Validation of the Cognition Test Battery for Spaceflight in a Sample of Highly Educated Adults

Tyler M. Moore; Mathias Basner; Jad Nasrini; Emanuel Hermosillo; Sushila Kabadi; David R. Roalf; Sarah McGuire; Adrian J. Ecker; Kosha Ruparel; Allison M. Port; Chad T. Jackson; David F. Dinges; Ruben C. Gur

BACKGROUND: Neuropsychological changes that may occur due to the environmental and psychological stressors of prolonged spaceflight motivated the development of the Cognition Test Battery. The battery was designed to assess multiple domains of neurocognitive functions linked to specific brain systems. Tests included in Cognition have been validated, but not in high-performing samples comparable to astronauts, which is an essential step toward ensuring their usefulness in long-duration space missions.

METHODS: We administered Cognition (on laptop and iPad) and the WinSCAT, counterbalanced for order and version, in a sample of 96 subjects (50% women; ages 25–56 yr) with at least a Master's degree in science, technology, engineering, or mathematics (STEM). We assessed the associations of age, sex, and administration device with neurocognitive performance, and compared the scores on the Cognition battery with those of WinSCAT. Confirmatory factor analysis compared the structure of the iPad and laptop administration methods using Wald tests.

RESULTS: Age was associated with longer response times (mean $\beta = 0.12$) and less accurate (mean $\beta = -0.12$) performance, women had longer response times on psychomotor ($\beta = 0.62$), emotion recognition ($\beta = 0.30$), and visuo-spatial ($\beta = 0.48$) tasks, men outperformed women on matrix reasoning ($\beta = -0.34$), and performance on an iPad was generally faster (mean $\beta = -0.55$). The WinSCAT appeared heavily loaded with tasks requiring executive control, whereas Cognition assessed a larger variety of neurocognitive domains.

DISCUSSION: Overall results supported the interpretation of Cognition scores as measuring their intended constructs in high performing astronaut analog samples.

KEYWORDS: spaceflight, neurocognition, Cognition Test Battery for Spaceflight, Penn Computerized Neurocognitive Battery, psychometrics, validity.

Moore TM, Basner M, Nasrini J, Hermosillo E, Kabadi S, Roalf DR, McGuire S, Ecker AJ, Ruparel K, Port AM, Jackson CT, Dinges DF, Gur RC. *Validation of the Cognition Test Battery for spaceflight in a sample of highly educated adults.* Aerosp Med Hum Perform. 2017; 88(10):937–946.

The Cognition Test Battery[3] was designed to assess multiple domains of neurocognition that are: 1) linked to specific brain areas and networks, and 2) likely of particular importance in spaceflight. Due to the myriad environmental and psychological stressors and hazards involved in spaceflight—circadian misalignment, confinement, isolation, decompression, dietary restriction, fluid shifts, hypercapnia, hypoxia, increased intracranial pressure, limited physical exercise, microgravity, noise, radiation, restricted sleep, risk of catastrophic technical failures, among others—it is possible that brain integrity is affected and as a result astronauts experience changes in neurocognitive performance ability, almost always for the worse. Astronauts themselves have confirmed as much subjectively, but objective, empirical evidence of such stressors on neurocognition is scarce.[30] Because decreased neurocognitive ability increases the risk of mission failure and physical harm to the crew, valid and reliable assessment of cognitive functioning in spaceflight is of obvious importance. This was the motivation for developing the Cognition Test Battery.

From the Department of Psychiatry, Neuropsychiatry Section, and the Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Perelman School of Medicine, University of Pennsylvania, and the VISN4 Mental Illness Research, Education, and Clinical Center at the Philadelphia VA Medical Center, Philadelphia, PA.

This manuscript was received for review in November 2016. It was accepted for publication in June 2017.

Address correspondence to: Tyler M. Moore, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, 3400 Spruce St., 10th Floor Gates Pavilion, Philadelphia, PA 19104; tymoore@upenn.edu.

DOI: https://doi.org/10.3357/AMHP.4801.2017

The Cognition Test Battery has been administered multiple times on the International Space Station and in various projects sponsored by NASA and the National Space Biomedical Research Institute[3] and consists of well-validated tests (see below). However, the full battery itself has not been validated in its current implementation in a larger group of healthy adults, which would likely provide valid normative data relevant to astronauts and other highly educated subjects in NASA studies of analog conditions. Moreover, there is a need to establish Cognition's factorial structure and sensitivity to age and sex differences in an astronaut-like population—namely, educated, high-performing adults ranging in age from 25 to 56 yr. A focus on such a population is needed to make the data comparable to the high-performing space travelers for whom the battery is intended. As Basner et al.[3] state:

"*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 [p. 943].*"

Validity of Cognition score interpretation can be assessed in various ways and, in this article, we report an assessment of the concurrent validity of Cognition by examining each test score's association with age and sex. There is well-established literature describing the effects of aging on cognition[14,15] and, in our age range of interest, it is clear that we should expect modest decreases in both accuracy and speed, especially the latter. There are also several well-established findings related to sex differences in neurocognition,[14] whereby male subjects tend to outperform female subjects on sensorimotor and psychomotor tasks[5] and visuo-spatial tasks,[33] and female subjects tend to outperform male subjects on emotion recognition tasks[34] and some memory tasks.[29] Additionally, it is important to measure gender differences in risk taking as male subjects tend to engage in more risky behaviors.[6]

A second purpose of this study was to compare performance on Cognition when administered on a laptop compared to an iPad, to ensure that validation data exist on the two technical platforms most likely to be available for cognitive testing in spaceflight and other operations environments. Some speed enhancement is expected from the tablet format, which does not require interface with a control device (e.g., the touchpad used for Cognition laptop administration). However, such effects may vary by test and the magnitude of these effects needs to be established for determining norms.

The third aim of the study was to compare Cognition to the WinSCAT[20] computerized cognitive battery long used by NASA operationally. The WinSCAT (mean administration time ≈ 13.8 min) is somewhat shorter and emphasizes executive control domains, while Cognition (mean administration time ≈ 19.9 min) emphasizes executive functioning, but is designed to cover a broader range of neurobehavioral domains. However, note that mean Cognition administration time for astronauts ≈ 16.5 min.

METHODS

Subjects

The study consisted of 96 subjects (50% female) with at least a Master's degree in science, technology, engineering, or mathematics (STEM). All were from the Philadelphia area, ranged in age from 25 to 56 (mean = 40.3 ± 9.5), and were screened (via self-report questionnaire) for serious medical and psychiatric disorders, such as schizophrenia, past cerebrovascular accident (stroke), epilepsy, and other medical disorders or conditions that can affect performance. This age range was chosen to approximate the age range of the current population of space travelers, with the lower limit extended down to 25 yr with the hope of pre-empting the possibility of younger space travelers in the future. While it is true that some astronauts have flown while outside this age range (e.g., John Glenn was 77 on his final mission), the range of 25−56 yr captures the vast majority of astronauts (mean age = 48, with range = 36−56).[7] This study was approved by the Institutional Review Board of the University of Pennsylvania, and subjects signed written informed consent prior to study participation.

Measures

Cognition Test Battery. The Cognition battery comprises 10 individual neurocognitive tests and has been described in detail elsewhere.[3] Briefly, Cognition contains a subset of tests from a widely used and validated neurocognitive battery, the Penn Computerized Neurocognitive Battery (CNB),[13–15,26] as well as additional tests—i.e., the Psychomotor Vigilance Test[22] and Digit Symbol Substitution Test[32]—that have either been used extensively in spaceflight or assess cognitive domains of particular interest in spaceflight. The CNB is currently being used in assessment of military servicemembers,[25,31] development in children,[11] and genomic research in populations with or at risk for psychiatric disorders.[10] A basic visualization of all the tests described below can be found in supplementary **Fig. A1**, and a summary of how each test score is calculated is in **Table A1** (https://doi.org/10.3357/amhp4801sd.2017). Also, note the key test-administration variables in this study (device, test version, and order) were exactly counterbalanced and person-level variables (sex and age group) were also balanced; i.e., the study design is a perfect "Latin Square." Below, we provide brief descriptions of the tests and more elaborate descriptions can be found in the supplementary materials (**Appendix A**, found online at https://doi.org/10.3357/amhp.4801sd.2017).

The Motor Praxis task. The Motor Praxis task (MP)[13] was administered at the start of testing to ensure that subjects had sufficient command of the computer interface and immediately thereafter as a measure of sensorimotor speed. Subjects were instructed to click on squares that appeared randomly on the screen, with each successive square being smaller and thus more difficult to track.

The Visual Object Learning Test. The Visual Object Learning Test (VOLT) assessed subject memory for complex figures.[9]

Subjects were asked to memorize 10 sequentially displayed three-dimensional figures. Later, they were instructed to select those objects they memorized from a set of 20 such objects also sequentially presented, half from the learning set and half new.

The Fractal 2-Back. The Fractal 2-back (F2B, or NBACK)[27] is a nonverbal variant of the standard Letter 2-Back, which is currently included in the core CNB. The Fractal NBACK consists of the sequential presentation of a set of figures (fractals), each potentially repeated multiple times. Subjects had to respond when the current stimulus matched the stimulus displayed two figures ago.

Abstract Matching. The Abstract Matching (AM) test[8] is a measure of the abstraction and flexibility components of executive function, including an ability to discern general rules from specific instances. Validity of the AM has been established mostly from its ability to distinguish patients with schizophrenia from healthy controls. The test paradigm presented subjects with two pairs of objects at the bottom left and right of the screen, varied on specific perceptual dimensions (i.e., shape and fill). 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 Line Orientation Test. The Line Orientation Test (LOT) is a measure of spatial orientation derived from the well-validated Judgment of Line Orientation Test,[4] the computerized version of which was among the first to be administered with functional neuroimaging[12] and is used in the core CNB. The LOT format consists of presenting two lines at a time, one stationary and the other can be rotated by clicking an arrow. Subjects rotated the movable line until it was parallel to the stationary line.

