficulty in the population of interest. After one night without sleep, medium to large effect sizes were observed for 3 of the 10 tests addressing vigilant attention (Cohen's d=1.00), cognitive throughput (d=0.68), and abstract reasoning (d=0.65).

Corresponding author: Mathias Basner, MD, PhD, MSc, Associate Professor of Sleep and Chronobiology in Psychiatry, Perelman School of Medicine at the University of Pennsylvania, 1013 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA, Tel: +1 215 573-5866, Fax: +1 215 573-6410, basner@upenn.edu.

*These authors contributed equally to the manuscript.

All other authors declare no conflicts of interest related to the work presented in this manuscript.

Conclusions—In addition to providing neuroimaging-based novel information on the effects of spaceflight on a range of cognitive functions, *Cognition* will facilitate comparing the effects of ground-based analogs to spaceflight, increase consistency across projects, and thus enable meta-analyses.

Keywords

neuropsychological test; cognitive test; cognition; space; spaceflight; performance; astronaut; microgravity; stress; confinement; isolation; sleep deprivation

Introduction

Successful human space exploration depends on the integrity of a range of cognitive abilities for unprecedented durations. Errors and accidents may have debilitating or fatal consequences, lead to the loss of expensive equipment, and compromise mission success. In addition to the physiological effects of microgravity, the spacecraft setting can involve exposure to a number of environmental toxicants and operational stressors that have the potential to degrade astronaut cognitive performance. Among these factors are radiation, noise, hypercapnia, hypoxia, decompression, dietary restrictions, fluid shifts, increased intracranial pressure, side effects of certain medications, and psychological factors related to isolation, confinement, and operational and interpersonal distress. Sleep of sufficient length and quality is of paramount importance for high levels of daytime performance, yet astronauts on Space Shuttle and International Space Station (ISS) missions have averaged less than 6.1 h sleep per 24 h (2). This amount of sleep is comparable to that of chronic sleep restriction, which has been shown to induce cognitive and neurobehavioral deficits and negative health outcomes (1). Although the reasons for reduced sleep durations in space flight are unknown, factors that contribute to sleep disturbance in spaceflight include non-24 h light–dark cycles, acute operational shifts in sleep timing, high workload and physical stress.

While astronauts have reported cognitive symptoms (often referred to as space fog or neurasthenia (23)), especially after initial exposure to the spacecraft environment, the results of objective cognitive testing in spaceflight have been inconclusive and most often fail to show statistically significant changes in cognitive performance (42). Therefore, the extent, etiology and persistence of these symptoms are still unknown. Several factors may contribute to the discrepancy between subjective symptom reports and objective assessment of cognitive functions in space.

Performance on most cognitive tests improves with repeated administration. This practice effect may confound (or mask) any cognitive deficits induced by the spacecraft environment. Depending on the complexity of the cognitive tests and given pre-mission time constraints, it will often not be possible to achieve asymptotic performance levels pre-flight. Additionally, existing studies often lack adequate ground-based control groups, and ground-based normative data usually do not exist for the astronaut population (42).

Spaceflight studies are often underpowered due to small sample sizes, and test batteries and cognitive domains they assess typically differ among studies, complicating systematic meta-

analyses that could increase statistical power. Furthermore, the tests used in the individual studies may simply lack sensitivity because they were designed for clinical populations or populations with lower aptitudes. They may be sensitive enough to detect symptoms associated with manifest disturbances such as severe brain trauma, but fail to detect sub-clinical deficits that can degrade optimal performance in high functioning individuals such as astronauts. Although sub-clinical deficits may not constitute an operational concern, they can be valuable in the early detection of environmental or psychological stressors and thus in the prevention of manifest cognitive deficits.

Another limitation of available test batteries is that they often only probe a few cognitive domains (i.e., they are not comprehensive), despite having multiple cognitive evaluations. It is thus possible that deficits in domains not covered by these batteries may have been overlooked. NASA currently uses the WinSCAT test battery operationally (24). WinSCAT consists of a 5-test subset of the larger Automated Neuropsychological Assessment Metrics (ANAM) test system developed by the Department of Defense (39): (1) Mathematical Processing; (2) Running Memory Continuous Performance; (3) Delayed Matching to Sample; (4) Code Substitution; and (5) Code Substitution Delayed Recognition. The cognitive domains assessed by these tests are: (1) basic computational skills and working memory; (2) attention and working memory; (3) spatial processing and visuo-spatial working memory; (4) complex scanning, visual tracking, and attention; and (5) memory. Therefore, the WinSCAT predominantly probes working memory, but fails to assess cognitive domains like spatial orientation, abstract reasoning, emotion processing, stability of sustained attention, and risk decision making that are also important for space mission success.

There are other reasons why cognitive deficits may go undetected in spaceflight. Well-educated, highly trained, motivated astronauts may be able to transiently compensate for deficits in cognitive performance induced during spaceflight by teamwork and other strategies. Countermeasures used by astronauts may reverse or mask a cognitive deficit. Astronauts may not subjectively be aware of some cognitive deficits that could be detected by a sensitive test battery. For example, sleep deprivation studies indicate that subjective and objective assessments of performance may differ substantially (45). Although performance capability is mostly overestimated during periods of sleep restriction, especially during the biological night, it is possible that the opposite can happen in spaceflight. This underscores the need for brief and valid objective assays of cognitive performance in spaceflight.

Here, we describe the development of an improved neurocognitive assessment tool, named *Cognition*, provide learning curves for each of its 10 cognitive tests, and show that each test differs in its sensitivity to acute total sleep deprivation. *Cognition* was specifically designed to assess cognitive functions in astronauts and address some of the >25 knowledge gaps and health risks that mention cognition in NASA's Human Research Roadmap (12). *Cognition* covers the main cognitive domains – executive, episodic memory, complex cognition, social cognition and sensorimotor speed – and is based on tests known to engage specific brain systems during functional neuroimaging (19, 41). The latter may provide information on the neurostructural origin of a cognitive deficit, which is important as we currently have no neuroimaging capability in spaceflight. *Cognition* is involved in several ongoing and soon-

to-be commenced validation studies. The goal is to advance knowledge on the cognitive effects of spaceflight by offering a brief and well-validated battery of tests that are acceptable to the astronaut population, feasible in spaceflight, and that provide crucial clinical feedback on neurobehavioral functions in space. This research tool will hopefully increase consistency across projects and facilitate meta-analyses.

