Age, Sex, and Repeated Measures Effects on NASA's "Cognition" Test Battery in STEM Educated Adults

Grace Lee; Tyler M. Moore; Mathias Basner; Jad Nasrini; David R. Roalf; Kosha Ruparel; Allison M. Port; David F. Dinges; Ruben C. Gur

BACKGROUND: Cognition is a neurocognitive test battery created at the University of Pennsylvania and adapted by the National Aeronautics and Space Administration (NASA). It comprises 10 neurocognitive tests that examine multiple domains, and has been validated in a normative sample of STEM-educated adults and compared to NASA's WinSCAT battery.

METHODS: The purpose of this study was to follow the original sample to assess Cognition and WinSCAT's test-retest reliability and age, sex, and test-retest interval effects on performance.

RESULTS: Performance on both Cognition and WinSCAT decreased with age but improved with repeated administration due to practice effects, and men had higher scores than women on tasks that required vigilant attention, spatial reasoning, and risk-taking behaviors. Assessment of test-retest reliability showed intraclass coefficients for efficiency ranging from 0.417 to 0.810, reflecting the broad nature of constructs assessed by Cognition.

DISCUSSION: Results largely matched predictions, with some counter-intuitive results for test-retest reliability interval.

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

Lee G, Moore TM, Basner M, Nasrini J, Roalf DR, Ruparel K, Port AM, Dinges DF, Gur RC. *Age, sex, and repeated measures effects on NASA's "Cognition" Test Battery in STEM educated adults.* Aerosp Med Hum Perform. 2020; 91(1):18–25.

Exploration of Mars has been a longtime goal of humanity; however, the duration of a mission to Mars would far exceed any previous space missions, the record being a 437-d mission by former Russian cosmonaut Valeri Polyakov. Along with the acute physiological and environmental risks of any long-duration space mission (high levels of radiation, constant danger of life-threatening equipment failure, etc.), prolonged exposure to these stressors could affect the wide-ranging cognitive abilities that astronauts must sustain for mission success and safety. So far, the evidence base for testing cognition in long-term spaceflight suggests that cognition is not impacted by long-term spaceflight, yet the evidence base is limited by the fact that the established test battery (WinSCAT) has not been used as a research tool, and the scheduling of its use may have missed critical periods (e.g., initial adaptation phase). Thus far it is clear that astronauts' self-reported symptoms of "space fog" (previously attributed to neurasthenia) can occur from a number of physiological and environmental stressors.[30]

Multiple factors can contribute to cognitive deficits in spaceflight: during initial adaptation to (and prolonged duration in) a microgravity environment, cognitive and motor behaviors may be significantly impaired, especially with increased demand over motor control while simultaneously performing cognitive tasks. Studies examining microgravity effects have consistently found deficits in executive function, memory, language, and visuospatial ability.[6]

The University of Pennsylvania (Penn Medicine) has been working in collaboration with NASA to develop a neurocognitive assessment suitable for studying these spaceflight-related effects. NASA uses The Spaceflight Cognitive Assessment Tool for Windows (WinSCAT) as its main operational test battery; however, WinSCAT focuses almost exclusively on executive (frontal lobe) functions. Titled "Cognition", the new assessment is a battery of 10 neurocognitive tests that are based on tasks used in functional neuroimaging. It has been administered

From the Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, and VISN4 Mental Illness Research, Education, and Clinical Center at the Philadelphia VA Medical Center, Philadelphia, PA, USA.

This manuscript was received for review in August 2019. It was accepted for publication in October 2019.

Address correspondence to: Tyler M. Moore, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104; tymoore@pennmedicine.upenn.edu.

DOI: https://doi.org/10.3357/AMHP.5485.2020

hundreds of times on the International Space Station and in various NASA-sponsored projects that include such spaceflight analogs as winter-over in Antarctica, the Human Exploration Research Analog (HERA) at Johnson Space Center, Hawai'i Space Exploration Analog & Simulation (HI-SEAS), and head-down tilt bedrest studies.[3] It is currently part of NASA's standardized behavioral measures, and is specifically designed to evaluate function in a broad range of cognitive domains relevant to spaceflight (executive, episodic memory, complex cognition, social cognition, and sensorimotor ability), as well as to address some of the $>$ 25 knowledge gaps and health risks related to cognitive functioning in NASA's Human Research Roadmap.[4,25] The objective is to ultimately create a brief, reliable battery of tests acceptable to the astronaut population, feasible in spaceflight, and consistent across projects and missions.

