Published in final edited form as:
J Clin Exp Neuropsychol. 2020 July ; 42(5): 516–529. doi:10.1080/13803395.2020.1773765.

Cognition Test Battery: Adjusting for Practice and Stimulus Set Effects for Varying Administration Intervals in High Performing Individuals

Mathias Basner, MD, PhD, MSc[a,*], **Emanuel Hermosillo, BA**[a], **Jad Nasrini, BS**[a], **Salil Saxena, MD**[a], **David F. Dinges, PhD**[a], **Tyler M. Moore, PhD**[b], **Ruben C. Gur, PhD**[b]

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

[b]Brain Behavior Laboratory, Neuropsychiatry Section, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Abstract

Introduction: Practice effects associated with the repeated administration of cognitive tests often confound true therapeutic or experimental effects. Alternate test forms help reduce practice effects, but generating stimulus sets with identical properties can be difficult. The main objective of this study was to disentangle practice and stimulus set effects for Cognition, a battery of 10 brief cognitive tests specifically designed for high-performing populations with 15 unique versions for repeated testing. A secondary objective was to investigate the effects of test-retest interval on practice effects.

Methods: The 15 versions of Cognition were administered in three groups of 15-16 subjects (total N=46, mean±SD age 32.5±7.2 years, range 25-54 years, 23 male) in a randomized but balanced fashion with administration intervals of ≥10 days, ≤5 days, or 4 times per day. Mixed effect models were used to investigate linear and logarithmic trends across repeated administrations in key speed and accuracy outcomes, whether these trends differed significantly between administration interval groups, and whether stimulus sets differed significantly in difficulty.

Results: Protracted, non-linear practice effects well beyond the second administration were observed for most of the 10 Cognition tests both in accuracy and speed, but test-retest administration interval significantly affected practice effects only for 3 out of the 10 tests and only in the speed domain. Stimulus set effects were observed for the 6 Cognition tests that use unique sets of stimuli. Factors were established that allow for correcting for both practice and stimulus set effects.

[*]Corresponding author: Mathias Basner, MD, PhD, MSc, Unit of Experimental Psychiatry, Division of Sleep and Chronobiology, Perelman School of Medicine, University of Pennsylvania, 1019 Blockley Hall, 423 Guardian Drive, Philadelphia PA 19104-6021, Phone: 215 573 5866, Fax: 215 573 6410, basner@pennmedicine.upenn.edu. Address where work was conducted: Unit of Experimental Psychiatry, Division of Sleep and Chronobiology, Perelman School of Medicine, University of Pennsylvania, 1019 Blockley Hall, 423 Guardian Drive, Philadelphia PA 19104-6021.

Disclosure Statement
Financial disclosure: none
Non-financial disclosure: none

Conclusions: Practice effects are pronounced and probably under-appreciated in cognitive testing. The correction factors established in this study are a unique feature of the Cognition battery that can help avoid masking practice effects, address noise generated by differences in stimulus set difficulty, and facilitate interpretation of results from studies with repeated assessments.

Keywords

cognition; practice; learning; test difficulty; testing interval

Introduction

Neuropsychological assessments are often performed only once for diagnostic purposes. However, sometimes these assessments occur repeatedly in the same patient or subject, e.g. to monitor disease progression or therapeutic success. Practice effects refer to the impact of repeated exposure to an instrument on an examinee's performance. More specifically, practice effects due to repeated test administration can be defined as any improvement in accuracy or speed that is not attributable to changes in age, test version[a], or events between administrations (including state changes of the test-taker, such as recovery from mental illness). Test performance typically improves gradually (i.e., higher accuracy and faster speed) with repeated administrations (Beglinger et al., 2005). These practice effects therefore likely confound the true neuropsychological ability of interest, as they may mask worsening of a cognitive deficit or overestimate therapeutic successes. Importantly, practice effects typically do not generalize beyond the test instrument itself (Owen et al., 2010), and they are affected by several variables associated with the test, the testing situation, and the individual (see Duff (2012) for an overview).