Methods

Subjects

A normative study in astronauts/astronaut candidates (N=8, mean age 44.1 years, range 34–53 years, 38% female) and mission controllers (N=11, mean age 28.0 years, range 22–38 years, 55% female) was completed in 2014 at Johnson Space Center, Houston, TX. This study was approved by NASA's Institutional Review Board and subjects signed written informed consent prior to study participation. The Cognition battery was also performed by a total of 44 different subjects (mean age ± SD 34.1 ± 8.7 years, 50% male) in two sleep restriction protocols that included acute total sleep deprivation (i.e., one night without sleep). These studies were approved by the Institutional Review Board of the University of Pennsylvania, and subjects signed written informed consent prior to study participation.

Equipment

The *Cognition* battery contains a subset of tests from a widely used and validated neurocognitive battery, the Penn Computerized Neurocognitive Battery (CNB) (18, 19, 31), as well as a number of additional tests that have either been used extensively in spaceflight (i.e., Psychomotor Vigilance Test (28), Digit Symbol Substitution Test (43)) or assess cognitive domains of particular interest in spaceflight. The CNB is currently being used in cognitive assessment of military enlistees, cognitive development in children, cognitive therapies for schizophrenia patients, and genomic research in populations with or at risk for schizophrenia (19).

Table I provides an overview of the cognitive domains assessed and brain regions primarily recruited by each of the 10 tests of the *Cognition* battery. It also shows average adminsitration time for each test based on a study performed in astronauts, astronaut candidates and mission controllers at Johnson Space Center (see 4 below). Screenshots of each of the 10 *Cognition* tests are shown in Figure 1. The tests are described in detail below:

The **Motor Praxis Task (MP)** (17) is administered at the start of testing to ensure that participants have sufficient command of the computer interface, and immediately thereafter as a measure of sensorimotor speed. Participants are instructed to click on squares that appear randomly on the screen, each successive square smaller and thus more difficult to track. Performance is assessed by the speed with which participants click each square. The current implementation uses 20 consecutive stimuli. As a screener for computer skill, the MP has been included in every implementation of the CNB and validated for sensitivity to age effects (18), sex-differences (40), and associations with psychopathology (33).

The **Visual Object Learning Test (VOLT)** assesses participant memory for complex figures (14). Participants are asked to memorize 10 sequentially displayed three-dimensional

figures. Later, they are instructed to select those objects they memorized from a set of 20 such objects also sequentially presented, some of them from the learning set and some of them new. Such tasks have been shown to activate frontal and bilateral anterior medial temporal lobe regions (22). As the hippocampus and medial temporal lobe are also adversely affected by chronic stress (42), the VOLT offers a validated tool for assessment of these temporo-limbic regions in operational settings.

The **Fractal 2-Back (F2B)** (36) is a nonverbal variant of the Letter 2-Back, which is currently included in the core CNB (19). N-back tasks have become standard probes of the working memory system, and activate canonical working memory brain areas. The F2B consists of the sequential presentation of a set of figures (fractals), each potentially repeated multiple times. Participants have to respond when the current stimulus matches the stimulus displayed two figures ago. The current implementation uses 62 consecutive stimuli. The fractal version was chosen for *Cognition* because of its increased difficulty (36) and the availability of algorithms with which new items can be generated. Traditional letter N-back tasks are restricted to 26 English letters, which limits the ability to generate novel stimuli for repeat administrations. The Running Memory Continuous Performance task implemented in WinSCAT is a 1-Back task that uses the numbers 0–9 as stimuli. The F2B implemented in *Cognition* is well-validated and shows robust activation of the dorsolateral prefrontal cortex (36).

The **Abstract Matching (AM)** test (13) 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 presents 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 are presented with a target object in the upper middle of the screen that they must classify as more belonging with one of the two pairs, based on a set of implicit, abstract rules. The current implementation uses 30 consecutive stimuli. Tasks assessing abstraction and cognitive flexibility activate the prefrontal cortex (7).

The **Line Orientation Test (LOT)** is measure of spatial orientation and derived from the well-validated Judgment of Line Orientation Test (6), the computerized version of which was among the first to be administered with functional neuroimaging (16) and is used in the core CNB (19). It has been shown to be sensitive to sex differences and age effects (18). The LOT format consists of presenting two lines at a time, one stationary and the other can be rotated by clicking an arrow. Participants rotate the movable line until it is parallel to the stationary line. The current implementation has 12 consecutive line pairs that vary in length and orientation. Difficulty is determined by the length of the rotating line, its distance from the stationary line, and number of degrees that the line rotates with each click. Spatial orientation and reasoning are crucial for success in space missions, being necessary for repairs, craft piloting, and safe maneuvering in microgravity.

The **Emotion Recognition Task (ERT)** was developed (20) and validated with neuroimaging (30) and is part of the Penn CNB (19). The ERT presents subjects with photographs of professional actors (adults of varying age and ethnicity) portraying emotional facial expressions of varying intensities (biased towards lower intensities and

balanced across the different versions of the test). Subjects are given a set of emotion labels ("happy"; "sad"; "angry"; "fearful"; and "no emotion") and must select the label that correctly describes the expressed emotion. The current implementation uses 40 consecutive stimuli, with 8 stimuli each representing one of the above 5 categories. ERT performance has been associated with amygdala activity in a variety of experimental contexts, showing its sensitivity to such diverse phenomena as menstrual cycle phase, mood and anxiety disorders, and schizophrenia (9, 21). Additionally, sensitivity to hippocampal activation has been demonstrated when positive and negative emotional valence performance is analyzed alone (on a specific task only probing valence) (21). The test has also shown sensitivity to sex differences and normal aging (15). The ERT was included in the Cognition battery because of its well-studied brain response and the importance of emotion identification abilities to long-term social interactions. Emotional STROOP tasks have been administered in spaceflight, and show that at least some components of emotional processing are adversely affected by spaceflight (34).

The **Matrix Reasoning Test (MRT)** is a measure of abstract reasoning and consists of increasingly difficult pattern matching tasks (17). It is analogous to Raven Progressive Matrices (38) and recruits prefrontal, parietal, and temporal cortices (35). It is based on a well-known measure of the "g" factor. The test consists of a series of patterns, overlaid on a grid. One element from the grid is missing and the participant must select the element that fits the pattern from a set of alternative options. The current implementation uses 12 consecutive stimuli. MRT administration will stop automatically if three consecutive stimuli are answered incorrectly. The MRT is included in the Penn CNB, and has been validated along with all other tests in major protocols using the CNB (19).