The Emotion Recognition Task. The Emotion Recognition Task (ERT) is a measure of visual emotion recognition that was developed[16] and validated with neuroimaging[23] and is part of the Penn CNB. The ERT presented subjects with photographs of professional actors (adults of varying age and ethnicity) portraying emotional facial expressions of varying intensities. 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 Matrix Reasoning Test. The Matrix Reasoning Test (MRT) is a measure of abstract reasoning and consists of increasingly difficult pattern matching tasks.[13] It is analogous to Raven's Progressive Matrices[28] and consists of a series of patterns overlaid on a grid. One element from the grid is missing and the subject had to select the element that fit the pattern from a set of alternative options.

The Digit-Symbol Substitution Task. The Digital-Symbol Substitution Task (DSST)[32] is a computerized adaptation of a paradigm used in the Wechsler Adult Intelligence Scale (WAIS-III). The DSST required the subject 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 subject had to select the corresponding number 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. The Balloon Analog Risk Test (BART) is a validated assessment of risk taking behavior[21] and requires subjects to either inflate an animated balloon or collect a reward. Subjects were 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.

The Psychomotor Vigilance Test. The Psychomotor Vigilance Test (PVT) measures vigilant attention by recording reaction times (RT) to visual stimuli that occur at random interstimulus intervals.[2] 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 1 s. Subjects are instructed to be as fast as possible without hitting the spacebar in absence of a stimulus (i.e., false starts or errors of commission).

Spaceflight Cognitive Assessment Tool for Windows. The Spaceflight Cognitive Assessment Tool for Windows (WinSCAT) is administered with a Windows laptop, and the left and right mouse button are the primary input method. The WinSCAT[20] comprises a subset of five tests from the Automated Neuropsychological Assessment Metrics battery.[18] They are as follows:

The Code Substitution Test (Codesub) is a measure of visual scanning. The subject was shown a number-symbol pair and asked to determine if that pair matches any of the pairs in a key that is presented on the same screen (and remained throughout the test). There were 72 trials and the main outcome measures are accuracy (percent correct) and mean response time for all responses. This test is very similar to Cognition's DSST test.

The Running Memory Continuous Performance Test (CPT) is a measure of working memory and attention. It uses the same paradigm as the NBACK described above used by Cognition, but with two differences: 1) the stimuli are numbers rather than fractals, and 2) the subject must respond when the present stimulus is the same as the one immediately before rather than two before. There are 180 trials and the main outcome measures are accuracy (percent correct, including both true positives and true negatives) and mean response time for all responses.

Mathematical Processing (Math) is a measure of computational processing and mathematical achievement. Subjects were shown a three-term math problem (e.g., $1 + 5 - 4 = ?$) and had to decide whether the missing expression was greater than or less than 5. There were 20 trials and the main outcome measures were accuracy (percent correct) and mean response time for all responses.

Delayed Matching to Sample (M2S) is a measure of visual memory. Subjects were shown a 4×4 grid comprising squares of two different colors. Five seconds later, they were shown two

such grids and decided which one matched the one they were initially shown. There were 20 trials, and the main outcome measures were accuracy (percent correct) and mean response time for all responses.

The Delayed Recognition of Code Substitution Test (DR) is a measure of short-term memory. Subjects were shown a number-symbol pair as in the Codesub task above, but without the presence of a key. They had to decide whether the pair matched any of the pairs in the key they were shown in the Codesub task. There were 36 trials and the main outcome measures were accuracy (percent correct) and mean response time for all responses.

Analyses

We performed linear regressions predicting each test's scores (accuracy and RT) with age, age-squared (to account for nonlinearity), sex, the order in which the battery was taken, and, for the Cognition battery, the device on which the test was taken (laptop vs. iPad). To further explore and visualize the age results, subjects were split into five age categories and, for each significant linear effect, mean performance was plotted by these categories. This procedure was also done for the overall accuracy and speed scores (combining all tests on each battery). To further explore these effects, efficiency scores were calculated by taking the mean of the accuracy and speed z-scores, where speed = RT * (−1) so that a higher score means faster performance. The reason for calculating efficiency is that optimal neurocognitive performance is characterized not only by the ability to be accurate, but to do so quickly. Thus, for example, if a person's speed score was 1 SD below the mean and accuracy score was 1 SD above the mean, he/she would receive an efficiency score of zero (average). These scores were then plotted by sex and differences were tested for statistical significance using *t*-tests. Finally, to probe for interaction effects in the regressions described above, they were also run including all two-way interaction terms.

To examine how the Cognition and WinSCAT batteries compare and contrast in what they measure, we estimated the accuracy score on each of the tests using the entire opposite battery and compared the mean R^2 (and adjusted R^2) values for each. That is, WinSCAT DR accuracy was predicted using all 10 Cognition tests and the R^2 recorded, then Codesub was predicted using all Cognition tests and R^2 recorded, etc., and the mean of these five R^2 values was taken to represent how well Cognition scores can predict WinSCAT scores. Then the same was done to predict each of the 10 Cognition scores with the full (5-test) WinSCAT and this mean of 10 R^2 values was taken to represent how well WinSCAT scores can predict Cognition scores. All the above was done 500 times with random bootstrapped subsamples to obtain 95% confidence intervals around the two means for the purpose of testing the statistical significance of the difference. That is, all the above analyses were performed on a random subsample of 48 out of the 96 subjects and the R^2 values recorded. Then, a different random sample of 48 was selected and the analyses run using this new random sample (and R^2 values again recorded). This was repeated 500 times to obtain a distribution (and therefore confidence interval) of the R^2 estimates.

To investigate the consistency of factor structures across the two Cognition devices (laptop and iPad), we performed confirmatory factor analysis (CFA) on the efficiency scores for each device separately and compared the factor loadings using Wald tests in Mplus. Wald tests were performed both at the omnibus level (a single test of differences across all coefficients) and at the level of individual loadings (one Wald test per loading). In addition to Wald tests, we calculated a Chi-squared difference test between two models: 1) all loadings in the laptop and iPad models were freely estimated, and 2) loadings on the laptop and iPad were constrained to equality using the Model Constraint command in Mplus. This tests whether allowing loadings to be freely estimated across administration methods significantly increases model fit.

The number of factors to extract was based on theory, such that the tests designed to measure executive functioning (AM, DSST, NBACK, and BART) were made to load on one factor, and tests designed to measure memory and more complex reasoning (VOLT, ERT, LOT, and MRT) were made to load on a second factor. The extraction of two factors was further supported by parallel analysis.[17] Because the assignment of tests to factors was based on theory, no exploratory factor analysis was performed before the CFA.

Finally, we examined test-retest reliability of Cognition efficiency scores. Specifically, administration device, order, and test version were regressed out of the efficiency scores, and the correlations were estimated between first and second administrations.

RESULTS

Multiple Regressions and *t*-Tests

Table I shows the results of the regression analyses predicting the 19 Cognition accuracy and RT scores and the 10 WinSCAT accuracy and RT scores. Without exception, all significant effects of age (MRT, NBACK, and VOLT) on Cognition accuracy are negative, indicating that older subjects performed worse. Likewise, for significant Cognition RT scores (DSST, ERT, VOLT, MP, and PVT), all are positive, indicating that older subjects performed not only less accurately but also more slowly (higher RT). For WinSCAT accuracy, only DR showed a significantly negative association with age, though three tests (DR, Codesub, and M2S) showed a significant positive association between age and RT. Note that the coefficients for age in Table I are linear main effects, where age-squared was included in the model but not shown. Because the linear term was standardized before squaring, the linear and nonlinear terms were orthogonal. The linear effect can, therefore, be interpreted as any other linear effect, and if a nonlinear effect is indicated in Table I, the reader should note that there is also some curvature to the relationship. To elucidate this, supplementary **Fig. A2** and **Fig. A3** (available online at https://doi.org/10.3357/amhp4801sd.2017) show the above significant effects graphically with age split into

Table I. Regression Results Describing the Associations Among Neurocognitive Performance Scores and Sex, Age, and Administration Device.

		AGE		SEX*		DEVICE**			
	SCORE	B	SIG.	B	SIG.	B	SIG.	UB	UB RAW
Cognition	AM Accuracy	−0.13†	0.06	0.02	0.88	−0.01	0.97	−0.001	0.002
	BART Risk	−0.04	0.57	−0.09	0.52	0.01	0.95	0.003	0.927
	DSST Accuracy	0.02†	0.83	0.04	0.78	**0.69**	**<0.005**	0.013	0.013
	ERT Accuracy	−0.06	0.40	0.27	0.06	**−0.44**	**<0.005**	−0.083	−0.018
	LOT Accuracy	−0.12	0.11	−0.21	0.14	−0.14	0.31	−0.015	−0.027
	MRT Accuracy	**−0.24**†	**<0.005**	**−0.34**	**0.01**	0.06	0.62	0.017	−0.007
	NBACK Accuracy	**−0.15**	**0.04**	−0.17	0.24	0.11	0.44	0.018	0.004
	VOLT Accuracy	**−0.23**	**<0.005**	−0.08	0.59	−0.02	0.89	−0.004	−0.007
	PVT Accuracy	−0.11†	0.12	−0.11	0.45	−0.27	0.06	−0.021	0.688
	AM RT	0.09	0.21	0.14	0.34	**−0.37**	**0.01**	−0.419	
	BART RT	−0.03	0.69	**0.40**	**<0.005**	**−0.45**	**<0.005**	−0.430	
	DSST RT	**0.26**†	**<0.005**	0.13	0.14	**−1.49**	**<0.005**	−0.575	
	ERT RT	**0.17**	**0.01**	**0.30**	**0.02**	**−0.65**	**<0.005**	−0.570	
	LOT RT	0.04	0.55	**0.48**	**<0.005**	**−0.60**	**<0.005**	−1.632	
	MRT RT	0.02†	0.80	−0.06	0.64	−0.13	0.34	−0.381	
	NBACK RT‡	0.10	0.16	0.19	0.20	0.21	0.16	0.025	
	VOLT RT	**0.16**†	**0.02**	0.07	0.61	**−0.33**	**0.02**	−0.202	
	MPT RT	**0.18**	**<0.005**	**0.13**	**0.01**	**−1.84**	**<0.005**	−0.571	
	PVT RT‡	**0.16**	**0.02**	**0.62**	**<0.005**	0.20	0.15	0.091	
WinSCAT	DR Mem Accuracy	**−0.24**	**0.02**	**−0.47**	**0.02**				
	Codesub Accuracy	−0.11†	0.29	−0.08	0.67				
	CPT Accuracy	−0.17	0.10	−0.33	0.10				
	M2S Accuracy	−0.17	0.09	**−0.41**	**0.04**				
	Math Accuracy	−0.16	0.13	−0.15	0.45				
	DR Mem RT	**0.34**	**<0.005**	0.06	0.76				
	Codesub RT	**0.37**	**<0.005**	**0.62**	**<0.005**				
	CPT RT	0.15	0.14	**0.42**	**0.04**				
	M2S RT	**0.31**	**<0.005**	0.32	0.10				
	Math RT	0.07	0.48	**0.54**	**0.01**				

* Reference group = male; **reference device = PC laptop; ‡these tests use the space bar on the laptop as the input method; †indicates a significant nonlinear effect of age was found, though the reported coefficient above is for the linear effect.