Cognition was previously validated in a normative sample of highly educated adults analogous to the space traveler population in qualifying criteria and age range. Specifically, Cognition's structure and sensitivity to age and sex differences was measured in a sample of 96 high-performing, STEM-educated adults (minimum of a Master's degree) ranging from 26 to 58 yrs.[25] Cognition's technical and psychometric performance were compared for consistency across platforms used during spaceflight (PC versus iPad) and compared to WinSCAT for efficacy. That is, subjects in this previous study were administered the Cognition battery twice in the same day on different devices. The results provided evidence that Cognition is an accurate predictor of WinSCAT scores, while WinSCAT predicted only the tasks of executive function on Cognition.[25] However, a second data collection time point was necessary to further examine the validity and reliability of the Cognition battery within 6+ month time spans between testing, so subjects were invited to return for a second trial. This second trial was subjects' third administration of Cognition (first visit included one previous iPad and one previous laptop administration) and second administration of WinSCAT (visit one administration on laptop).

Well-established literature has suggested that within our neurocognitive domains of interest, there is a reliable negative correlation between age and speed, as well as decreasing accuracy on some tasks with increasing age. Previous administration of the neurocognitive battery concluded that there was no age effect on working memory; however, a likely explanation would be a drastically different age range (18–84, versus 26–58).[16] We therefore expected speed to decrease with age on all tasks. Furthermore, we expected marked sex differences in the first time point, with better performance by men on spatial and motor tasks and better performance by women on memory and emotion processing tasks.[15] While age and performance correlations all matched their predicted outcomes at time point 1, the sex differences at time point 1 were surprising: women were marginally more accurate at emotion recognition, with no effect when accuracy and speed were combined, and men significantly outperformed women on an abstract reasoning task. These unexpected sex effects may have been due to shared characteristics of the self-selecting STEM-educated population, compared to previous randomized community samples; however, they could have also been due to random chance (Type I and Type II error). Therefore, for the present study, we retained the sex difference hypotheses supported by literature—i.e., the same expectations as in the previous (time point 1) study.[25]

Secondly, the first time point showed poor test-retest reliability for some Cognition tests, likely due to having two test forms on two devices. Re-administering the test a second time on a single platform (PC only) between the two trials would provide some more consistency and basis for comparison for the strength of the entire assessment. While scores may improve due to practice effects, overall test-retest reliability was expected to be higher than reported in the first study, as the first study used two different devices. Note that, for all test-retest reliability calculations, only laptop administrations were used (no second time point for iPad). A second question in the present manuscript was how test-retest interval (time between administrations) affects test-retest reliability of the tests. We hypothesized that longer test-retest intervals would result in poorer test-retest reliability.

METHODS

Subjects

In a previously published study, 96 subjects with the criteria stated below were administered the Cognition battery (twice) and WinSCAT battery (once) on a laptop and iPad. Of the original 96 subjects from the first study, 78 returned to take the Cognition battery again, and were re-administered both WinSCAT and Cognition on the PC only. These subjects all held at least a Master's degree in science, technology, engineering, or mathematics (STEM), were from the Philadelphia area, ranged in age from 26 to 58 yr at first performance, and were screened (via self-report questionnaire) for serious medical and psychiatric disorders that could affect performance. This age range was chosen because it best mirrors the age range of the majority of the current and recent space traveler population, but with an extended lower limit with the consideration that there may be younger space travelers in the future. Note that we recruited 18 additional subjects to achieve a total sample size of 96 at the second time point, but because these 18 were not assessed at the first time point, their data was not used in the present study. 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** shows the subject demographic characteristics.

Materials

Cognition Test Battery. Please see **Appendix A** online (DOI: https://doi.org/10.3357/AMHP.5485sd.2020) for a full description of Cognition tests.

The Cognition battery comprises 10 neurocognitive tests that have been previously validated and described in detail.[4] Most of the tests in the Cognition battery are from the Penn Computerized Neurocognitive Battery (CNB),[15,16,27] which has

Table I. Demographics by Sex, Age, and Duration Between Time Points.

PARTICIPANTS ($N = 78$)		
	MALE	**FEMALE**
Sex	46%	54%
Age (years)		
Range	26-57	26-58
Mean, SD	41.19 ± 9.37	42.21 ± 9.27
Duration (days)*		
Range	60.0-677.9	21.0-651.9
Mean, SD	439.46 ± 147.80	458.15 ± 134.24

* Duration = time between Time Point 1 and Time Point 2.

been widely used and validated in assessment of military service members,[26,31] childhood development,[14] and genomic research in populations with or at risk for psychiatric disorders.[11] In addition to these tests, Cognition uses the Psychomotor Vigilance Test[21] and the Digital Symbol Substitution Test,[29] which have been used extensively in spaceflight. A brief description of all tests follows.

The Motor Praxis Task *(MP).*[13] MP was administered at the start of testing to ensure that subjects have sufficient command of the computer interface, and as a measure of sensorimotor speed. Subjects were instructed to click on squares that appeared randomly on the screen, each square being successively smaller with each click (and therefore harder to track). The first bout was to familiarize the participant with the interface and task, and for the second round they were instructed to do the task as fast as they could. Only the second bout was used in analyses.