The **Digit-Symbol Substitution Task (DSST)** (43) is a computerized adaptation of a paradigm used in the Wechsler Adult Intelligence Scale (WAIS-III). The DSST requires the participant to refer to a displayed legend relating each of the digits one through nine to specific symbols. One of the nine symbols appears on the screen and the participant must select the corresponding number as quickly as possible. The test duration is fixed at 90 s, and the legend key is randomly re-assigned with each administration. Fronto-parietal activation associated with DSST performance has been interpreted as reflecting both onboard processing in working memory and low-level visual search (43). The DSST is also part of WinSCAT where it is called "Code Substitution". In WinSCAT's "Delayed Recognition" version, the test is repeated without the legend to probe memory.

The **Balloon Analog Risk Test (BART)** is a validated assessment of risk taking behavior and has been shown to robustly activate striatal mesolimbic-frontal regions not covered by existing batteries (27, 37). The BART requires participants to either inflate an animated balloon or collect a reward. Participants are rewarded in proportion to the final size of each balloon, but a balloon will pop after a hidden number of pumps, which changes from trial to trial (27). The current implementation uses 30 consecutive stimuli. The average tendency of balloons to pop is systematically varied between test administrations. This requires subjects to adjust the level of risk they take based on the behavior of the balloons, and prevents subjects from identifying a strategy during the first administrations of the battery and carrying it through to later administrations. The ability to effectively weigh risks is

compromised in conditions of sleep deprivation (26). Risk taking behavior is crucially important to measure in spaceflight as alterations in self-monitoring and introspection may lead astronauts to accept risks that would otherwise be rejected.

The **Psychomotor Vigilance Test (PVT)** records reaction times (RT) to visual stimuli that occur at random inter-stimulus intervals (3). Subjects are instructed to monitor a box on the screen, and hit the space bar once a millisecond counter appears in the box and starts incrementing. The reaction time will then be displayed for one second. Subjects are instructed to be as fast as possible without hitting the spacebar without a stimulus (i.e., false starts or errors of commission). The PVT is a sensitive measure of vigilant attention and the effects of acute and chronic sleep deprivation and circadian misalignment, conditions highly prevalent in spaceflight (2). The PVT has negligible aptitude and learning effects (3), and is ecologically valid as sustained attention deficits and slow reactions affect many real-world tasks (e.g., operating a vehicle) (10). Differential activation to PVT performance has been shown across sleep-deprivation conditions, displaying increased activation in right fronto-parietal sustained attention regions when performing optimally, and increased default-mode activation after sleep deprivation, thought to be a compensatory mechanism (11). *Cognition* uses a validated 3-min. version of the PVT with shorter inter-stimulus intervals (2–5 s instead of 2–10 s) (5). This version has been administered >2,500 times in 24 astronauts during 6-month ISS missions.

Procedure

For the normative study, we installed the *Cognition* software on the NASA-issued laptops of each astronauts/astronaut candidate. The mission controllers performed the tests on one of two designated laptops in a quiet room of the mission control building at JSC. Each participant performed all 15 unique versions of the battery. All laptops were calibrated for timing precision. The scheduled intervals between test administrations were either two weeks (tests 1–5 and 9–15) or one week (tests 6–8) approximating the intervals used during 6-month ISS missions. Participants were instructed not to take the test within one hour of waking up or after being awake more than 16 hours. They were also instructed not to perform *Cognition* if they were sick. Sleep times on the day of test administration, the consumption of stimulants and depressants in the 6 hours preceding the test, and alertness levels were assessed with a questionnaire prior to test administration. Subjects were asked to use the trackpad of the laptop and not to attach a computer mouse, as the latter is not available in space. Astronauts/astronaut candidates were asked to take the test in an environment without distractions. All participants were instructed to leave comments that could explain any irregularity in the data (e.g., distractions).

In the sleep deprivation protocol, subjects performed *Cognition* on a daily basis on calibrated laptops shortly after 11 am. They were free of acute and chronic medical and psychological conditions, as established by interviews, clinical history, questionnaires, physical exams, and blood and urine tests. They were studied in small groups (4–5) while they remained in the Sleep and Chronobiology Laboratory at the Hospital of the University of Pennsylvania. Subjects were continuously monitored by trained staff to ensure adherence to each experimental protocol. They wore wrist actigraphs throughout each protocol. Meals

were provided at regular times, caffeinated foods and drinks were not allowed, and light levels in the laboratory were held constant during scheduled wakefulness (<50 lux) and sleep periods (<1 lux). Ambient temperature was maintained between 22° and 24° C.

Statistical Analysis

For the normative study, we generated key speed and accuracy outcomes for each of the 10 Cognition tests. To facilitate comparisons between tests, all outcomes were standardized based on the mean and standard deviation of test outcomes of the first administration.

For the sleep deprivation study, effect sizes (Cohen's d) were calculated by subtracting test scores under alert conditions (i.e., on the day immediately preceding the night without sleep) from test scores at the same time of day after one night without sleep and dividing by the standard deviation of the differences. In addition to effect size point estimates, 95% bootstrap confidence intervals based on 100,000 replications were generated.

Results

Standardized performance across the 15 administrations is shown separately for key speed and accuracy outcomes of the 10 *Cognition* tests in Figure 2. The trajectory of speed and accuracy measures across repeated administrations varied greatly between tests. Although the greatest performance changes were usually observed during the first two or three administrations of the tests (e.g., VOLT, AM, PVT), some outcomes continued to change systematically with an increasing number of test administrations: On the subject-paced tasks, the time required to respond to a given number of stimuli decreased on the VOLT, ERT, BART, and DSST, and, to a lesser degree, on the MP and LOT, while the AM and MRT showed no systematic variation across administrations. Accuracy measures continued to improve with repeated administration on the VOLT, F2B, and AM, and, to a lesser degree, on the MRT, while MP, LOT, ERT, DSST, BART, and PVT either showed no systematic variation or even slight deterioration of test accuracy with repeated administration. Furthermore, test-retest variability differed between tests and was somewhat larger for the VOLT and ERT compared to the other tests.

Acute total sleep deprivation effects sizes are shown for several Cognition outcome metrics in Figure 3. Only slower response speed on the PVT reached a large effect size ($d > 0.8$), followed by lower DSST throughput, shorter MRT duration, and a greater number of lapses on the PVT with medium effect sizes ($0.5 < d < 0.8$). Shorter test durations on the VOLT, LOT, ERT, and AM as well as lower MRT accuracy and more errors on the DSST reached small effect sizes ($0.2 < d < 0.5$). This suggests that subjects tended to rush through the subject-paced tasks while sleep deprived. All other outcomes reached small or negligible effect sizes that were not significantly different from 0.