B = standardized beta; UB = unstandardized beta; AM = Abstract Matching; BART = Balloon Analog Risk Task; DSST = Digit Symbol Substitution Task; ERT = Emotion Recognition Task; LOT = Line Orientation Task; MRT = Matrix Reasoning Task; VOLT = Visual Object Learning Test; DR = Delayed Recognition; Codesub = Code Substitution; CPT = Continuous Performance Test; M2S = Match to Sample; Acc = Accuracy; Neut = Neutral; RT = response time; Sig = Significance Level (*P*-value).

Significant effects are bolded; the following were included in the model as covariates but not shown: test version (form), order of administration, and age-squared (to account for nonlinearity).

five categories, allowing better grasp of the nature of the associations. Note that RT has been converted to speed by z-transforming and multiplying by −1 so as to be consistent with accuracy in that a higher score means better performance. For Cognition accuracy (Fig. A2), the MRT shows a significant nonlinear relationship, whereas for Cognition speed (also Fig. A2), the DSST and VOLT show significant nonlinear relationships. For WinSCAT accuracy (Fig. A3), Codesub shows a significant nonlinear association; for WinSCAT speed (also Fig. A3), no nonlinear effect was found. When accuracy and speed scores are combined for an efficiency score (by taking the mean of the accuracy and speed z-scores, where speed is keyed such that higher = faster) and overall battery scores are considered (**Fig. 1**), the Cognition and WinSCAT batteries show age-related decreases in both accuracy and speed, with a notable apparent nonlinear effect whereby efficiency drops precipitously after age 43.

Note that in Fig. 1, because WinSCAT was administered only on the laptop, Cognition scores were limited to the laptop administration for comparison. All negative linear relationships between continuous age and overall efficiency are significant at the $P < 0.001$ level for both Cognition (both devices, though iPad not shown in Fig. 1) and WinSCAT.

For associations with sex, the only Cognition test that showed significant effects for accuracy was the MRT (men > women; $P = 0.01$). For RT, the BART, ERT, LOT, MP, and PVT all showed significant effects, with men performing faster than women. For WinSCAT accuracy, men significantly outperformed women on the DR and M2S, and for RT, men performed faster on the Codesub, CPT, and Math.

Fig. 2 shows the results using accuracy and speed scores combined to form efficiency scores. For Cognition, the BART, LOT and PVT all show significant effects, with men outperforming women. For WinSCAT, four of the five tests (Codesub, CPT, M2S, and Math) showed significantly better performance in men.

For the effect of device (laptop vs. iPad), subjects performed significantly more accurately on the DSST when using the iPad and significantly less accurately on the ERT when using the iPad. For RT, subjects performed significantly faster on the iPad on seven tests: AM, BART, DSST, ERT, LOT, VOLT, and MP. To provide a convenient translation of scores between devices, the

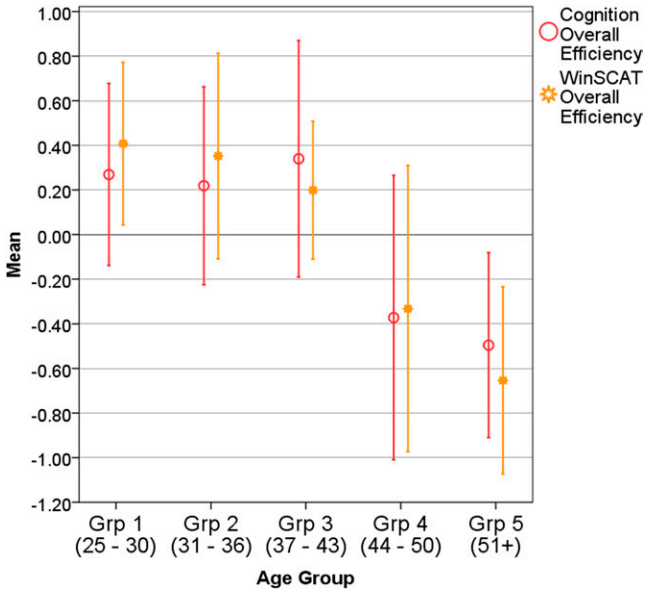

Fig. 1. Relationship between age and overall efficiency on the Cognition (laptop version) and WinSCAT test batteries. Grp = Group. Lines with white circles (red in the online article) is Cognition overall efficiency and lines with asterisks (orange in the online article) are WinSCAT overall efficiency.

second-to-rightmost column in Table I (labeled "UB") shows the unstandardized beta coefficients for the effect of device. Because the reference device is a laptop, the coefficients show how much should be added or subtracted from a score on the laptop to estimate how well the person would have done if he/she had taken the test on the iPad. For example, the coefficient for VOLT RT is −0.202 s (or −202 ms), so if someone had a mean response time of 1.000 s on the laptop, that score could be translated to a score of 1.000 – 0.202 = 0.798 s on the iPad. Note that, for accuracy, the unstandardized betas are proportions, not percentages (e.g., a beta of 0.002 translates to 0.2%). The WinSCAT was administered only on a laptop, so the Device variable does not apply. In addition, the rightmost column of Table I ("UB Raw") shows the unstandardized beta coefficients when basic raw scores are used. These are proportion correct (0 to 1) for the VOLT, NBACK, AM, LOT, ERT, MRT, and DSST; total number of pumps for the BART; and total number of false starts and lapses for the PVT. Note that the coefficient for the PVT in the rightmost column has the opposite sign from its other Device accuracy coefficients because its raw score (False Starts + Lapses) is scored such that higher = worse. Finally, supplementary **Table A2** (found online at https://doi.org/10.3357/amhp.4801sd.2017) gives a set of Cognition raw score norms (and SDs) by age, sex, and device for the first administration (therefore no practice effects). So, for example, if a 25-yr-old man took the VOLT on an iPad and scored 0.40 for accuracy, he would be performing substantially worse than would be expected given his age, sex, and the administration device (expected score = 0.62).

For the significant effects of order (not shown in Table I), all were in the expected direction for both Cognition and WinSCAT—namely, on the second administration, subjects tended to be both faster and more accurate. Though some effects of order were nonsignificant, all significant effects were in this direction.

Cross-Battery Prediction

Fig. 3 shows the R^2 values for each test being predicted by all tests of the opposite battery simultaneously. The solid (blue in the online article) bars indicate how well WinSCAT could predict the scores on each of the Cognition tests, and the checkered (green in the online article) bars indicate how well Cognition could predict each WinSCAT test. For prediction of Cognition by WinSCAT, the best-predicted was the NBACK (R^2 = 0.25), followed by the MRT (R^2 = 0.19). The least well-predicted scores from Cognition were the BART and ERT (R^2 = 0.05 and 0.03, respectively). For prediction of WinSCAT by Cognition, Codesub and CPT were equally well-predicted (R^2 = 0.26 for both), followed by M2S (R^2 = 0.24). The least well-predicted were DR and Math (R^2 = 0.15 and 0.14, respectively). The upper (green in the online article) horizontal line represents the mean of the five predictions of WinSCAT by Cognition, and the lower (blue in the online article) line represents the mean of the 10 predictions of Cognition by WinSCAT. The means have nonoverlapping bootstrapped confidence intervals (not shown), suggesting that the difference between them (0.21 − 0.13 = 0.08) is statistically significant ($P < 0.005$). When the above analyses were performed on the adjusted R^2, which penalizes Cognition for using more predictors when predicting WinSCAT, all R^2 values decrease, but the difference between the averages (0.12 − 0.08 = 0.04) remains significant ($P < 0.01$).

Factor Analysis Comparing Laptop and iPad Administrations

Fig. 4 shows the CFA results for the Cognition Test Battery administered on the laptop and iPad. Because the laptop and iPad versions were analyzed jointly to be able to compare coefficients using Wald tests, a single set of fit statistics was generated to describe both models. These suggest acceptable fit, with a $\chi^2(59) = 57.8$ ($P = 0.52$), comparative fit index = 1.00, Tucker-Lewis index = 1.00, standardized root mean square residual = 0.055, and root mean square error of approximation = 0.000 ± 0.060. Results of the Wald tests indicate that none of the coefficients differ significantly between administration devices, either at the omnibus level or at the level of the individual loadings. Results of the Chi-squared difference test between constrained and unconstrained models support the same conclusion ($\Delta\chi^2 = 3.4$; $\Delta df = 10$; $P = 0.97$), suggesting no significant improvement in fit when loadings are freely estimated vs. constrained.

Test-Retest Reliability

Table II shows the correlations between efficiency scores on the first and second administrations of Cognition, after regressing out order, test version, and administration device. Mean correlation is 0.50, with the two lowest for the DSST and ERT (0.31 and 0.34, respectively) and two highest for the AM and BART (0.69 and 0.68, respectively).