Variables associated with the test include test reliability, especially test-retest reliability, the type of neuropsychological measure, its novelty, and floor and ceiling effects. Thus, potential adjustments for practice effects must be specific for each unique test. Variables associated with the testing situation include the retest interval and performance on initial evaluation (due to regression to the mean). Practice effects typically decrease with increasing retest intervals, but it is unclear whether, or after what duration, they are diminished, and what an optimal retest interval would be for any given practice curve (Falleti, Maruff, Collie, & Darby, 2006; Heilbronner et al., 2010). Variables associated with the individual include demographics (e.g., age, sex, and educational attainment), the clinical condition, and prior experience with similar cognitive tests.

Several statistical methods that correct for practice effects have been developed (Duff, 2012; Maassen, Bossema, & Brand, 2009). These include the Simple Discrepancy Score, the Standard Deviation Index, the Reliable Change Index (Jacobson & Truax, 1991), and standardized regression-based formula (McSweeny, Naugle, Chelune, & Lüders, 1993). Some of these methods rely on normative data from a population of interest, and they often only address situations with a single repeated administration. In general, studies that

[a] Different "versions" of a test refer to different stimulus sets. Another term that is frequently used for "version" in cognitive testing is "form".

investigated practice effects across more than two test administrations are rare (Falleti et al., 2006). This is a problematic gap, as several clinical, operational, and research settings require multiple test administrations. Studies that did investigate multiple test administrations often find curvilinear learning curves with a more rapid improvement of performance during early repeated administrations and a decelerating improvement during later test administrations, with test performance eventually reaching a plateau or asymptote (Beglinger et al., 2005; Schlegel, Shehab, Gilliland, Eddy, & Schiflett, 1995). In many clinical and research settings it is impractical to train subjects until they reach a performance plateau, nor is it possible to investigate a control group that could be used to adjust for some of the effects of repeated testing.

One strategy to reduce practice effects is the use of alternate forms (Benedict & Zgaljardic, 1998). Instead of always presenting identical stimuli, each form of the test presents similar but different stimuli that are unique for each version of the test. This methodology may reduce, but not eliminate practice effects. Optimally, alternative test versions share the same psychometric properties, especially difficulty. However, it can be challenging to develop alternative versions that are also equally difficult.

We recently developed a computerized neurocognitive test battery for the National Aeronautics and Space Administration (NASA) called *Cognition* (Basner et al., 2015; Lee et al., 2020; Moore et al., 2017). It is based on tests from the Penn Computerized Neurocognitive Battery (CNB) that were modified to reflect the high aptitude and motivation of astronauts (Gur et al., 2012; Gur et al., 2010; Moore, Reise, Gur, Hakonarson, & Gur, 2014). The 10 *Cognition* tests assess a range of cognitive domains, and the brain regions primarily recruited by each test have been previously established (Roalf et al., 2014) (the 10 tests are described in greater detail below and in Basner et al. (2015)). *Cognition* may be used to monitor astronaut behavioral health on long-duration exploration-type space missions, an isolated, confined, and extreme environment with several environmental and operational stressors, which requires regular and repeated testing before, during, and after the mission. For this reason, we developed 15 versions of the battery. Four out of the 10 tests (MP, LOT, DSST, and PVT; see below for meaning of test abbreviations) randomly generate stimuli while the tests are being performed. The remaining 6 tests (VOLT, F2B, AM, ERT, MRT, and BART) have stimuli that are unique[b] for each test version and always presented in the same order. The stimuli of the latter 6 tests are unique enough that, if administered repeatedly in the same way, subjects would recall them and the answer they provided during prior administration(s). In that sense, the tests would increasingly reflect memory rather than the specific construct targeted by the test.