Discussion

In this manuscript, we describe the first normative and sleep deprivation data on the *Cognition* test battery. In the normative study in astronauts/astronaut candidates and mission controllers, practice effects and test-retest variability differed substantially between the 10

different *Cognition* tests. These data illustrate two important points. First, for a subset of the tests it will not be possible to achieve asymptotic performance levels pre-flight even with the 6 administrations currently used for WinSCAT. Astronauts will continue to learn in space, and these effects may confound any spaceflight effect if not properly taken into account. Second, although we specifically designed stimulus sets with comparable difficulty, the data show that variability in stimulus difficulty was still high for the VOLT and ERT. This variability needs to be taken into account when comparing data across several administrations of the *Cognition* battery. These points underscore the importance of normative data that both reflect practice effects and differences in stimulus set difficulty in the population of interest.

The *Cognition* data gathered before and after one night without sleep demonstrate the differential sensitivity of individual Cognition tests to acute total sleep deprivation. Only the PVT, DSST, and MRT reached medium to large effect sizes, while the other tests reached small or negligible effect sizes. Overall, these findings are consistent with the sleep deprivation literature (29).

The data presented in this manuscript address two important aspects of a cognitive test battery for spaceflight, i.e., the existence of normative data in the population of interest and the sensitivity of the battery to common spaceflight stressors. However, there are other important criteria a cognitive test battery for spaceflight needs to meet:

One important criterion is the breadth of cognitive domains covered. Spaceflight is characterized by many unique environmental and psychological stressors related to living in an isolated, confined, and extreme setting that may affect aspects of cognitive performance. Therefore, a cognitive test battery needs to cover a wide range of cognitive domains relevant for spaceflight. Current batteries lack emotional domain coverage and fail to include tasks sensitive to disruptions of particularly vulnerable brain regions (24, 42). The hippocampus and medial temporal lobe, as well as striatal regions including the basal ganglia, are vulnerable to damage caused by radiation and chronic stress (8, 42). The latter can be brought on by high workload, sleep disruption, and other characteristic risks in spaceflight missions. *Cognition* contains tests that have been experimentally associated with activity in these brain regions using functional neuroimaging, and it covers a wide breadth of cognitive abilities including emotion recognition, spatial orientation, and risk decision making (see Table I), which are capabilities that solidly contribute to mission success.

Another important criterion is test administration time, which is limited both operationally (e.g., high workload during 6-month ISS missions) and in terms of astronauts' limited willingness to perform prolonged cognitive testing in space. However, the duration of individual tests cannot be shortened ad libitum. This specifically applies to tests that measure vigilance attention (like the PVT), as the time-on-task related vigilance decrement unmasks fatigue, and the likelihood for subjects being able to compensate decreases with increasing test duration. We reduced the standard 10-min. administration time of the PVT to 3-min, and changed the properties of the test at the same time (e.g., shorter inter-stimulus intervals, lower lapse threshold) to conserve the test's sensitivity (5).

For Cognition, the test duration was optimized in a data driven manner where possible (i.e., maximizing accuracy of performance prediction while minimizing test duration). We are currently working on adaptive versions of the ERT and MRT. Adaptive testing is based on Item Response Theory (IRT), a psychometric technique based on the application of mathematical models to testing data; it is considered to supersede classical test theory (44), and it is the method now used for the Graduate Record Exam (GRE) and Graduate Management Admission Test (GMAT). In adaptive testing, the difficulty of test items is chosen based on prior responses of an individual subject. This results in fewer responses being needed to reliably estimate the subject's test score, which reduces test time considerably. We have already developed an adaptive version of the PVT (4) and the LOT (32) for future implementations of the *Cognition* battery. Finally, we implemented time outs for the subject-paced tasks MPT and MRT to prevent subjects from spending excessive amounts of time on individual stimuli and inflating administration time of the whole battery. Median administration times for WinSCAT (based on 1,300 sleep laboratory administrations) and Cognition are 13.5 minutes and 17.8 minutes, respectively.

The *Cognition* tests were specifically designed for the high performing astronaut population. For example, WinSCAT uses a 1-Back paradigm to probe working memory, where the left mouse button is pushed whenever the number on the previous screen matches the number on the current screen, and the right mouse button is pushed otherwise. This is likely too easy for the astronaut population. By contrast, in the Fractal 2-Back paradigm we used for *Cognition*, instead of numbers, astronauts must remember abstract fractals, and they have to push the spacebar if they have seen the same fractal two screens ago. This is more difficult and less likely to produce ceiling effects.

A spaceflight battery needs to allow for repeated administration to test for systematic changes in cognitive performance over time in mission. Administering the same version of each test across multiple time points is a suboptimal strategy for assessing cognitive abilities. For simple motor response tests (like the PVT), repeated administration is not problematic, as the test is not influenced by either learning or aptitude (28) (see Figure 2). For others that rely on specific stimuli (like the VOLT and MRT) large item pools need to be generated and validated. For example, we generated more than 600 unique stimuli for the ERT and more than 300 unique stimuli for the VOLT and the MRT. These were then subjected to crowd sourcing (25) to verify the psychometric properties of each item, but also of the combination of items (e.g., targets and decoys in the VOLT). Currently, there are 15 unique versions of the *Cognition* battery available.

Practice effects, especially in those tests assessing memory confound (or mask) the true effect of stressors on cognitive abilities. They therefore need to be taken into account for data analysis and interpretation. *Cognition* ameliorates this problem by generating multiple unique and comparable versions of each test (see above). In general, practice effects can be addressed in several ways. One common approach is to have astronauts perform the tests multiple times before launch. For some tests, asymptotic performance levels may be reached after a few administrations. For others (e.g., see VOLT below), practice effects continue to occur even after 10 or more administrations, and it is impractical to have astronauts perform a test battery that many times pre-flight (WinSCAT is currently performed six times pre-

flight). A second approach is to compare the performance of astronauts in space to ground-based control groups that perform the test battery at similar intervals. This is a powerful design, but it is costly, and it can sometimes be hard to recruit suitable astronaut controls. Control groups have been rare in cognitive testing in spaceflight (42). A third approach generates astronaut data pre-, in-, and post-flight that is compared to normative data derived in the population of interest (i.e., astronauts, see 4 below). We are currently developing this normative data set which reflects both practice effects and random differences in stimulus set difficulty.