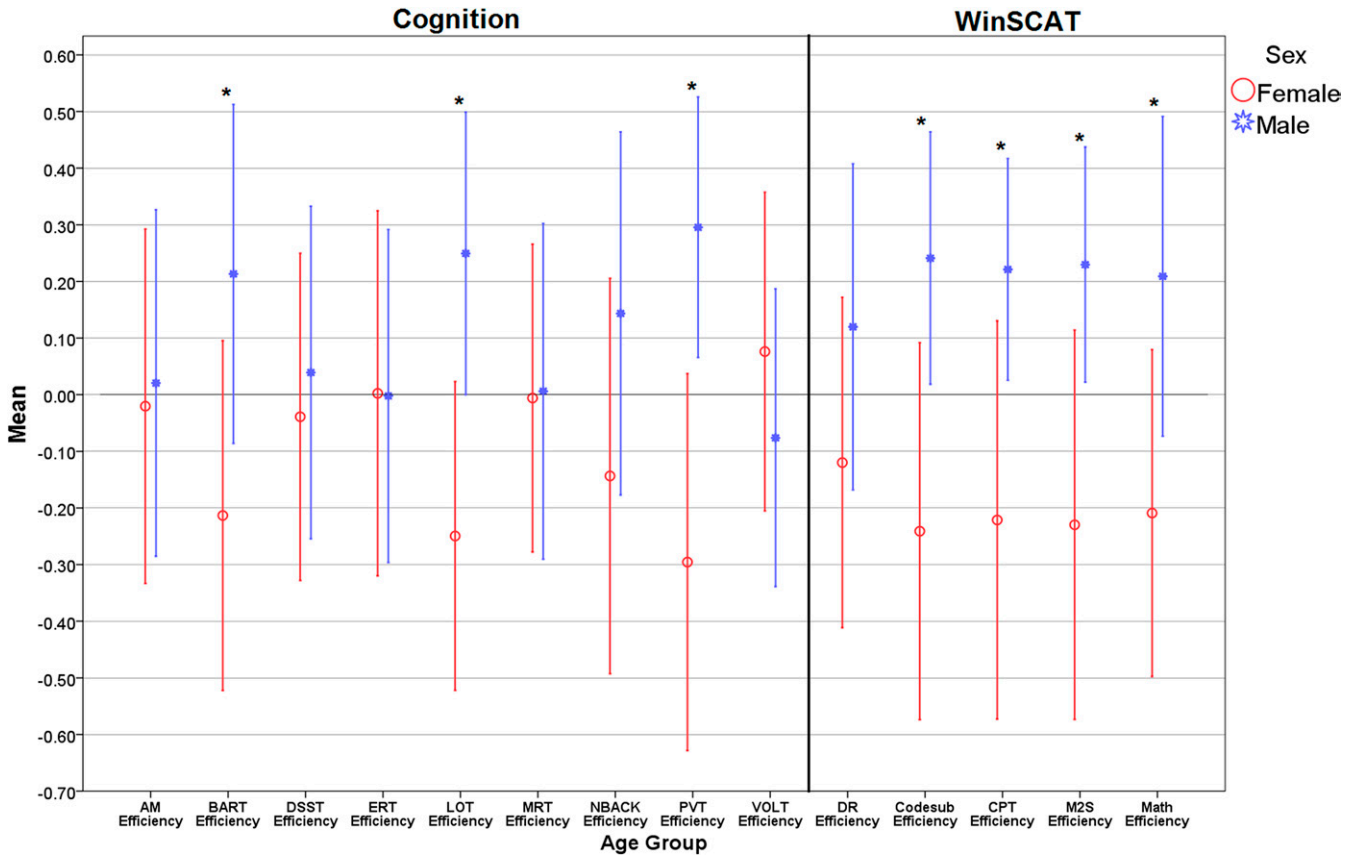

Fig. 2. Neurocognitive performance efficiency on the Cognition and WinSCAT test batteries by sex. AM = Abstract Matching; BART = Balloon Analog Risk Task; DSST = Digit Symbol Substitution Task; ERT = Emotion Recognition Task; LOT = Line Orientation Task; MRT = Matrix Reasoning Task; VOLT = Visual Object Learning Test; DR = Delayed Recognition; Codesub = Code Substitution; CPT = Continuous Performance Test; M2S = Match to Sample. Lines with white circles (red in the online article) denote women and lines with asterisks (blue in the online article) denote men.

DISCUSSION

This study in a highly educated sample analogous to astronauts and other subjects in space analog conditions attempted to establish whether Cognition scores in such a sample are sensitive to age effects and sex differences reported in normative samples of average educational levels. Most results support the validity of the interpretation of Cognition test scores as measuring their intended target neuropsychological phenomena in the present sample. All significant correlations with age were in the predicted direction of slower and less accurate performance with increasing age. The associations with sex were also almost all in the expected direction; i.e., men were faster on the ERT, LOT, MPT, and PVT, more accurate on the MRT, and showed faster risk-taking on the BART. When accuracy and speed were combined into efficiency, men showed more efficient performance on the BART, LOT, and PVT. Three sex difference findings (or lack thereof) were surprising. First, we expected women to outperform men on both ERT accuracy and speed (i.e., women faster than men). However, while women were more accurate on the ERT, the effect was marginal ($P = 0.03$, one-tailed) and they performed more slowly, resulting in no significant sex difference when examining accuracy and speed combined (efficiency). Perhaps women who enter STEM fields are different in some ways from the women tested in previous studies (from community samples); i.e., women self-select into STEM, and it might therefore be unwise to assume they are like most women. Another possibility is that STEM education itself influences neurocognitive abilities, perhaps in the direction of poorer emotion recognition. A second surprising sex difference was that men performed faster not only on the predicted tests (MPT and PVT), but on all tests except MRT. This effect is consistent with evidence that men perform faster not only on reaction time tests, but also on choice (response time) tests.[1] Third, based on the results of Gur et al.,[14]

Table II. Correlations between First and Second Administrations of the Cognition Test Battery, Controlling for Order, Test Version, Age, and Administration Device.

TEST	CORRELATION
AM Efficiency	0.69
BART Efficiency	0.68
DSST Efficiency	0.31
ERT Efficiency	0.34
LOT Efficiency	0.50
MRT Efficiency	0.37
NBCK Efficiency	0.51
PVT Efficiency	0.61
VOLT Efficiency	0.49

All correlations are significant at the $P < 0.05$ level.

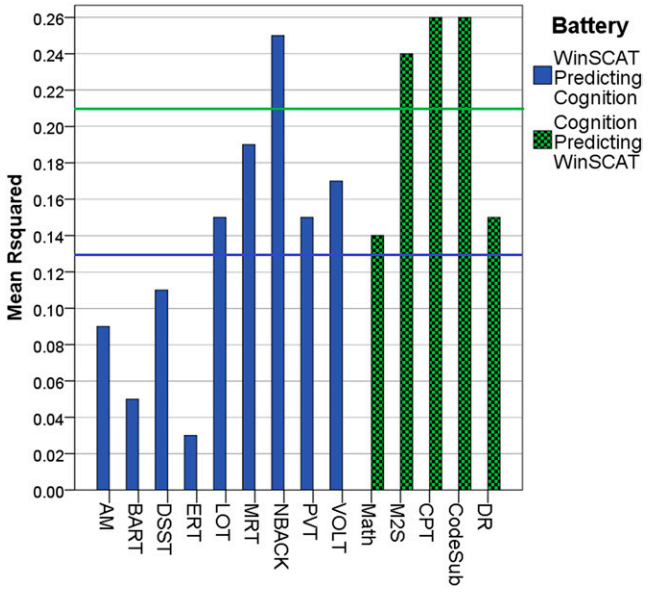

Fig. 3. R^2 values for the prediction of each of the Cognition and WinSCAT tests using all tests from the opposite battery. AM = Abstract Matching; BART = Balloon Analog Risk Task; DSST = Digit Symbol Substitution Task; ERT = Emotion Recognition Task; LOT = Line Orientation Task; MRT = Matrix Reasoning Task; VOLT = Visual Object Learning Test; DR = Delayed Recognition; Codesub = Code Substitution; CPT = Continuous Performance Test; M2S = Match to Sample. Solid lines (blue in the online article) are WinSCAT prediciting Cognition while the checkered lines (green in the online article) are Cognition predicting WinSCAT.

we did not expect to see a sex difference on MRT accuracy, but men significantly outperformed women in the present study. This effect is consistent with some evidence that men tend to score higher on "g" measures than women,[19] and the MRT is known to have a high loading on "g." Further, the absence of the effect in Gur et al.[14] is corroborated by the fact that the sample was fairly young (8–21 yr), and sex differences in "g" do not seem to appear until adulthood.[24] Also, note that when accuracy and speed were combined to form efficiency, men no longer performed better. Overall, most age- and sex-difference results from the present study were expected and support the validity of Cognition score interpretation.

Results of the prediction of each battery's tests with the complete opposite battery revealed a notable bias in WinSCAT toward executive function. This is perhaps not surprising given that the WinSCAT does not have analogs to Cognition's tests of mental flexibility (AM), risk-taking (BART), emotion recognition (ERT), visuo-spatial processing (LOT), or complex reasoning (MRT). Further, although Cognition and WinSCAT do both have tests of memory (VOLT and DR, respectively), each battery could predict the other's memory test only moderately well, suggesting that the neurocognitive phenomena driving performance on the two memory tests only show some overlap. Because neurocognitive demands during spaceflight are likely to be diverse, especially on extended missions with multiple crewmembers, the above findings suggest that WinSCAT is likely too narrow a measure of neurocognitive status.

It is noteworthy that, although many tests of mental ability are designed to measure a single construct (e.g., IQ) from multiple angles, that is not the purpose of the Cognition battery. Instead, the Cognition battery was designed to measure multiple, varied neurocognitive phenomena such that each test in the battery is uniquely important in itself rather than being one more probe of a single, battery-wide construct. This intended multidimensionality is why we have not calculated measures of unidimensionality or internal consistency at the level of the battery. While internal consistency within a single test is desirable, high internal consistency at the level of the battery would be undesirable, because it would suggest that its measurement target is too narrow. While some of the intertest correlations are undoubtedly caused by a single overarching "performance" factor, as suggested by the strong interfactor correlations in Fig. 4, examination of the individual intertest correlations reveals wide variation in their magnitudes, with some even being negative. This issue is of overall importance because it relates to the discussion above about the possible over-emphasis of WinSCAT on executive function while measuring other neurocognitive phenomena only tangentially, if at all.