The Visual Object Learning Test *(VOLT)*[10]. The VOLT was administered to assess subjects' memory for complex figures. Subjects were asked to memorize 10 sequentially presented three-dimensional figures to the best of their ability, and then to identify these figures among 20 sequentially presented figures, half from the learning set and half new.

The Fractal 2-Back *(NBACK).*[28] The NBACK is a nonverbal variant of the standard Letter 2-Back test, a working memory assessment, that is currently included in the CNB. The NBACK sequentially displayed a set of images (fractals), each potentially repeated multiple times. The participant was asked to respond when the current fractal matched the fractal displayed two figures before.

Abstract Matching *(AM).*[9] The AM test is a measure of abstraction and flexibility components of executive function, including an ability to discern general rules from specific instances. The test presented one pair of objects on each side of the screen, with variations in object shape and fill. Subjects were also presented with a target object above these shapes, and asked to classify the target object with one of the two pairs.

The Line Orientation Test *(LOT).*[5] The LOT is a measure of spatial orientation derived from the Judgment of Line Orientation Test of the CNB. The LOT presented two lines at a time; one remained stationary, while the participant would rotate the other line around to make it parallel with the stationary line.

The Emotion Recognition Test *(ERT).* The ERT is a measure of visual emotion recognition that is part of the CNB.[22] The ERT presented subjects with photographs of professional actors of varying age and ethnicity, portraying emotional facial expressions of varying intensities. Subjects were asked to choose from a set of emotional labels ("happy," "sad," "angry," "fearful," "no emotion") the one that they felt most closely matched the facial expression being displayed.

The Matrix Reasoning Test *(MRT).*[13] The MRT is a measure of abstract reasoning that consists of increasingly difficult pattern matching tasks. Patterns were overlaid on a matrix, with one element missing; the subject was asked to select the element that fits the pattern.

The Digit-Symbol Substitution Task *(DSST).*[32] The DSST is a computerized adaptation of a paradigm used in the Wechsler Adult Intelligence Scale. The DSST required the subject to refer to a legend relating each of the digits 1–9 to specific symbols. One of the symbols would appear on the screen, and the subject was asked to select the corresponding digit as quickly as possible. The legend key was randomly assigned to new symbols with each administration, and test duration was fixed at 90 s.

The Balloon Analog Risk Test *(BART).*[20] The BART assesses risk taking behavior where subjects are asked to either inflate an animated balloon or collect the current reward; subjects were rewarded by points in proportion to the final size of the balloon, but the balloon could pop after a random number of pumps that changed with each trial, voiding that reward.

The Psychomotor Vigilance Test *(PVT).*[2] The PVT is a measure of vigilant attention. Subjects were instructed to monitor a box on the screen and hit the space bar once a millisecond counter (stimulus) appeared in the box as fast as possible without hitting the spacebar in the absence of the stimulus.

The Spaceflight Cognitive Assessment Tool for Windows (WinSCAT)[18] is described below:

The Code Substitution Test *(Codesub).* The Codesub is a measure of visual scanning, and is very similar to the DSST. The subject was shown a number-symbol pair and asked to determine if it matched any of the pairs presented on the same screen.

The Running Memory Continuous Performance Test *(CPT).* The CPT is a measure of working memory and attention, similar to the Fractal 2-Back used in Cognition. However, the stimuli in the CPT are numbers rather than fractals, and the subject was asked to respond when the number was the same as the one immediately before (rather than 2-back).

Mathematical Processing *(Math).* This test is a measure of computational processing and mathematical achievement. Subjects were given a three term math problem and asked to decide whether the answer was greater or less than 5.

Delayed Matching to Sample *(M2S).* The M2S is a measure of visual memory. Subjects were shown a 4 × 4 grid comprised of squares with different colors. Five seconds later, they were shown two different, similarly comprised grids and asked to decide which one matched the first grid shown.

The Delayed Recognition Code Substitution Test *(DR).* The DR is a measure of short-term memory. Subjects were shown a number-symbol pair similar to the Codesub task above, but without the presence of the key. Subjects were then asked to decide whether the pair matched any of the pairs in the key shown in the previous Codesub task.

Statistical Analysis

We performed linear mixed models predicting each test's speed, accuracy, and efficiency scores using visit (time point) as an independent variable, with age and sex as covariates of interest. Efficiency scores were calculated as follows: speed and accuracy scores for each test were z-transformed and averaged, where Speed = RT * (-1) so that a higher Average RT score would indicate faster performance. Efficiency scores are necessary because optimal performance is characterized by both the ability to perform accurately, and the ability to perform quickly. We also examined test-retest reliability for accuracy, speed, and efficiency scores on the 10 Cognition tests and 5 WinSCAT tests. Time point was regressed out of the scores (to account for practice effects), and intraclass correlations (ICCs) were estimated between the first and second time points. The above was done separately for those with a "long" test-retest interval (482+ d) vs. "short" interval (< 482 d) to examine the effect of time lag on test-retest reliability. The thresholds for "long" and "short" intervals were determined by median split of the sample.