A cognitive battery for spaceflight needs to meet several operational requirements. The software needs to be easy to use, designed for self-administration, appealing and provide meaningful and immediate feedback to increase astronaut compliance. Based on >40 de-briefs of astronauts and astronaut surrogates, *Cognition* fulfills these criteria. Eight of the ten tests have practice bouts that can be performed immediately prior to test administration—this is useful for first-time participants and those who have not taken the tests for a longer period of time. If a user skips out of the battery before all tests are finished, the software will remember the last completed test bout and start with the next test the next time it is started. Feedback is provided immediately after each test as a standardized score ranging from 0 (worst possible performance) to 1000 (best possible performance). This score is based on both accuracy and speed, i.e., in order to receive a perfect score someone has to be both accurate and fast. The speed/accuracy weighting differs among tests. Historical test data are also displayed. Therefore, the astronaut can compare his current performance to past performances on the same test. A final score (sum of all standardized tests) is displayed at the end of the battery. Each test produces a number of summary metrics (e.g., average reaction time on the PVT), but information on individual stimuli is also generated by default (e.g., individual reaction times on the PVT).

Test administration should be flexible (e.g., a single test or the whole battery can be administered). The software needs to conform to ISS software standards to allow implementation on the ISS. Ideally, the software supports several hardware platforms (including handheld/tablet devices). Real time, remote access of the data for quality control purposes should be possible. *Cognition* was implemented on a cognitive testing platform that provides all these attributes. It is currently running on an HRF laptop in the Columbus module of the ISS, and both a Microsoft Windows 7 and an Apple iPad version of the battery exist. Response latencies of the iPad have been established, and Windows 7 laptops are calibrated with a robotic calibrator prior to deployment to ensure timing accuracy of the system. English, German, French, Italian, and Russian language versions were generated. *Cognition* has a brief survey module that can be administered before the first test. Our current ISS survey asks about sleep time, stimulant use, and the momentary alertness level. It was expanded for our Antarctic protocols (see below) to capture, among others, workload, stress, monotony, and crew conflicts. When a test is completed, the results are encrypted, stored on the local hard drive, and also transmitted to a central server if Internet connection is available. This allows for real-time quality control checks of the data.

A cognitive test battery for spaceflight does not only need to meet standard validity criteria. In addition, administration in space needs to be feasible, and it needs to be acceptable to the

astronaut population. Furthermore, sensitivity of the battery to the cognitive effects of common spaceflight stressors needs to be established. Several efforts are underway to establish the validity, feasibility, and sensitivity of *Cognition*. They are summarized below.

A basic validation study in an astronaut surrogate population (psychiatrically screened, age 25–56, half male, with Master's degree or higher) was started in early 2015. This study focuses on test reliability, correlations among the tests, item consistency, and criterion validity relative to gender. Each subject performs, in a balanced and randomized fashion, the Windows 7 version of *Cognition*, the iPad version of *Cognition*, and WinSCAT.

Data acquisition to assess the feasibility of *Cognition* on ISS started in November 2014 (the *Cognition* software is installed on one HRF laptop in the ISS Columbus module). *Cognition* is also part of the 12-month ISS mission that launched in March 2015 (both the US astronaut and the Russian cosmonaut signed up for *Cognition*) and of NASA's TWINS project. In the latter, one astronaut flies on a 12-month ISS mission, while his twin brother stays on the ground. Both brothers perform *Cognition* at similar intervals.

Cognition is also deployed in several space analog environments. NASA is performing several isolation studies over 1, 2, and 4 week periods in its Human Exploration Research Analog (HERA) at Johnson Space Center with a crew size of N=4. At the end of 2016, we will have gathered Cognition data on N=48 HERA subjects. Cognition is also part of 3 isolation studies (4, 8, and 12 month duration) performed in the Hawai'i Space Exploration Analog and Simulation (HI-SEAS) on the slopes of Mauna Loa on the Big Island of Hawai'i at approximately 8,200 feet above sea level. Each crew consists of N=6 subjects serving as astronaut surrogates. HERA and HI-SEAS crews are mixed-gender. Finally, *Cognition* is currently deployed at 4 Antarctic stations (Concordia, Neumayer, Halley, and SANAE) and is performed by the crew on a monthly basis during the winter-over periods of 2015 (all stations) and 2016 (Concordia only). Finally, it is planned to include *Cognition* in two long-term bed-rest studies, in a study on the effects of hypercapnia on cognitive performance, and in a study on the pharmacodynamics and pharmacokinetics of zaleplon and azithromycin in space.

In conclusion, a sustained high level of cognitive performance is of paramount importance for the success of space missions. The spaceflight milieu is characterized by several unique environmental and psychological stressors related to living in an isolated, confined, and extreme environment. These stressors may likely affect cognitive performance. Past research on the cognitive effects of spaceflight failed to show consistent changes in cognitive performance in space, despite frequent subjective symptom reports. This inconsistency could reflect the use of cognitive tests that lack sensitivity and validity for the astronaut population. Here we introduced a new cognitive test battery for spaceflight, *Cognition*, which was specifically designed for the high performing astronaut population. Preliminary evidence suggests that Cognition is a feasible, sensitive, and valid research tool for investigating the effects of spaceflight on astronaut performance.

Acknowledgments

The research was supported by the National Space Biomedical Research Institute (NSBRI) through NASA NCC 9-58 and by NASA through grants NNX14AM81G, NNX14AH27G, and NNX14AH98G. The sleep deprivation protocols were supported by NIH through R01NR00428 and the Office of Naval Research through N00014-11-1-0361. The authors thank John Hansen and Jason Schneiderman for their involvement in the early stages of the development of the Cognition test battery. *Cognition* was programmed by Pulsar Informatics, Inc. with the support of the NSBRI, project NBPF02501. Pulsar Informatics Inc. (Daniel J. Mollicone, Christopher M. Mott) and Joggle Research (Thom McCann) have a commercial interest in the *Cognition* battery.