Two limitations of the present study are worth noting. First, some of the Cognition tests showed only poor-to-moderate test-retest reliability. However, this could be due to the complicating factors of having two test forms, two devices, and confirmed practice effects. Although our models adjusted for these effects, it is unlikely that they could fully account for them due to the limited sample size. Additionally, given that the sample was limited to highly educated, high-performing individuals, the narrow range of ability likely deflated test-retest reliability further. To demonstrate this phenomenon, using the mirt package in R, we simulated response patterns from two groups of hypothetical examinees (one high-ability-only and one comprising a full range of abilities) and estimated test-retest reliability. Using identical tests for the two simulated samples, estimates of test-retest reliability are consistently higher when examinees come from the full ability range. The R script used to run this simulation is available in the **Appendix B** (available online at https://doi.org/10.3357/amhp.4801sd.2017). A second limitation is that, although we did estimate practice effects (not shown) for two administrations of Cognition, this does not allow us to extrapolate to further administrations. Because Cognition was designed to be administered up to 15 times during spaceflight and analog spaceflight missions, both of the above limitations will be addressed in future research.

In summary, the present study in highly educated individuals found that Cognition provides reliable measures of performance both when administered on a laptop and on an iPad, which are sensitive to age cohort effects and to sex differences. When comparing Cognition to WinSCAT, we found that WinSCAT scores can be well predicted from Cognition performance while WinSCAT can predict only executive performance on Cognition.

ACKNOWLEDGMENTS

This research was supported by the National Space Biomedical Research Institute (NSBRI) through NASA NCC 9-58; by NASA through grants NNX14AM81G,

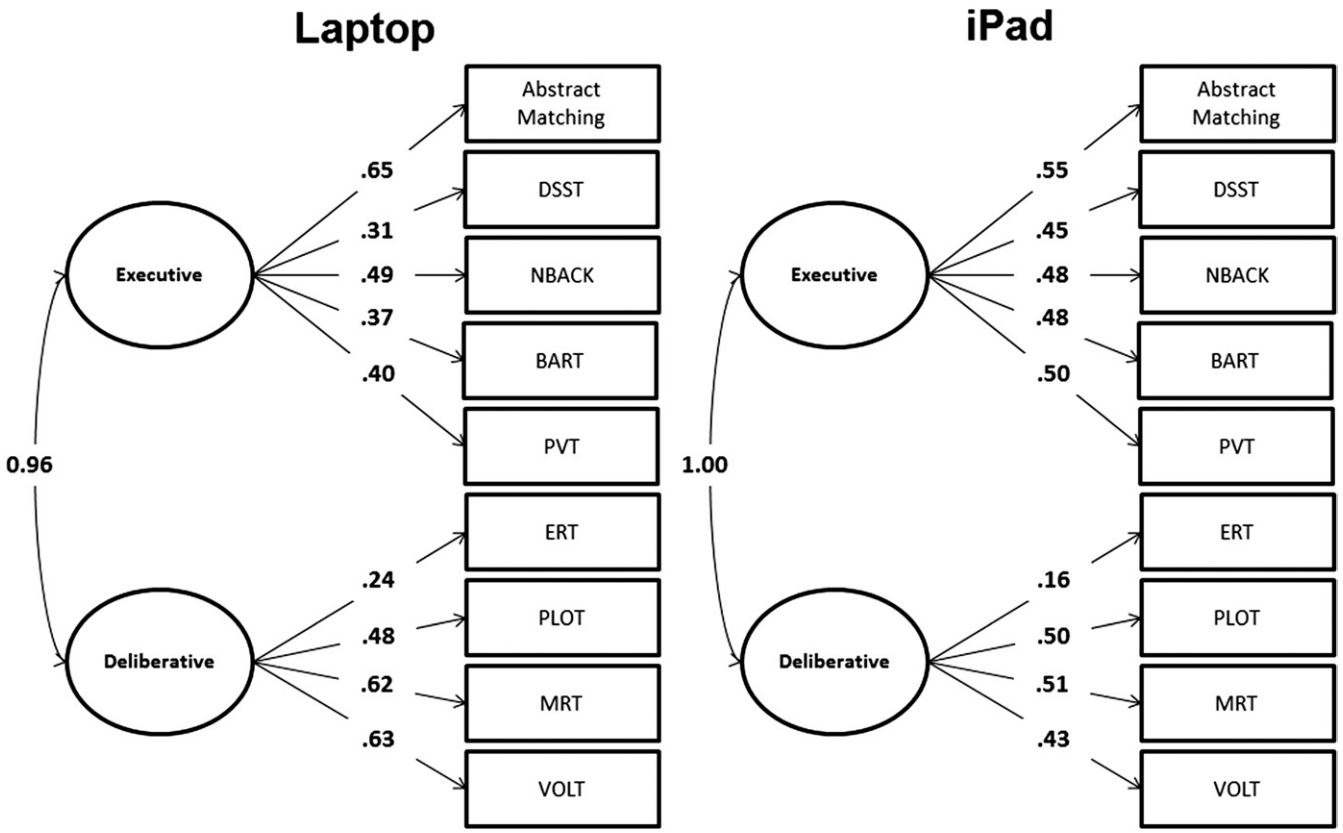

Fig. 4. Joint confirmatory factor analysis of the laptop and iPad versions of the Cognition Test Battery. AM = Abstract Matching; BART = Balloon Analog Risk Task; DSST = Digit Symbol Substitution Task; ERT = Emotion Recognition Task; PLOT = Penn Line Orientation Task; MRT = Matrix Reasoning Task; VOLT = Visual Object Learning Test.

NNX14AH27G, and NNX14AH98G; by NIMH through grants MH089983, MH019112, MH096891, and MH042228; and by the Dowshen Program for Neuroscience.

Authors and affiliations: Tyler M. Moore, Ph.D., M.Sc., Sushila Kabadi, B.A., David R. Roalf, Ph.D., Kosha Ruparel, M.S.E., Allison M. Port, B.A., Chad T. Jackson, M.S.E., and Ruben C. Gur, Ph.D., Department of Psychiatry, Neuropsychiatry Section, and Mathias Basner, M.D., Ph.D., Jad Nasrini, B.A., Emanuel Hermosillo, B.A., Sarah McGuire, Ph.D., Adrian J. Ecker, B.A., and David F. Dinges, Ph.D., Department of Psychiatry, Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, and Ruben C. Gur, Ph.D., VISN4 Mental Illness Research, Education, and Clinical Center at the Philadelphia VA Medical Center, Philadelphia, PA.

REFERENCES

1. Adam JJ, Paas FG, Buekers MJ, Wuyts IJ, Spijkers WA, Wallmeyer P. Gender differences in choice reaction time: evidence for differential strategies. Ergonomics. 1999; 42(2):327–335.
2. Basner M, Dinges DF. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. Sleep. 2011; 34(5):581–591.
3. Basner M, Savitt A, Moore TM, Port AM, McGuire S, et al. Development and validation of the Cognition Test Battery for spaceflight. Aerosp Med Hum Perform. 2015; 86(11):942–952.
4. Benton AL, Varney NR, Hamsher KD. Visuospatial judgment. A clinical test. Arch Neurol. 1978; 35(6):364–367.
5. Blatter K, Graw P, Münch M, Knoblauch V, Wirz-Justice A, Cajochen C. Gender and age differences in psychomotor vigilance performance under differential sleep pressure conditions. Behav Brain Res. 2006; 168(2): 312–317.
6. Byrnes JP, Miller DC, Schafer WD. Gender differences in risk taking: a meta-analysis. Psychol Bull. 1999;125(3):367–383.
7. Dinges DF, Basner M, Mollicone DJ. Reaction SelfTest on ISS: 6-month missions. Houston (TX): NASA; 2016. Final report for NASA project NNX08AY09G.
8. Glahn DC, Cannon TD, Gur RE, Ragland JD, Gur RC. Working memory constrains abstraction in schizophrenia. Biol Psychiatry. 2000; 47(1):34–42.
9. 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):602–612.
10. Greenwood TA, Badner JA, Byerley W, Keck PE, McElroy SL, et al. Heritability and linkage analysis of personality in bipolar disorder. J Affect Disord. 2013; 151(2):748–755.
11. Gur RC, Calkins ME, Satterthwaite TD, Ruparel K, Bilker WB, et al. Neurocognitive growth charting in psychosis spectrum youths. JAMA Psychiatry. 2014; 71(4):366–374.
12. Gur RC, Gur RE, Obrist WD, Hungerbuhler JP, Younkin D, et al. Sex and handedness differences in cerebral blood flow during rest and cognitive activity. Science. 1982; 217(4560):659–661.
13. 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):766–776.
14. Gur RC, Richard J, Calkins ME, Chiavacci R, Hansen JA, et al. Age group and sex differences in performance on a computerized neurocognitive battery in children age 8–21. Neuropsychology. 2012; 26(2):251–265.
15. Gur RC, Richard J, Hughett P, Calkins ME, Macy L, et al. A cognitive neuroscience-based computerized battery for efficient measurement of

individual differences: standardization and initial construct validation. J Neurosci Methods. 2010; 187(2):254–262.
16. Gur RC, Sara R, Hagendoorn M, Marom O, Hughett P, et al. A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. J Neurosci Methods. 2002; 115(2): 137–143.
17. Horn J. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965; 30(2):179–185.
18. Ibarra S. Automated Neuropsychological Assessment Metrics. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology. New York: Springer; 2011:325–327.
19. Jackson DN, Rushton JP. Males have greater g: sex differences in general mental ability from 100,000 17-to 18-year-olds on the Scholastic Assessment Test. Intelligence. 2006; 34(5):479–486.
20. 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–B191.
21. Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, et al. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). J Exp Psychol Appl. 2002; 8(2):75–84.
22. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Ann N Y Acad Sci. 2008; 1129:305–322.
23. Loughead J, Gur RC, Elliott M, Gur RE. Neural circuitry for accurate identification of facial emotions. Brain Res. 2008; 1194:37–44.
24. Lynn R, Irwing P. Sex differences in mental arithmetic, digit span, and g defined as working memory capacity. Intelligence. 2008; 36(3):226–235.
25. Moore TM, Gur RC, Thomas ML, Brown GG, Nock MK, et al. Development, administration, and structural validity of a brief, computerized neurocognitive battery: results From the Army Study to Assess Risk and Resilience in Servicemembers. Assessment. 2017; Jan. 1: 1073191116689820.
26. Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. 2015; 29(2):235–246.
27. Ragland JD, Turetsky BI, Gur RC, Gunning-Dixon F, Turner T, et al. Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. Neuropsychology. 2002; 16(3):370–379.
28. Raven JC. Advanced progressive matrices: sets I and II. London: HK Lewis; 1962.
29. Saykin AJ, Gur RC, Gur RE, Shtasel DL, Flannery KA, et al. Normative neuropsychological test performance: effects of age, education, gender and ethnicity. Appl Neuropsychol. 1995; 2(2):79–88.
30. Strangman GE, Sipes W, Beven G. Human cognitive performance in spaceflight and analogue environments. Aviat Space Environ Med. 2014; 85(10):1033–1048.
31. Thomas ML, Brown GG, Gur RC, Moore TM, Patt VM, et al. Measurement of latent cognitive abilities involved in concept identification learning. J Clin Exp Neuropsychol. 2015; 37(6):653–669.
32. Usui N, Haji T, Maruyama M, Katsuyama N, Uchida S, et al. Cortical areas related to performance of WAIS Digit Symbol Test: a functional imaging study. Neurosci Lett. 2009; 463(1):1–5.
33. Voyer D, Voyer S, Bryden MP. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychol Bull. 1995; 117(2):250–270.
34. Williams LM, Mathersul D, Palmer DM, Gur RC, Gur RE, Gordon E. Explicit identification and implicit recognition of facial emotions: I. Age effects in males and females across 10 decades. J Clin Exp Neuropsychol. 2009;31(3):257–277.