RESULTS

Table II shows mixed model results predicting age, sex, and visit effects on accuracy, speed, and efficiency scores for all 10 Cognition tests. For all significant results between efficiency and age, the negative coefficients indicated that subjects reliably performed less efficiently with each decade of age. Subjects were also less accurate on the NBACK with increasing age. Subjects were significantly slower with increasing age on the DSST, MP, MRT, and VOLT. Men, on average, were significantly faster than women on the BART, DSST, LOT, MP, and PVT, and more efficient on the BART and LOT. From Visit 1 to Visit 2, subjects improved significantly in accuracy on the MRT, NBACK, and VOLT, were faster on the LOT and PVT, and more efficient on the VOLT and BART.

Table III shows mixed model results for associations between WinSCAT accuracy, speed, and efficiency scores and age, sex, and visit for all five tests. Subjects were significantly less accurate on the M2S with increasing age, and less efficient on the DR, Codesub, M2S, and CPT. Speed for the DR, Codesub, and

Table II. Mixed Model Results Predicting Cognition Performance (Accuracy, Average Response Time (AvRT), and Efficiency) Using Age, Sex, and Time Point (Visit).

	AGE (DECADES)			FEMALE SEX			VISIT		
SCORE	B	SE	SIG.	B	SE	SIG.	B	SE	SIG.
MP Accuracy	-0.091	0.107	0.398	0.193	0.196	0.328	0.104	0.117	0.376
VOLT Accuracy	-0.158	0.105	0.137	0.075	0.193	0.700	**0.418**	**0.112**	**< 0.0005**
NBACK Accuracy	**-0.303**	**0.105**	**0.005**	-0.048	0.194	0.805	**0.328**	**0.099**	**0.002**
AM Accuracy	-0.057	0.102	0.576	-0.160	0.188	0.398	-0.129	0.130	0.323
LOT Accuracy	-0.054	0.114	0.637	-0.006	0.210	0.977	-0.134	0.094	0.161
ERT Accuracy	-0.082	0.099	0.408	0.264	0.182	0.150	-0.236	0.134	0.081
MRT Accuracy	-0.176	0.100	0.082	-0.193	0.185	0.299	**0.307**	**0.126**	**0.017**
DSST Accuracy	0.126	0.112	0.263	0.145	0.206	0.483	0.073	0.098	0.462
BART Risk	-0.011	0.105	0.917	-0.148	0.194	0.446	0.177	0.122	0.152
PVT Accuracy	-0.010	0.115	0.931	-0.179	0.213	0.402	0.115	0.086	0.187
MP Speed	**-0.498**	**0.096**	**< 0.0005**	**-0.354**	**0.176**	**0.048**	-0.082	0.093	0.380
VOLT Speed	**-0.223**	**0.109**	**0.044**	-0.011	0.201	0.955	0.031	0.103	0.768
NBACK Speed	0.023	0.107	0.831	-0.090	0.198	0.649	-0.150	0.124	0.233
AM Speed	-0.123	0.110	0.270	0.034	0.204	0.868	0.070	0.106	0.510
LOT Speed	-0.188	0.098	0.058	**-0.619**	**0.180**	**0.001**	**0.266**	**0.113**	**0.021**
ERT Speed	-0.186	0.110	0.096	-0.062	0.204	0.762	0.049	0.100	0.629
MRT Speed	**-0.230**	**0.110**	**0.040**	0.104	0.203	0.609	-0.020	0.099	0.840
DSST Speed	**-0.421**	**0.105**	**< 0.0005**	**-0.498**	**0.194**	**0.012**	0.103	0.060	0.089
BART Speed	-0.151	0.100	0.136	**-0.668**	**0.184**	**0.001**	0.140	0.109	0.200
PVT Speed	-0.010	0.113	0.931	**-0.597**	**0.210**	**0.006**	**0.135**	**0.067**	**0.047**
MP Efficiency	**-0.436**	**0.097**	**< 0.0005**	-0.119	0.179	0.508	0.016	0.110	0.883
VOLT Efficiency	**-0.272**	**0.105**	**0.012**	0.045	0.194	0.815	**0.321**	**0.107**	**0.004**
NBACK Efficiency	-0.181	0.112	0.110	-0.130	0.207	0.533	0.089	0.105	0.401
AM Efficiency	-0.123	0.108	0.256	-0.087	0.199	0.663	-0.041	0.114	0.720
LOT Efficiency	-0.187	0.105	0.078	**-0.485**	**0.193**	**0.014**	0.103	0.103	0.322
ERT Efficiency	-0.204	0.104	0.054	0.153	0.192	0.430	-0.141	0.117	0.233
MRT Efficiency	**-0.293**	**0.103**	**0.006**	-0.065	0.190	0.733	0.208	0.113	0.070
DSST Efficiency	-0.202	0.114	0.081	-0.241	0.210	0.256	0.120	0.078	0.128
BART Efficiency	-0.113	0.104	0.282	**-0.584**	**0.193**	**0.003**	**0.226**	**0.101**	**0.028**
PVT Efficiency	-0.010	0.115	0.929	-0.421	0.212	0.051	0.136	0.075	0.076

All coefficients are in standard deviation (SD) units; for age, the base unit is decade rather than year because the per-year increment change was extremely small; B = Coefficient; SE = standard error; Sig. = significance. Significant effects are bolded.