References

1. Banks S, Dinges DF. Behavioral and physiological consequences of sleep restriction. JClinSleep Med. 2007; 3(5):519–28.
2. Barger LK, Flynn-Evans EE, Kubey A, Walsh L, Ronda JM, Wang W, et al. Prevalence of sleep deficiency and hypnotic use among astronauts before, during and after spaceflight: an observational study. Lancet Neurology. 2014; 13(9):904–12. [PubMed: 25127232]
3. Basner M, Dinges DF. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep. 2011; 34(5):581–91. [PubMed: 21532951]
4. Basner M, Dinges DF. An adaptive duration version of the PVT accurately tracks changes in psychomotor vigilance induced by sleep restriction. Sleep. 2012; 35(2):193–202. [PubMed: 22294809]
5. Basner M, Mollicone DJ, Dinges DF. Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. Acta Astronautica. 2011; 69:949–59. [PubMed: 22025811]
6. Benton AL, Varney NR, Hamsher KD. Visuospatial judgment. A clinical test. Archives of neurology. 1978; 35(6):364–7. [PubMed: 655909]
7. Berman KF, Ostrem JL, Randolph C, Gold J, Goldberg TE, Coppola R, et al. Physiological activation of a cortical network during performance of the Wisconsin Card Sorting Test: a positron emission tomography study. Neuropsychologia. 1995; 33(8):1027–46. [PubMed: 8524452]
8. Chetty S, Friedman AR, Taravosh-Lahn K, Kirby ED, Mirescu C, Guo F, et al. Stress and glucocorticoids promote oligodendrogenesis in the adult hippocampus. Molecular psychiatry. 2014; 19(12):1275–83.10.1038/mp.2013.190 [PubMed: 24514565]
9. Derntl B, Windischberger C, Robinson S, Lamplmayr E, Kryspin-Exner I, Gur RC, et al. Facial emotion recognition and amygdala activation are associated with menstrual cycle phase. Psychoneuroendocrinology. 2008; 33(8):1031–40. [PubMed: 18675521]
10. Dinges DF. An overview of sleepiness and accidents. JSleep Res. 1995; 4(S2):4–14. [PubMed: 10607205]
11. Drummond SP, Bischoff-Grethe A, Dinges DF, Ayalon L, Mednick SC, Meloy M. The neural basis of the psychomotor vigilance task. SLEEP-NEW YORK THEN WESTCHESTER-. 2005; 28(9):1059.
12. Foster, J. NASA Human Research Roadmap. 2010. Retrieved 12 August 2015 from http://humanresearchroadmap.nasa.gov/
13. Glahn DC, Cannon TD, Gur RE, Ragland JD, Gur RC. Working memory constrains abstraction in schizophrenia. Biological psychiatry. 2000; 47(1):34–42. Epub 2000/01/29. [PubMed: 10650447]
14. 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. [PubMed: 9345704]
15. Gunning-Dixon FM, Gur RC, Perkins AC, Schroeder L, Turner T, Turetsky BI, et al. Age-related differences in brain activation during emotional face processing. Neurobiology of aging. 2003; 24(2):285–95. [PubMed: 12498962]
16. Gur RC, Gur RE, Obrist WD, Hungerbuhler JP, Younkin D, Rosen AD, et al. Sex and handedness differences in cerebral blood flow during rest and cognitive activity. Science. 1982; 217(4560): 659–61. [PubMed: 7089587]

17. 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):766–76.10.1016/S0893-133X(01)00278-0 [PubMed: 11682260]
18. Gur RC, Richard J, Calkins ME, Chiavacci R, Hansen JA, Bilker WB, et al. Age group and sex differences in performance on a computerized neurocognitive battery in children age 8–21. Neuropsychology. 2012; 26(2):251–65.10.1037/a0026712 [PubMed: 22251308]
19. 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. Epub 2009/12/01. 10.1016/j.jneumeth.2009.11.017 [PubMed: 19945485]
20. Gur RC, Sara R, Hagendoorn M, Marom O, Hughett P, Macy L, et al. A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. Journal of neuroscience methods. 2002; 115(2):137–43. [PubMed: 11992665]
21. Gur RE, McGrath C, Chan RM, Schroeder L, Turner T, Turetsky BI, et al. An fMRI study of facial emotion processing in patients with schizophrenia. American Journal of Psychiatry. 2002; 159(12):1992–9. [PubMed: 12450947]
22. Jackson O III, Schacter DL. Encoding activity in anterior medial temporal lobe supports subsequent associative recognition. Neuroimage. 2004; 21(1):456–62. [PubMed: 14741683]
23. Kanas N, Salnitskiy V, Gushin V, Weiss DS, Grund EM, Flynn C, et al. Asthenia--does it exist in space? Psychosom Med. 2001; 63(6):874–80. Epub 2001/11/24. [PubMed: 11719624]
24. Kane RL, Short P, Sipes W, Flynn CF. Development and validation of the spaceflight cognitive assessment tool for windows (WinSCAT). Aviat Space Environ Med. 2005; 76(6 Suppl):B183–91. Epub 2005/06/10. [PubMed: 15943211]
25. Keutmann MK, Moore SL, Savitt A, Gur RC. Generating an item pool for translational social cognition research: Methodology and initial validation. Behavior research methods. 2015; 47(1): 228–34.10.3758/s13428-014-0464-0 [PubMed: 24719265]
26. Killgore WDS, Kamimori GH, Balkin TJ. Caffeine protects against increased risk-taking propensity during severe sleep deprivation. Journal of sleep research. 2011; 20(3):395–403.10.1111/j.1365-2869.2010.00893.x [PubMed: 20946437]
27. 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]
28. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Ann N Y Acad Sci. 2008; 1129:305–22.10.1196/annals.1417.002 [PubMed: 18591490]
29. Lim J, Dinges DF. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. Psychol Bull. 2010; 136(3):375–89. Epub 2010/05/05. 2010-07936-006 [pii]. 10.1037/a0018883 [PubMed: 20438143]
30. Loughead J, Gur RC, Elliott M, Gur RE. Neural circuitry for accurate identification of facial emotions. Brain research. 2008; 1194:37–44.10.1016/j.brainres.2007.10.105 [PubMed: 18191116]
31. Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric Properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. 201410.1037/neu0000093
32. Moore TM, Scott JC, Reise SP, Port AM, Jackson CT, Ruparel K, et al. Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. Psychological Assessment. 2015 in press.
33. de Neves MC, Albuquerque MR, Neves FS, Lage GM, Malloy-Diniz L, Nicolato R, et al. Sensorimotor performance in euthymic bipolar disorder: the MPraxis (PennCNP) analysis. Rev Bras Psiquiatr. 2014; 36(3):248–50. [PubMed: 24676046]
34. Pattyn N, Migeotte P-F, Morais J, Soetens E, Cluydts R, Kolinsky R. Crew performance monitoring: Putting some feeling into it. Acta astronautica. 2009; 65(3):325–9.
35. Perfetti B, Saggino A, Ferretti A, Caulo M, Romani GL, Onofrj M. Differential patterns of cortical activation as a function of fluid reasoning complexity. Hum Brain Mapp. 2009; 30(2):497–510. Epub 2007/12/21. 10.1002/hbm.20519 [PubMed: 18095280]

36. 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]