This study had three main objectives:

1. Determine practice effects in both the accuracy and speed domain for each of the 10 Cognition tests across 15 administrations;

[b] When we call tests "unique" here and throughout the manuscript, we refer to the test stimuli and their order. The 15 versions of each test still have the same psychometric properties, i.e., they are based on the same rules, they come with the same instructions, and they target the same construct, but each with novel and different content.

2. Determine the effects of varying test-retest intervals (≥10 days, ≤5 days, 15 minutes) on practice effects; and
3. Determine differences in stimulus set difficulty between the 15 unique versions of *Cognition*.

The overarching goal of the study was to generate algorithms that would allow *Cognition* users to adjust their test data for practice effects and stimulus set difficulty effects.

Methods

Cognition Test Battery

A detailed description of the 10 *Cognition* tests can be found in Basner et al. (2015). Screenshots of the 10 tests can be found in supplementary Figure S1. Here, we provide a brief characterization of each test (modified from Basner et al. (2017)). The 10 tests were always performed in the order they are listed below.

The **Motor Praxis Task (MP)** (Gur et al., 2001) was 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 were instructed to click on squares that appear in random locations on the screen, with each successive square smaller and thus more difficult to track. Performance was assessed by the speed with which participants click each square.

The **Visual Object Learning Test (VOLT)** (David C Glahn, Gur, Ragland, Censits, & Gur, 1997) assessed participant memory for complex figures. Participants were asked to memorize 10 sequentially displayed three-dimensional shapes. Later, they were instructed to select the objects they memorized from a set of 20 such objects, also sequentially presented, half of which were from the learning set and half new.

The **Fractal 2-Back (F2B)** (Ragland et al., 2002) is a nonverbal variant of the Letter 2-Back. N-back tasks have become standard probes of the working memory system and activate canonical working memory brain areas. The F2B consisted of the sequential presentation of a set of figures (fractals), each potentially repeated multiple times. Participants were instructed to respond when the current stimulus matches the stimulus displayed two figures ago. The current implementation used 62 consecutive stimuli.

The **Abstract Matching (AM)** test (D. C. Glahn, Cannon, Gur, Ragland, & Gur, 2000) is a validated measure of the abstraction and flexibility components of executive function, including an ability to discern general rules from specific instances. The test paradigm presented subjects with two pairs of objects at the bottom left and right of the screen, varied on perceptual dimensions (e.g., color and shape). Subjects were presented with a target object in the upper middle of the screen that they had to classify as belonging more with one of the two pairs, based on a set of implicit, abstract rules. The current implementation used 30 consecutive stimuli.

The **Line Orientation Test (LOT)** (Benton, Varney, & Hamsher, 1978) is a measure of spatial orientation and derived from the well-validated Judgment of Line Orientation Test.

The LOT format consisted of presenting two lines at a time, one stationary while the other could be rotated by clicking an arrow. Participants could rotate the movable line until they perceived it to be parallel to the stationary line. The implementation used in this study had 12 consecutive line pairs that varied in length and orientation.

The **Emotion Recognition Task (ERT)** (Gur et al., 2010) is a measure of facial emotion recognition. It presented subjects with photographs of professional actors (adults of varying age and ethnicity) portraying emotional facial expressions of varying types and intensities (biased towards lower intensities, and with the prevalence of emotion categories balanced within each version of the test). Subjects were given a set of emotion labels (“happy”; “sad”; “angry”; “fearful”; and “no emotion”) and had to select the label that correctly described the expressed emotion. The implementation for the study used 40 consecutive stimuli, with 8 stimuli each representing one of the above 5 emotion categories.

The **Matrix Reasoning Test (MRT)** (Gur et al., 2001) is a measure of abstract reasoning and consists of increasingly difficult pattern matching tasks. It is analogous to Raven’s Progressive Matrices (Raven, 1965), a well-known measure of general intelligence. The test consisted of a series of patterns, overlaid on a grid. One element from the grid was missing and the participant had to select the element that fits the pattern from a set of several options. The implementation in the study used 12 consecutive stimuli.