Table III. Mixed Model Results Predicting WinSCAT Performance (Accuracy, Speed, and Efficiency) Using Age, Sex, and Time Point (Visit).

	AGE (DECADES)			MALE/FEMALE			VISIT		
SCORE	B	SE	SIG.	B	SE	SIG.	B	SE	SIG.
DR Accuracy	-0.219	0.012	0.066	-0.375	0.210	0.079	**0.377**	**0.108**	**0.001**
Codesub Accuracy	-0.199	0.011	0.081	0.040	0.202	0.842	0.057	0.140	0.686
M2S Accuracy	**-0.230**	**0.011**	**0.049**	-0.092	0.205	0.655	-0.034	0.129	0.794
Math Accuracy	-0.118	0.012	0.328	-0.148	0.214	0.492	0.075	0.123	0.546
CPT Accuracy	-0.246	0.013	0.062	-0.281	0.233	0.232	**0.280**	**0.094**	**0.004**
DR Speed	**-0.365**	**0.012**	**0.003**	-0.065	0.210	0.759	**0.304**	**0.101**	**0.004**
Codesub Speed	**-0.417**	**0.011**	**0.001**	**-0.615**	**0.205**	**0.004**	**0.251**	**0.068**	**< 0.0005**
M2S Speed	**-0.341**	**0.012**	**0.005**	-0.417	0.211	0.052	0.114	0.093	0.222
Math Speed	-0.098	0.013	0.445	**-0.460**	**0.228**	**0.047**	**0.254**	**0.077**	**0.002**
CPT Speed	-0.230	0.012	0.061	**-0.680**	**0.216**	**0.003**	0.127	0.073	0.085
DR Efficiency	**-0.378**	**0.011**	**0.002**	-0.285	0.204	0.167	**0.441**	**0.098**	**< 0.0005**
Codesub Efficiency	**-0.426**	**0.011**	**< 0.0005**	-0.391	0.201	0.056	**0.211**	**0.100**	**0.038**
M2S Efficiency	**-0.357**	**0.012**	**0.003**	-0.318	0.207	0.131	0.050	0.104	0.631
Math Efficiency	-0.128	0.013	0.319	-0.371	0.228	0.109	**0.200**	**0.083**	**0.019**
CPT Efficiency	**-0.276**	**0.012**	**0.030**	**-0.547**	**0.224**	**0.017**	**0.229**	**0.059**	**< 0.0005**

All coefficients are in standard deviation (SD) units; for age, the base unit is decade rather than year because the per-year increment change was extremely small; B = Coefficient; SE = standard error; Sig. = significance. Significant effects are bolded.

M2S tasks all decreased with age as well, indicating slower performance. Men were faster than women on the Codesub, Math, and CPT tasks, and more efficient than women on the CPT. From Visit 1 to Visit 2, subjects' accuracy scores improved on the DR and CPT, and were faster on the DR, Codesub, and Math tasks. Subjects were more efficient on the second visit for all tests except for the M2S.

Table IV shows test-retest reliability (ICCs) for all Cognition and WinSCAT tests. Among the Cognition scores, the lowest ICC was 0.136 for ERT Accuracy at the longer interval, and the highest was 0.882 for the DSST Response Time (RT) at the longer interval. Among the WinSCAT scores, the lowest ICC was 0.287 for Codesub Accuracy at the shorter interval and highest was 0.892 for CPT Efficiency at the shorter interval. Focusing on efficiency scores, for the shorter interval, PVT and CPT accuracy were most highly correlated between visits for Cognition and WinSCAT, respectively, while MP and M2S had the lowest ICCs. **Fig. 1** shows the tests with the best and worst test-retest reliability for Cognition and WinSCAT, along with test-retest reliability for the full batteries (efficiency).

Table IV. Test-Retest Reliability (Intra-Class Coefficient) for 10 Cognition Tests and Five WinSCAT Tests (Accuracy, Response Time, and Efficiency), by Length of Inter-Visit Interval.