37. Rao H, Korczykowski M, Pluta J, Hoang A, Detre JA. Neural correlates of voluntary and involuntary risk taking in the human brain: an fMRI Study of the Balloon Analog Risk Task (BART). Neuroimage. 2008; 42(2):902–10. [PubMed: 18582578]

38. Raven, JC. Advanced Progressive Matrices: Sets I and II. London: Lewis; 1965.

39. Reeves DL, Winter KP, Bleiberg J, Kane RL. ANAM genogram: historical perspectives, description, and current endeavors. Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists. 2007; 22(Suppl 1):S15–37.10.1016/j.acn.2006.10.013 [PubMed: 17276030]

40. Roalf DR, Gur RE, Ruparel K, Calkins ME, Satterthwaite TD, Bilker WB, et al. Within-individual variability in neurocognitive performance: age- and sex-related differences in children and youths from ages 8 to 21. Neuropsychology. 2014; 28(4):506–18.10.1037/neu0000067 [PubMed: 24773417]

41. Roalf DR, Ruparel K, Gur RE, Bilker W, Gerraty R, Elliott MA, et al. Neuroimaging predictors of cognitive performance across a standardized neurocognitive battery. Neuropsychology. 2014; 28(2):161–76.10.1037/neu0000011 [PubMed: 24364396]

42. Strangman GE, Sipes W, Beven G. Human cognitive performance in spaceflight and analogue environments. Aviation, Space, and Environmental Medicine. 2014; 85(10):1033–48.10.3357/asem.3961.2014

43. 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):1–5.10.1016/j.neulet.2009.07.048 [PubMed: 19631255]

44. Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. Health Educ Res. 2006; 21(Suppl 1):i19–32. Epub 2006/08/02. cyl053 [pii]. 10.1093/her/cyl053 [PubMed: 16880221]

45. Zhou X, Ferguson SA, Matthews RW, Sargent C, Darwent D, Kennaway DJ, et al. Mismatch between subjective alertness and objective performance under sleep restriction is greatest during the biological night. Journal of sleep research. 2012; 21(1):40–9. Epub 2011/05/14. 10.1111/j.1365-2869.2011.00924.x [PubMed: 21564364]

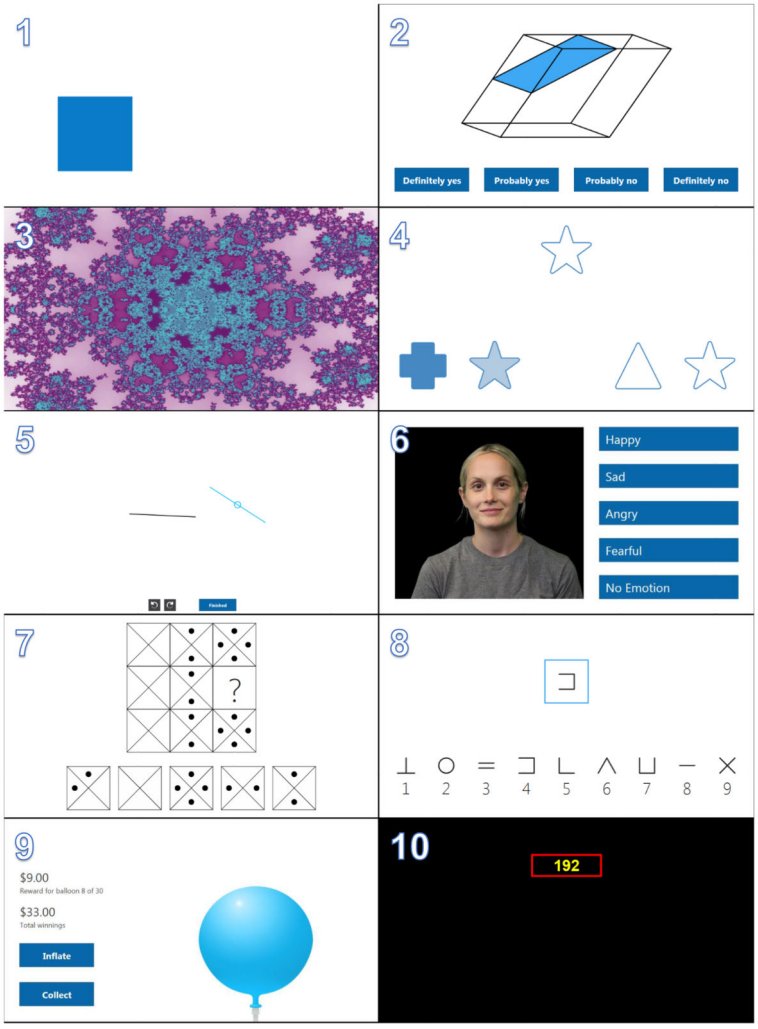

Figure 1.

Screenshots of the 10 individual tests comprising the *Cognition* test battery. The tests are listed in the standard order of administration: 1) Motor Praxis (MP), 2) Visual Object Learning (VOLT); 3) Fractal 2-Back (F2B); 4) Abstract Matching (AM); 5) Line Orientation (LOT); 6) Emotion Recognition (ERT); 7) Matrix Reasoning (MRT); 8) Digit Symbol Substitution (DSST); 9) Balloon Analog Risk (BART); and 10) Psychomotor Vigilance (PVT).