The **Digit-Symbol Substitution Task (DSST)** (Usui et al., 2009) is a measure of complex scanning, visual tracking and working memory, and a computerized adaptation of a paradigm used in the Wechsler Adult Intelligence Scale (WAIS-III). The DSST required the participant to refer to a displayed legend relating each of the digits one through nine to specific symbols. One of the nine symbols appeared on the screen and the participant had to type the corresponding number using the keyboard as quickly as possible. The test duration was fixed at 90 s, and the legend key was randomly re-assigned with each administration.

The **Balloon Analog Risk Test (BART)** is a validated assessment of risk taking behavior (Lejuez et al., 2002). The BART required participants to either inflate an animated balloon or stop inflating and collect a reward. Participants were rewarded in proportion to the final size of each balloon, but a balloon popped after a hidden number of pumps, which changed across stimuli (Lejuez et al., 2002). The implementation in the study used 30 consecutive stimuli. The average tendency of balloons to pop was systematically varied between test administrations. This required subjects to adjust the level of risk based on the behavior of the balloons.

The **Psychomotor Vigilance Test (PVT)** is a validated measure of sustained attention based on reaction time (RT) to visual stimuli that occur at random inter-stimulus intervals (Basner & Dinges, 2011; Dinges et al., 1997). Subjects were instructed to monitor a box on the screen and press the space bar once a millisecond counter appeared in the box and started incrementing. The reaction time was then displayed for one second. Subjects were instructed to be as fast as possible without hitting the spacebar in the absence of a stimulus (i.e., false starts or errors of commission). The PVT is a sensitive measure of vigilant attention, and has been well-established as a tool to detect acute and chronic sleep deprivation and circadian misalignment, conditions highly prevalent in spaceflight (Barger et al., 2014). The PVT has negligible aptitude and learning effects (Basner & Dinges, 2011), and is ecologically relevant as sustained attention deficits and slow reactions affect many real-world tasks (e.g., operating a moving vehicle) (Dinges, 1995). In the current study, *Cognition* contained a validated 3-min. brief PVT-B with 2–5 s inter-stimulus intervals and a 355 millisecond lapse threshold (Basner, Mollicone, & Dinges, 2011).

Subjects and Protocol

A total of N=46 healthy adult subjects (23 male, mean±SD age for the entire sample 32.5±7.2 years, range 25–54 years) were recruited for the study through on campus and online postings. Subjects filled out a screening questionnaire prior to enrollment. They had to be proficient in English, between 25 and 55 years old, and have at least a Master's degree (or equivalent) to be comparable to high-performing astronauts for which Cognition was developed. Subjects with a history of a medical disorder that may significantly impact brain function (e.g. neurological disorders, severe head trauma) as well as subjects with a diagnosed psychiatric illness were excluded from study participation. The study was reviewed by the Institutional Review Board of the University of Pennsylvania and considered exempt. Written informed consent was obtained from study participants prior to data collection. Subjects were compensated for each session they attended. Depending on group assignment they received between $235 and $390 if they completed all test sessions.

Subjects performed the *Cognition* test battery for a total of 15 times in an office located in the Unit for Experimental Psychiatry at the University of Pennsylvania. The office door was kept closed to minimize noise and distraction. Before the first test administration, subjects watched a standardized familiarization video and were asked to perform a brief practice version of 8 out of the 10 *Cognition* tests before each of those tests for the first administration (the VOLT and BART have no practice versions). The practice versions were available before their respective tasks throughout the study, but were not required starting with the second battery administration. In the video, the principal investigator (MB) performed a short version of all 10 *Cognition* tests after reading out loud the standardized instructions for each test. During all test administrations, subjects were supervised by a research coordinator who could answer questions and address technical problems at any time. Subjects were not instructed to follow a specific sleep schedule. They were allowed to go about their normal lives. Subjects were free to choose the time of day for testing (i.e., they did not perform *Cognition* at the same time of day every time they were tested). However, 97.1% of testing sessions started between 8 am and 6 pm (earliest 7:10 am, latest 8:24 pm).