	SHORT INTERVAL (UP TO 483 d)			LONG INTERVAL (OVER 483 d)		
TEST	ACCURACY	RT	EFFICIENCY	ACCURACY	RT	EFFICIENCY
			Cognition			
MP	0.547	0.637	0.417	0.341	0.519	0.456
VOLT	0.582	0.563	0.590	0.347	0.608	0.451
NBACK	0.724	0.532	0.710	0.389	0.318	0.423
AM	0.433	0.591	0.514	0.152	0.525	0.450
LOT	0.755	0.601	0.678	0.416	0.416	0.473
ERT	0.426	0.596	0.428	0.136	0.618	0.475
MRT	0.316	0.627	0.464	0.407	0.550	0.487
DSST	0.698	0.794	0.785	0.249	0.882	0.676
BART	0.581	0.430	0.621	0.251	0.645	0.550
PVT	0.749	0.857	0.810	0.576	0.786	0.701
			WinSCAT			
DR	0.513	0.659	0.567	0.620	0.521	0.629
Codesub	0.287	0.780	0.606	0.339	0.866	0.642
CPT	0.887	0.798	0.892	0.628	0.843	0.868
M2S	0.341	0.578	0.547	0.455	0.738	0.597
Math	0.421	0.820	0.748	0.515	0.806	0.796

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; RT = response time. Cognition mean ICCs (Accuracy, Speed, Efficiency): 0.454, 0.605, 0.558; WinSCAT mean ICCs (Accuracy, Speed, Efficiency): 0.500, 0.741, 0.689.

DISCUSSION

Repeated administration of the Cognition and WinSCAT tests was used to assess the reliability of Cognition and examine age, sex, and test-retest interval duration effects on performance. Overall, many of the test results of significance were in the predicted direction, and test-retest reliability scores suggest moderately sustained reliability of Cognition as a predictor of neurocognitive ability.

Our results showed the predicted negative correlation between performance and age on both tests. On Cognition, subjects were slower on the MRT and DSST with every 1-yr increase in age; on WinSCAT, subjects were similarly slower, less accurate, and less efficient with increasing age. These findings are consistent with previous evidence of age effects on accuracy and speed; abilities such as conceptual reasoning, memory, and processing speed have often been shown to decline over time, and can be impacted by both age-related gray and white matter loss.[17] Decreased lateral