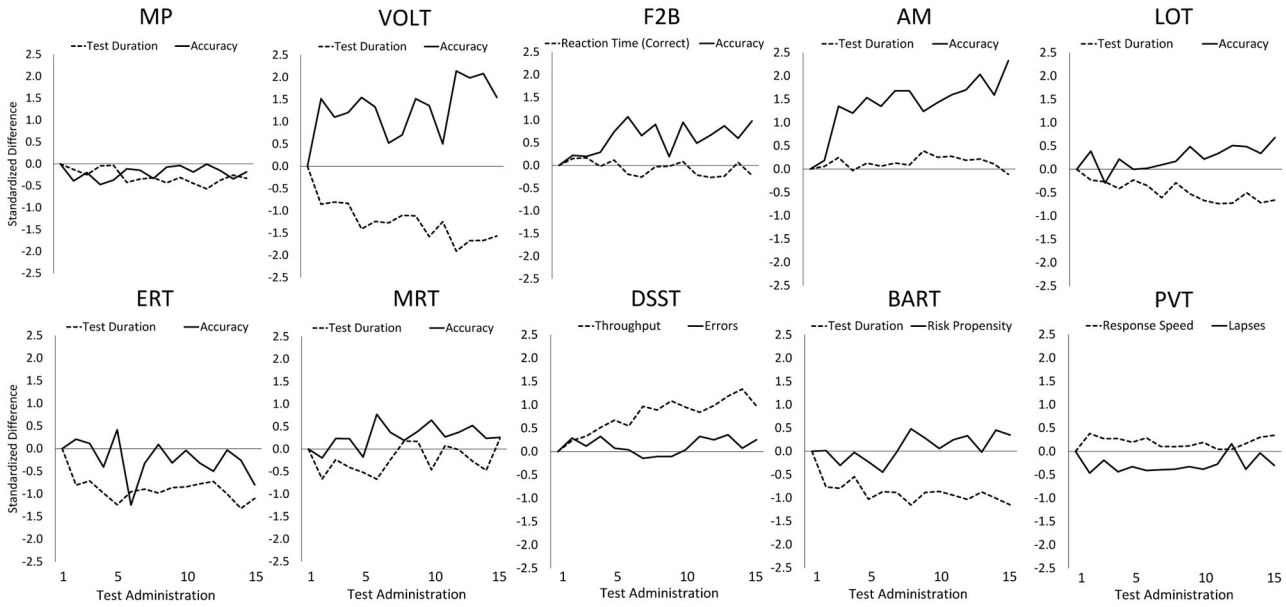

Figure 2.

Changes in performance with repeated administration are shown for key accuracy and speed outcomes for each of the 10 *Cognition* tests. Data were sampled from 8 astronauts and astronaut candidates and 11 mission controllers who performed all 15 unique versions of the battery in the same order with 1–2 week intervals between test administrations. Mean and standard deviation of scores of the first test administration were used for standardization to facilitate comparisons across tests. Motor Praxis (MP); Visual Object Learning (VOLT); Fractal 2-Back (F2B); Abstract Matching (AM); Line Orientation (LOT); Emotion Recognition (ERT); Matrix Reasoning (MRT); Digit Symbol Substitution (DSST); Balloon Analog Risk (BART); Psychomotor Vigilance (PVT)

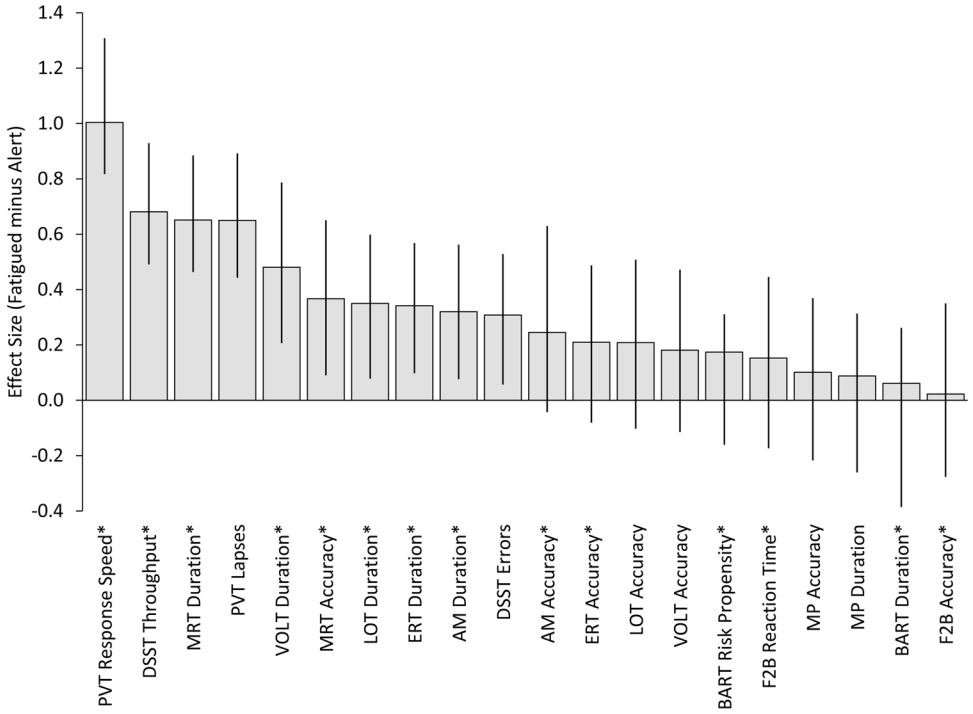

Figure 3.

Effects of one night of acute total sleep deprivation on *Cognition* performance. Error bars represent 95% bootstrap confidence intervals based on 10,000 replications. * indicates that the effect size was multiplied by −1 to facilitate comparisons across variables. This study was approved by the Institutional Review Board of the University of Pennsylvania and subjects signed written informed consent prior to study participation.

Table I

Overview of the Cognition test battery

Test	Cognitive Domains Assessed	Brain Regions Primarily Recruited	Administration Time [Minutes] Median (Range)
Motor Praxis (MP)	Sensory-motor speed	Sensorimotor cortex	0.4 (0.3 – 2.3)
Visual Object Learning (VOLT)	Spatial learning and memory	Medial temporal cortex, hippocampus	1.7 (1.4 – 8.2)
Fractal 2-Back (F2B)	Working memory	Dorsolateral prefrontal cortex, cingulate, hippocampus	2.0 (1.7 – 16.5)
Abstract Matching (AM)	Abstraction, concept formation	Prefrontal cortex	1.8 (1.3 – 7.9)
Line Orientation (LOT)	Spatial orientation	Right temporo-parietal cortex, visual cortex	1.2 (0.8 – 2.4)
Emotion Recognition (ERT)	Emotion identification	Cingulate, amygdala, hippocampus, fusiform face area	1.7 (1.2 – 3.1)
Matrix Reasoning (MRT)	Abstract reasoning	Prefrontal cortex, parietal cortex, temporal cortex	2.1 (0.6 – 3.9)
Digit Symbol Substitution (DSST)	Complex scanning and visual tracking	Temporal cortex, prefrontal cortex, motor cortex	1.6 (1.6 – 2.6)
Balloon Analog Risk (BART)	Risk decision making	Orbital frontal and ventromedial prefrontal cortex, amygdala, hippocampus, anterior cingulate cortex, ventral striatum	2.1 (1.7 – 4.1)
Psychomotor Vigilance (PVT)	Vigilant attention	Prefrontal cortex, motor cortex, inferior parietal and some visual cortex	3.2 (3.1 – 4.5)

Administration times based on N=15 administrations of the Cognition battery in each of N=19 astronauts, astronaut candidates and mission controllers (N=285 total administrations; see text for details). Administration times include the time needed to input comments and any pause taken by the subject before proceeding to the next test.