APPENDIX A. VALIDATION OF THE COGNITION TEST BATTERY FOR SPACEFLIGHT IN A SAMPLE OF HIGHLY EDUCATED ADULTS

Full Description of Cognition Tests

The Motor Praxis task. The Motor Praxis task (MP)[7] 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 is 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 and the final score is mean response time for all responses. Because it is highly unusual for someone to click outside of a square on this task and subjects are not instructed to hit the center of the square, there is no accuracy score.

The Visual Object Learning Test. The Visual Object Learning Test (VOLT) assesses participant memory for complex

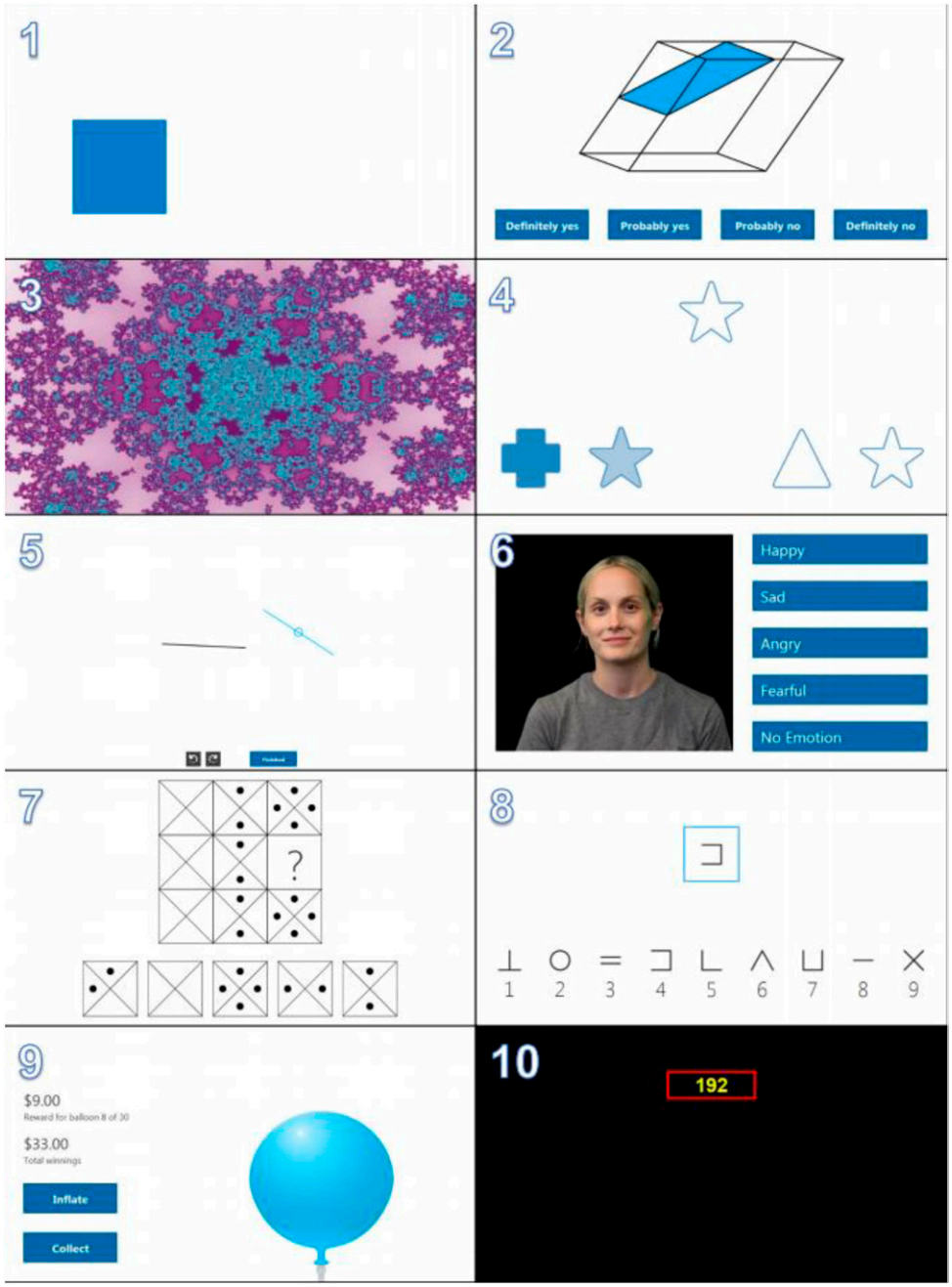

Fig. A1. Visualization of the 10 Cognition battery tests. 1 = Motor Practice; 2 = Visual Object Learning Test; 3 = Fractal NBACK; 4 = Abstract Matching; 5 = Line Orientation Test; 6 = Emotion Recognition Test; 7 = Matrix Reasoning Test; 8 = Digit Symbol Substitution Test; 9 = Balloon Analog Risk Task; 10 = Psychomotor Vigilance Test.

Table A1. Score Calculation for the 10 Cognition Tests.

TEST	ACCURACY MEASURE	SPEED MEASURE
MP	N/A	Mean response time for all responses
VOLT	Weighted percent correct*	Mean response time for all responses
NBACK	Weighted percent correct*	Mean response time for all responses
AM	Weighted percent correct*	Mean response time for all responses
LOT	(3 − average number of clicks off) / 3; 1 for ≤3 clicks off; 0 for >3 clicks off	Mean response time for all responses
ERT	IRT weighted percent correct†	Mean response time for all responses
MRT	IRT weighted percent correct†	Mean response time for all responses
DSST	Percent correct	Mean response time for all responses
BART	Risk taking propensity (0 = minimal; 1 = maximal)‡	Mean response time for all responses
PVT	1 − [(Lapses + FS) / (nStimuli + FS)]	10 − Mean of reciprocal response times.

* Weights based on average percent correct derived from a comparative population of Cognition study participants.
† Weights based on factor loading and item difficulty derived from a 2-parameter item response theory model.
‡ The risk score for the BART is based on the total number of pumps taken by the subject. This is compared to a cumulative relative frequency distribution of pumps taken by a comparative group of subjects. Each battery is associated with a unique cumulative relative frequency distribution, reflecting the unique order of balloons in each battery.
Lapses: response times ≥355 ms; FS: false starts (premature responses or responses without a stimulus); nStimuli: number of valid stimuli (false starts are not counted as valid stimuli).

figures.[5] 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, half from the learning set and half new. The final scores are weighted percent correct and mean response time for all responses.

The Fractal 2-Back. The Fractal 2-Back (F2B or NBACK)[13] is a nonverbal variant of the standard Letter 2-Back, which is currently included in the core Penn Computerized Neurocognitive Battery (CNB). NBACK tasks have become standard probes of the working memory system and activate canonical working memory brain areas. The Fractal NBACK 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 final scores are weighted percent correct and mean response time for all responses.

Abstract Matching. The Abstract Matching (AM) test[4] is a measure of the abstraction and flexibility components of executive function, including an ability to discern general rules from specific instances. Validity of the AM has been established mostly from its ability to distinguish patients with schizophrenia from healthy controls.[4,15] The test paradigm presents subjects with two pairs of objects at the bottom left and right of the screen, varied on specific perceptual dimensions (i.e., shape and fill). Subjects are presented with a target object in the upper middle of the screen that they must classify as belonging more with one of the two pairs based on a set of implicit, abstract rules. The current implementation uses 30 consecutive stimuli and the final scores are weighted percent correct and mean response time for all responses.

The Line Orientation Test. The Line Orientation Test (LOT) is a measure of spatial orientation and is derived from the well-validated Judgment of Line Orientation Test,[3] the computerized version of which was among the first to be administered with functional neuroimaging[6] and is used in the core CNB. 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 the number of degrees of rotation associated with each mouse click. Specifically, shorter line length, further distance from the stationary line, and small numbers of degrees per click are associated with increased difficulty,[12] such that the easiest possible item is one in which the lines are long, the lines are very close together, and each mouse click rotates the line a lot (e.g., 9° vs. 3°). Spatial orientation and reasoning are crucial for success in space missions, being necessary for repairs, craft piloting, and safe maneuvering in microgravity. The final accuracy score is 3 minus the average number of mouse clicks away from the correct response, divided by 3. Thus, a person who was exactly correct on all responses (0 clicks off) would have a score of (3 – 0)/3 = 1.0, and a person who was off by 3 on average would have a score of (3 – 3)/3 = 0.0; more than 3 clicks off on average also results in a score of 0. The rationale for this scoring method is that number of clicks away from a perfect answer is more informative than a simple correct/incorrect score per item; i.e., in addition to knowing whether someone got an item wrong, we can gain additional information by knowing how wrong he/she was. The final speed score is mean response time for all responses.