The 15 versions of Cognition were administered in a randomized but balanced fashion, i.e., each subject performed each version of Cognition exactly once, and each version of Cognition was administered in each administration order position exactly once. The randomization of administration order allowed us to disentangle practice effects from stimulus set difficulty effects. In an optimal study matrix, each battery would have been preceded by each other battery exactly once. However, there is no known solution for N>9 batteries (Archdeacon, Dinitz, Stinson, & Tillson, 1980). We therefore randomly generated

50,000 balanced matrices and selected the final study matrix based on how often a single battery was preceded by the same other battery. In the final matrix shown in Table 1, a battery was maximally preceded three times by the same other battery, and the number of instances where a battery was preceded by another battery more than once was minimized.

A secondary objective of the study was to test the effect of the duration of the interval between two *Cognition* administrations on practice effects. Our goal was to recruit 15 subjects that followed the test administration order outlined in Table 1 for each of the following three groups:

1. ≥10 days between test administrations (long group),
2. ≤5 days between test administrations (short group), or
3. four times on a single day with 15 minute breaks between test battery administrations for a total of 4 visits (ultrashort group).

In the short and ultrashort groups, there had to be at least one day without testing between study visits. In the ultrashort group, sessions were scheduled with 2–6 day intervals. Due to last minute scheduling conflicts, this interval had to be extended to seven days in two subjects in one instance each.

Overall, all 15 versions of *Cognition* were administered in the planned order in 93.3% of subjects. In the long interval group, one subject discontinued after the sixth *Cognition* administration. In another subject of the long group, *Cognition* version #10 was erroneously administered a second time during administration #11. The data of the six *Cognition* tests that have unique stimuli (see below) were excluded from data analysis for this participant. In the same subject, version #1 instead of #2 of *Cognition* was administered during administration #12. In the short group, the administration order of versions #6 and #11 was inverted in one subject. An additional subject with the correct administration order was recruited, increasing the sample size in the short group to N=16. We used block-randomization based on sex to assign subjects to one of three groups, which were comparable in age and sex composition (Table 2).

Measurement and Outcomes

Cognition was administered on a calibrated laptop computer (Dell Latitude E6430 with a 14” screen diagonal, 16:9 aspect ratio) using the *Cognition* software (version 2.6.0.4) (Basner et al., 2015). The average response latency of the spacebar and mouse button was determined with a robotic calibrator before the start of the study and subtracted from each response time. After completion of each of the 10 tests, subjects were presented with a feedback score ranging between 0 (worst possible performance) and 1000 (best possible performance). Depending on the test, the feedback score was based on accuracy, speed, or both. It took subjects on average 19.2 minutes (SD 2.9 minutes) to complete all 10 tests.