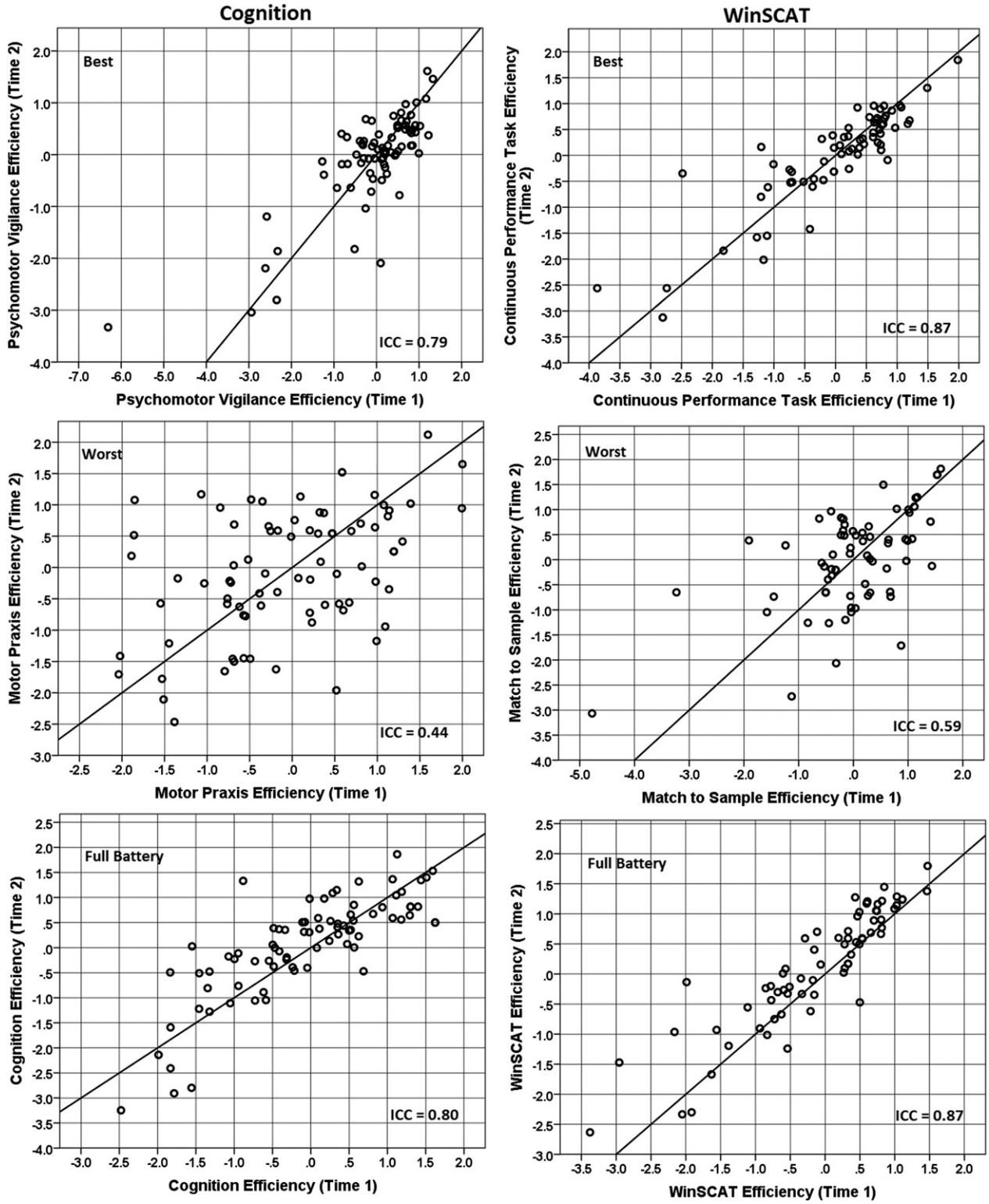

Fig. 1. Best, worst, and full-battery test-retest reliability for Cognition and WinSCAT.

frontal gray matter volume has been associated with a decrease in the ability to organize and execute strategies in attention and executive function tests;[32,36] declines in white matter tract integrity in anterior white matter have been linked to executive function deficits, and loss of integrity in the corpus callosum may lead to age-related cognitive decline.[17]

Sex effects also matched many of the predicted outcomes, although women again did not show better performance over men in emotion identification. Men were faster on the BART, DSST, LOT, MP, and PVT, and significantly more efficient on the LOT and BART. Previous studies are consistent with these findings; a previous meta-analysis strongly supports the presence

of sex differences in spatial ability,[33] as tested in the LOT. Men are also generally more inclined to engage in risk-taking behaviors as evidenced by more efficient scores on the BART.[7] Faster scores on multiple tests on both Cognition and WinSCAT are consistent with results from the first time point[25] and previous literature that suggests that men perform faster on both reaction and response time tests.[1] As in the first time point, women did not significantly outperform men on the ERT, as our hypotheses had expected based on earlier findings with this test.[15,35] It is impossible to conclude at this point whether the lack of this sex difference is due to power or whether high-functioning men are better or high-functioning women worse to the point of diminishing differences. The lack of sex differences on the memory task (VOLT) is consistent with earlier findings in a community sample[15] where spatial memory did not show sex differences while word and face memory were performed significantly better by women. Memory for complex figures has not traditionally shown sex effects, perhaps due to its equal focus on memory function and visual-spatial orientation, the former favoring women and the latter favoring men.[10] Word and face memory were not included in Cognition. Our results with the NBACK are likewise consistent with previous literature that supported a male advantage among adults in working memory capacity.[24] Lastly, we did not see any sex effects on MRT accuracy in Cognition at the second time point, so the effect seen in time point 1 may have been due to random error.

Results of re-administration (Visit 1 to Visit 2) also demonstrated strong practice effects. In all results of significance for accuracy, speed, and efficiency in Cognition and WinSCAT, subjects improved during their second visit, but note that the second-visit administration was subjects' third time taking Cognition since they took it twice at time point 1. Practice effects are often strongest between the first and second administrations of cognitive test batteries and become insignificant with further administrations.[8] We are currently examining the effects of administering Cognition 15 times in another NASA study.

Another motive for re-administration was to examine test-retest reliability of repeated administrations of the Cognition battery. Results for Cognition showed a wide range of test-retest reliability across tests and score types (accuracy, speed, efficiency), ranging from poor to moderate/good. Test-retest reliability for WinSCAT was generally higher, which was expected given its narrow focus on executive tasks. Higher test-retest reliability of executive tasks has been demonstrated, and studies of batteries containing a mix of executive and nonexecutive tasks (e.g., Cambridge Automated Neuropsychological Test Battery, among others) have found a range of test-retest reliability coefficients (poor to moderate/good) similar to the range found here for Cognition.[12,23,34] Another confounding factor that persisted throughout both time points was the relatively small sample size: the sample was further limited to a narrow range of abilities, due to similar education and training criteria; a simulation done in the first time point comparing high ability-only and full-range ability showed that test-retest reliability statistics were consistently higher for the full-range ability sample, and further research can be done by examining differences between STEM-educated and normative populations.[25]

Also central to the test-retest findings was the effect of test-retest interval on reliability. We hypothesized that test-retest reliability would decrease as the length of the test-retest interval increased, and results from these analyses were mixed. For Cognition, 22 (73%) of 30 scores showed lower test-retest reliability for the longer interval, lending moderate support to our hypothesis. For WinSCAT, 4 (27%) of 15 scores showed lower test-retest reliability for the longer interval, providing evidence directly contrary to our hypothesis. Feasible explanations include chance (small sample size) and different numbers of administrations of Cognition (3) versus WinSCAT (2), but further research on test-retest interval effects is clearly needed.