The Emotion Recognition Task. The Emotion Recognition Task (ERT) is a measure of visual emotion recognition that was developed[8] and validated with neuroimaging[10] and is part of the Penn CNB. The ERT presents subjects with photographs of professional actors (adults of varying age and ethnicity) portraying emotional facial expressions of varying intensities (biased toward 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. Item-Factor analysis of the ERT itemwise data revealed that some items had very small loadings and/or extremely low difficulty

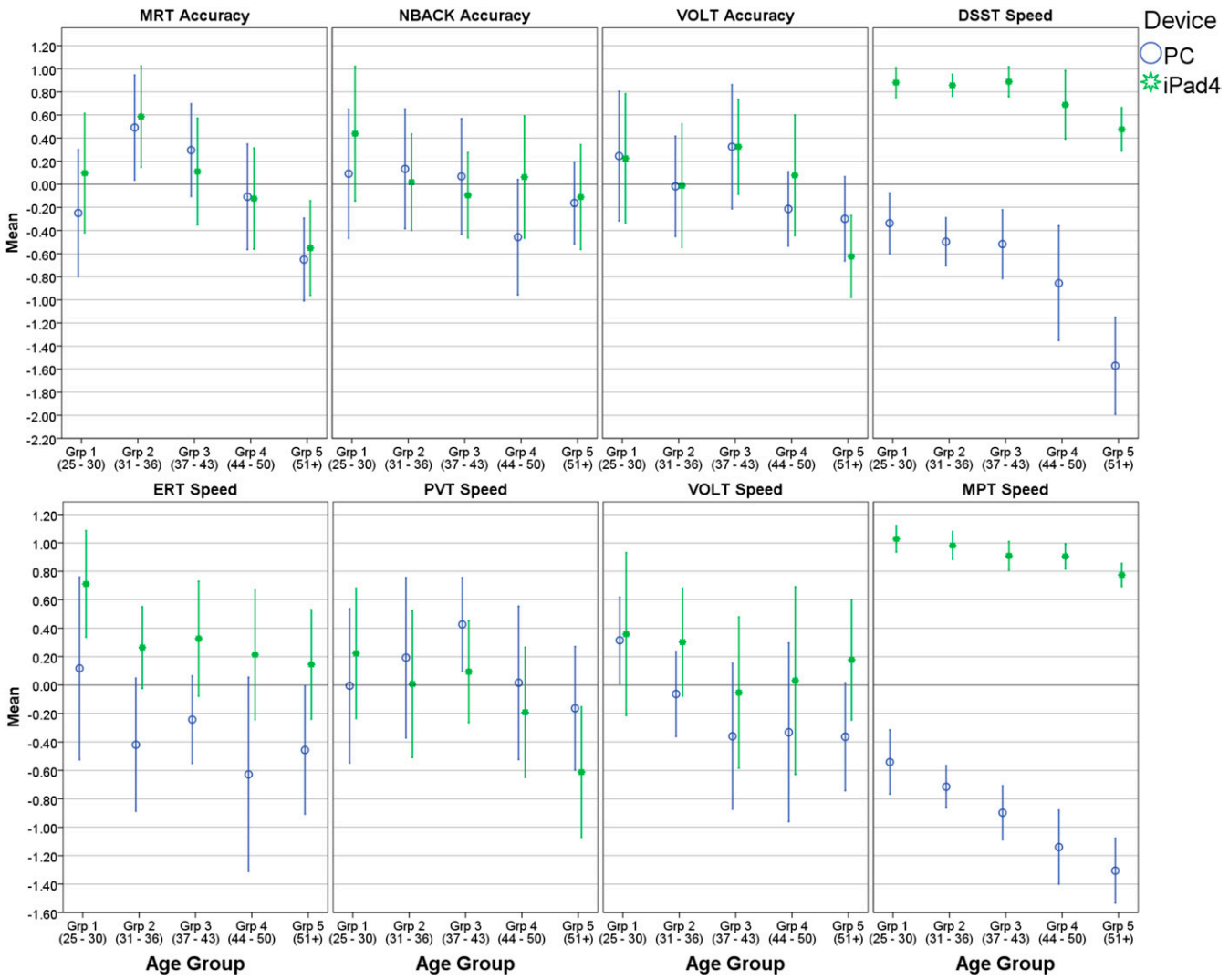

Fig. A2. Significant associations of accuracy and speed with age on the Cognition Test Battery. Grp = group; AM = Abstract Matching; MRT = Matrix Reasoning Test; VOLT = Visual Object Learning Test; DSST = Digit-Symbol Substitution Test; ERT = Emotion Recognition Test; PVT = Psychomotor Vigilance Test; MPT = Motor Praxis Task. The blue lines with white circles denote laptop use while the green lines with an asterisk denote iPad 4 use.

(>97% correct), indicating "bad" stimuli. These stimuli (loading <0.1 or >97% correct) were not considered in the calculation of the accuracy or speed score, which were weighted percent correct and mean response time for all responses, respectively.

The Matrix Reasoning Test. The Matrix Reasoning Test (MRT) is a measure of abstract reasoning and consists of increasingly difficult pattern matching tasks.[7] It is analogous to Raven's Progressive Matrices,[14] and 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. The MRT is included in the Penn CNB and has been validated along with all other tests in protocols using the CNB.[11] The final scores are weighted percent correct and mean response time for all responses.

The Digit-Symbol Substitution Task. The Digit-Symbol Substitution Task (DSST)[16] is a computerized adaptation of a paradigm used in the Wechsler Adult Intelligence Scale. 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 reassigned with each administration. Because participants are administered different numbers of items depending on how quickly they work, the final accuracy score is percent correct. The final speed score is the mean response time for all responses.

The Balloon Analog Risk Test. The Balloon Analog Risk Test (BART) is a validated assessment of risk taking behavior[9] and requires participants to either inflate an animated balloon or

Table A2. Mean Accuracy and Speed Scores (and Standard Deviations) for the 10 Cognition Battery Tests, by Device, Sex, and Age Group.

	DEVICE							
	IPAD4				PC			
	SEX				SEX			
	MALE		FEMALE		MALE		FEMALE	
	AGE GROUP		AGE GROUP		AGE GROUP		AGE GROUP	
	25–40	41+	25–40	41+	25–40	41+	25–40	41+
Test Score	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
AM Accuracy	0.48 (0.20)	0.45 (0.17)	0.53 (0.16)	0.47 (0.16)	0.51 (0.12)	0.46 (0.15)	0.50 (0.13)	0.46 (0.25)
BART Risk	0.40 (0.26)	0.46 (0.32)	0.52 (0.32)	0.52 (0.28)	0.46 (0.38)	0.38 (0.32)	0.49 (0.24)	0.25 (0.22)
DSST Accuracy	0.99 (0.01)	1.00 (0.00)	1.00 (0.01)	1.00 (0.01)	0.98 (0.02)	1.00 (0.01)	0.98 (0.03)	0.98 (0.03)
ERT Accuracy	0.44 (0.19)	0.41 (0.16)	0.51 (0.19)	0.45 (0.17)	0.51 (0.18)	0.52 (0.18)	0.57 (0.21)	0.54 (0.21)
LOT Accuracy	0.77 (0.10)	0.76 (0.10)	0.75 (0.11)	0.79 (0.10)	0.79 (0.12)	0.73 (0.09)	0.74 (0.10)	0.70 (0.09)
MRT Accuracy	0.51 (0.28)	0.48 (0.23)	0.51 (0.21)	0.43 (0.21)	0.55 (0.26)	0.51 (0.18)	0.52 (0.27)	0.35 (0.22)
NBACK Accuracy	0.46 (0.20)	0.56 (0.15)	0.52 (0.14)	0.42 (0.14)	0.53 (0.17)	0.48 (0.16)	0.48 (0.18)	0.39 (0.12)
PVT Accuracy	0.92 (0.08)	0.89 (0.10)	0.93 (0.06)	0.87 (0.11)	0.96 (0.04)	0.92 (0.06)	0.92 (0.07)	0.93 (0.05)
VOLT Accuracy	0.62 (0.26)	0.48 (0.22)	0.51 (0.25)	0.46 (0.24)	0.54 (0.06)	0.38 (0.15)	0.50 (0.29)	0.51 (0.19)
AM RT	2.56 (0.72)	2.69 (1.20)	2.56 (1.43)	3.32 (1.23)	3.17 (0.96)	2.88 (1.11)	3.09 (1.13)	3.21 (1.39)
BART RT	2.31 (0.86)	2.28 (0.92)	2.56 (0.63)	2.96 (1.25)	2.85 (0.89)	2.46 (1.23)	2.82 (1.06)	3.07 (1.12)
DSST RT	0.97 (0.09)	0.99 (0.15)	0.94 (0.08)	1.07 (0.20)	1.46 (0.21)	1.60 (0.39)	1.43 (0.18)	1.78 (0.38)
ERT RT	2.25 (0.43)	2.37 (0.80)	2.39 (0.89)	3.05 (0.41)	3.10 (0.92)	3.19 (1.30)	3.08 (0.98)	3.19 (0.82)
LOT RT	7.33 (2.36)	5.82 (1.91)	7.23 (1.71)	9.25 (3.80)	7.79 (1.98)	8.16 (2.09)	9.47 (3.20)	8.33 (1.52)
MRT RT	12.29 (3.41)	11.15 (1.79)	11.95 (3.21)	12.31 (3.27)	12.84 (3.42)	11.42 (2.24)	11.30 (2.35)	10.28 (3.66)
NBACK RT	0.58 (0.09)	0.64 (0.08)	0.63 (0.05)	0.67 (0.11)	0.60 (0.06)	0.61 (0.14)	0.63 (0.07)	0.64 (0.14)
PVT RT[†]	5.38 (0.35)	5.50 (0.54)	5.62 (0.47)	5.82 (0.40)	5.07 (0.31)	5.44 (0.33)	5.66 (0.56)	5.56 (0.39)
VOLT RT	2.04 (0.45)	2.00 (0.54)	2.27 (0.97)	2.36 (0.69)	2.30 (0.50)	2.41 (0.71)	2.02 (0.44)	2.30 (0.62)
MPT RT	0.46 (0.05)	0.53 (0.07)	0.52 (0.05)	0.52 (0.05)	1.03 (0.08)	1.17 (0.13)	0.99 (0.16)	1.18 (0.12)

[†] See Table A1 for how PVT performance speed (slowness) is calculated.

SD = standard deviation; AM = Abstract Matching; BART = Balloon Analog Risk Task; DSST = Digit Symbol Substitution Task; ERT = Emotion Recognition Task; LOT = Line Orientation Task; MRT = Matrix Reasoning Task; VOLT = Visual Object Learning Test; RT = response time (in seconds).