In summary, administration of Cognition and WinSCAT at a second time point largely supported the original hypotheses around age and sex effects on performance. Marginal sex differences in performance on emotion identification persisted in this readministration; this, as well as moderate test-retest reliability ICCs, can be further examined in future research on the characteristics of a STEM-educated population[19,29] compared to a community sample. The mean test-retest reliability ICCs for Cognition, while slightly lower than WinSCAT on average, suggest that Cognition is a reliable measure of performance, although continued administrations of the test would be helpful in further studies.

ACKNOWLEDGMENTS

This research was supported by the National Space Biomedical Research Institute (NSBRI) through NASA NCC 9-58; by NASA through grants NNX14AM81G, NNX14AH27G, and NNX14AH98G; by NIMH through grants MH089983, MH019112, MH096891, and MH042228; through the Office of Naval Research through grant ONR N00014-11-1-0361; by NIH through grant R01NR004281-14; and by the Dowshen Neuroscience Fund.

Financial Disclosure Statement: The authors have no competing interests to disclose.

Authors and affiliations: Grace Lee, B.A., Tyler M. Moore, Ph.D., David R. Roalf, Ph.D., Kosha Ruparel, M.S.E., Allison M. Port, B.A., and Ruben C. Gur, Ph.D., Department of Psychiatry, Neuropsychiatry Section, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; Mathias Basner, M.D., Ph.D., M.Sc., Jad Nasrini, 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; andRuben C. Gur, Ph.D., VISN4 Mental Illness Research, Education, and Clinical Center at the Philadelphia VA Medical Center, Philadelphia, PA, USA.

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 (Basel). 2011; 34(5):581–591.
3. Basner M, Nasrini J, Hermosillo E, McGuire S, Dinges DF, et al. SPACECOT Investigator Group. Effects of -12° head-down tilt with and without elevated levels of CO_2 on cognitive performance: the SPACECOT study. J Appl Physiol (1985). 2018; 124(3):750–760.

4. 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.
5. Benton AL, Varney NR, Hamsher K. Visuospatial judgment-reply. Arch Neurol. 1979; 36(1):59.
6. Bigelow RT, Agrawal Y. Vestibular involvement in cognition: Visuospatial ability, attention, executive function, and memory. J Vestib Res. 2015; 25(2):73–89.
7. Byrnes JP, Miller DC, Schafer WD. Gender differences in risk taking: A meta-analysis. Psychol Bull. 1999; 125(3):367–383.
8. Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. J Int Neuropsychol Soc. 2003; 9(3):419–428.
9. Glahn DC, Cannon TD, Gur RE, Ragland JD, Gur RC. Working memory constrains abstraction in schizophrenia. Biol Psychiatry. 2000; 47(1): 34–42.
10. 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.
11. 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.
12. Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS vital signs. Arch Clin Neuropsychol. 2006; 21(7):623–643.
13. 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.
14. 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.
15. 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.
16. 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.
17. Harada CN, Natelson Love MC, Triebel KL. Normal cognitive aging. Clin Geriatr Med. 2013; 29(4):737–752.
18. 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.
19. Lamb R, Akmal T, Petrie K. Development of a cognition-priming model describing learning in a STEM classroom. J Res Sci Teaching. 2015; 52(3):410–437.
20. 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.
21. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Ann N Y Acad Sci. 2008; 1129(1):305–322.
22. Loughead J, Gur RC, Elliott M, Gur RE. Neural circuitry for accurate identification of facial emotions. Brain Res. 2008; 1194:37–44.
23. Lowe C, Rabbitt P. Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. Cambridge Neuropsychological Test Automated Battery. International Study of Post-Operative Cognitive Dysfunction. Neuropsychologia. 1998; 36(9):915–923.
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, Basner M, Nasrini J, Hermosillo E, Kabadi S, et al. Validation of the cognition test battery for spaceflight in a sample of highly educated adults. Aerosp Med Hum Perform. 2017; 88(10):937–946.
26. 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. 2019; 26(1):125–143.
27. Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric properties of the Penn computerized neurocognitive battery. Neuropsychology. 2015; 29(2):235–246.
28. 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.
29. Sharobeam M. Research and teaching: The variation in spatial visualization abilities of college male and female students in STEM fields versus non-STEM fields. J Coll Sci Teach. 2016; 046(02).
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. White N, Forsyth B, Lee A, Machado L. Repeated computerized cognitive testing: Performance shifts and test-retest reliability in healthy young adults. Psychol Assess. 2018; 30(4):539–549.
35. 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.
36. Zimmerman ME, Brickman AM, Paul RH, Grieve SM, Tate DF, et al. The relationship between frontal gray matter volume and cognition varies across the healthy adult lifespan. Am J Geriatr Psychiatry. 2006; 14(10):823–833.