See Table A1 for details on how accuracy is calculated for each test.

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. 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 risk score for the BART is based on the total number of pumps taken by the subject. This is compared to a cumulative relative frequency distribution of pumps taken by a comparative group of subjects. Each unique battery is associated with a unique cumulative relative frequency distribution, reflecting the unique order of balloons in each battery. This score varies between 0 (no risk taken, all balloons collected) and 1 (maximum risk taken, all balloons popped). A score of 0.5 indicates that half of the comparative group took less risk while the other half took more risk. Here we treat BART risk-taking as an "accuracy" score along with the other accuracy scores. This might seem unusual, but we found that BART risk-taking correlates positively with all other accuracy measures except ERT, with some reaching statistical significance. This positive association of risk-taking with the other measures is apparent in the factor analysis described in the Results section of the article. The final speed measure is mean response time.

The Psychomotor Vigilance Test. The Psychomotor Vigilance Test (PVT) records reaction times (RT) to visual stimuli that occur at random interstimulus intervals.[2] 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 1 s. Subjects are instructed to be as fast as possible without hitting the spacebar in an absence of 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.[1] The final accuracy measure is:

$$A = 1 - \frac{(\text{Lapses} + \text{False Starts})}{(\text{Total Stimuli} + \text{False Starts})},$$

where "False Starts" are responses before the stimulus appears or within 130 ms of its appearing (i.e., false starts and coincident false starts; errors of commission), and the "Lapses" are failures to respond within 355 ms (i.e., errors of omission). The final speed metric is 10 minus the mean of the reciprocal response times.[3]

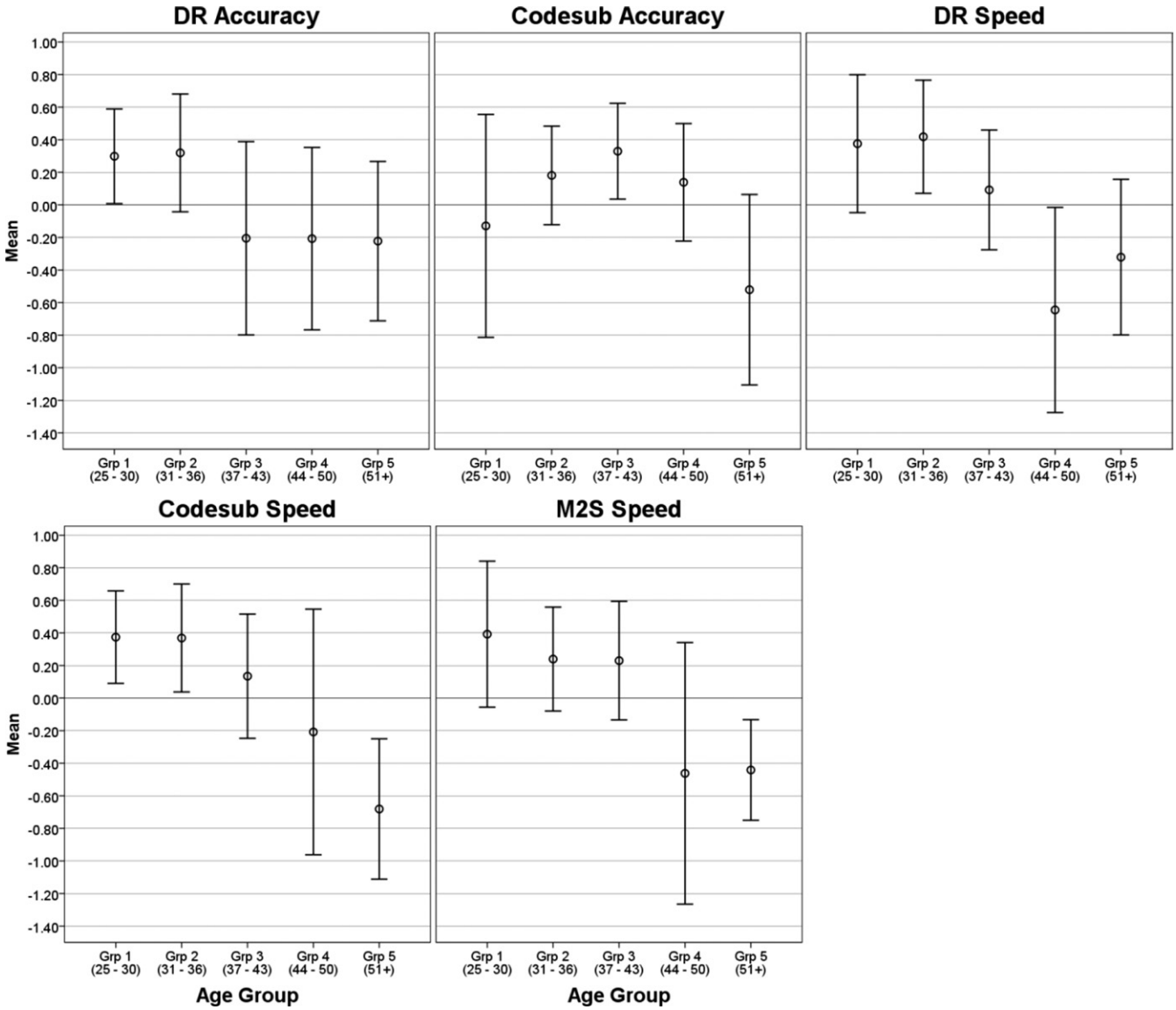

Fig. A3. Significant linear and nonlinear associations of accuracy and speed with age on the WinSCAT test battery. Grp = Group; DR = Delayed Recognition Memory; Codesub = Code Substitution; M2S = Match-to-Sample.

REFERENCES

1. Barger LK, Flynn-Evans EE, Kubey A, Walsh L, Ronda JM, et al. Prevalence of sleep deficiency and use of hypnotic drugs in astronauts before, during, and after spaceflight: an observational study. Lancet Neurol. 2014; 13(9):904–912.
2. Basner M, Dinges DF. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. Sleep. 2011; 34(5):581–591.
3. Benton AL, Varney NR, Hamsher KD. Visuospatial judgment. A clinical test. Arch Neurol. 1978; 35(6):364–367.
4. Glahn DC, Cannon TD, Gur RE, Ragland JD, Gur RC. Working memory constrains abstraction in schizophrenia. Biol Psychiatry. 2000; 47(1):34–42.
5. 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):602–612.
6. Gur RC, Gur RE, Obrist WD, Hungerbuhler JP, Younkin D, et al. Sex and handedness differences in cerebral blood flow during rest and cognitive activity. Science. 1982; 217(4560):659–661.
7. Gur RC, Ragland JD, Moberg PJ, Turner TH, Bilker WB, et al. Computerized neurocognitive scanning: I. Methodology and validation in healthy people. Neuropsychopharmacology. 2001; 25(5): 766–776.
8. Gur RC, Sara R, Hagendoorn M, Marom O, Hughett P, et al. A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. J Neurosci Methods. 2002; 115(2): 137–143.
9. Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, et al. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). J Exp Psychol Appl. 2002; 8(2):75–84.
10. Loughead J, Gur RC, Elliott M, Gur RE. Neural circuitry for accurate identification of facial emotions. Brain Res. 2008; 1194:37–44.
11. Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. 2015; 29(2):235–246.
12. Moore TM, Scott JC, Reise SP, Port AM, Jackson CT, et al. Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. Psychol Assess. 2015; 27(3):955–964.

13. Ragland JD, Turetsky BI, Gur RC, Gunning-Dixon F, Turner T, et al. Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. Neuropsychology. 2002; 16(3):370–379.
14. Raven JC. Advanced progressive matrices: sets I and II. London: HK Lewis; 1962.
15. Ridler K, Veijola JM, Tanskanen P, Miettunen J, Chitnis X, et al. Fronto-cerebellar systems are associated with infant motor and adult executive functions in healthy adults but not in schizophrenia. Proc Natl Acad Sci USA. 2006; 103(42):15651–15656.
16. Usui N, Haji T, Maruyama M, Katsuyama N, Uchida S, et al. Cortical areas related to performance of WAIS Digit Symbol Test: a functional imaging study. Neurosci Lett. 2009; 463(1):1–5.

APPENDIX B. TEST-RETEST RELIABILITY SIMULATION SCRIPT

```
# This script requires the mirt package.

library(mirt)

# We first simulate 10000 response patterns
  to 1000 items. Items are
# set to have discrimination parameters of
  0.3 and difficulty parameters
# ranging from extremely easy to extremely
  difficulty (identical on items
# 1-500 and items 501-1000). Because all
  items are of equal discrimination
# and difficulty across the first 500 and last
  500 items, the first 500 can
# be taken to be the first administration, and
  the last 500 items as the
# second administration of the same test.
  Correlations between sum scores on
# these two sets of 500 items indicates
  test-retest reliability. The first
# simulated set (x1 below) is generated from
  examinees only in the high
# ability range (1 to 3 standard deviations
  above the mean). The second
# simulated set (x2 below) is generated from
  examinees in the full ability
# range (−3 to 3 standard deviations below/
  above the mean).

x1 <- simdata(a=matrix(runif(1000,0.299,
   0.301),1000,1), d=matrix(c(seq(-2,2,
   length.out=500),seq(-2,2,length.out=5
   00)),1000,1),N=10000, itemtype="dich",
   Theta=matrix(runif(10000,1,3),
   10000,1))

x2 <- simdata(a=matrix(runif(1000,0.299,
   0.301),1000,1), d=matrix(c(seq(-2,2,
   length.out=500),seq(-2,2,length.out=5
   00)),1000,1),N=10000, itemtype="dich",
   Theta=matrix(runif(10000,-3,3),
   10000,1))

cor(cbind(rowSums(x1[,1:500]),rowSum
   s(x1[,501:1000])))
cor(cbind(rowSums(x2[,1:500]),rowSum
   s(x2[,501:1000])))

# Here we see that the first group (x1) with
  the narrower ability range
# yields lower estimates of test-retest
  reliability (~0.72) despite the fact
# that they were administered the same
  test as the second group, who yielded
# higher estimates (~0.96).
```