Our analyses concentrated on one speed and one accuracy outcome for each *Cognition* test. All accuracy outcomes ranged from 0% to 100% with 100% representing best possible performance. For all speed outcomes, lower values reflect shorter response times and thus higher speed. Average response time [milliseconds] was the speed outcome for all tests but the PVT. In the latter, 10 minus reciprocal response time [1/RT] was used as the speed outcome, as this metric was shown to be a superior outcome for the PVT relative to average RT (Basner & Dinges, 2011). Percent correct was the accuracy outcome for five *Cognition* tests. For the MP, the distance from the center of each square (in pixels) was averaged across all responses. The center of the square translates to 100% accuracy, 50 pixels or more away from the center translate to an accuracy score of 0%, with linear scaling between 0 and 50 pixels. For the F2B, we averaged the percent correct for matches and non-matches to avoid subjects achieving good accuracy scores even if they never hit the spacebar. For the LOT, the accuracy measure was calculated as 3 minus the average number of clicks off, which was then divided by 3 (subjects are on average 0.8 clicks off). For tests with more than 3 clicks off on average, the accuracy score was set to 0%. For the ERT, stimuli that loaded negatively in an Item Response Theory (IRT) analysis were excluded for the calculation of both speed and accuracy on the ERT (see Supplementary Tables S1 for a list of excluded stimuli). Likewise, the following stimuli were excluded for the calculation of both speed and accuracy on the MRT due to negative IRT loadings: battery #3 stimulus #11; battery #4 stimulus #12; battery #7 stimulus #12; battery #11 stimulus #6; battery #14 stimulus #11; and battery #15 stimulus #11. For each pump on the BART, a value of 1 divided by the number of possible pumps across all 30 balloons was added to the Risk Score. This Risk Score therefore takes into account that different sets of balloons popped at different inflation rates. Here, we list BART risk-taking as an “accuracy” outcome despite the fact that it inherently measures risk taking. For the PVT, the accuracy score was calculated as 1 - ((# of Lapses + # of False Starts) / (Number of Stimuli + # of False Starts)). Any response time not falling in-between the false start threshold (100 ms) and the lapse threshold (355 ms) thus decreased accuracy on the PVT.

Data analyses

All Cognition tests were inspected for subject non-adherence, but none of the tests needed to be excluded from data analysis. All data were analyzed with linear mixed effects models in SAS (Version 9.3, Cary, NC). Degrees of freedom were corrected with Satterthwaite’s method (Satterthwaite, 1946). Test statistics were considered significant at p<0.05.

Practice and Administration Interval Effects

Administration number, or the logarithmic transform (base e) of administration number, were entered as a continuous variable in a random subject intercept model with random slopes (unstructured covariance). Because of the data loss described above, the study matrix was not completely balanced, and therefore battery version was also included as a categorical variable in the model. The Akaike Information Criterion (AIC) was used to determine whether linear or logarithmic administration number fit the data better (Bozdogan, 1987). As the logarithmic fit outperformed the linear fit in every instance, we only report results on the logarithmic fit here. To investigate the effects of administration interval on practice effects, a categorical variable for administration interval (long, short, ultra-short) was added to the model together with an interaction term with administration number (logarithmic transform).

Stimulus Set Effects

Cognition version was entered as a class variable in a linear mixed effect model with random subject intercept. Because of the data loss described above, the study matrix was not completely balanced, and therefore administration number was also included as a categorical variable in the model. Based on the Type-III test of fixed effects for battery version it was determined whether stimulus sets differed significantly.

Outlier Handling

One of the main purposes of this study was to create estimates that would allow for correcting test data for practice and stimulus set effects. We wanted to make sure that individual subjects did not unduly influence these corrections, without being too lenient in excluding subjects. For each model, we therefore repeated the analyses described above 46 times, each time leaving out one of the 46 subjects. We then calculated the mean and standard deviation for each of the 46 estimates, and excluded from our final models subjects (for practice effects) or individual test bouts (for stimulus set effects) who caused the estimate to deviate >4 standard deviations from the mean across all estimates. For practice effect corrections, one subject was excluded for ERT and LOT speed, another subject was excluded for AM speed, and a third subject was excluded for MP accuracy, representing 0.4% of the input data. For stimulus set effects, 16 tests representing 0.2% of the input data were excluded. All final models and corrections were based on datasets excluding these outliers.

Results

Repeated-measures within-subject correlations (Bland & Altman, 1995) and cross-sectional Pearson correlations for the 10 Cognition tests are shown in Supplementary Table S2.

Practice and Administration Interval Effects

Practice effects for the 10 Cognition tests are visualized in Figure 1 for the speed domain and in Figure 2 for the accuracy domain. Except for the PVT, all tests demonstrated a significant practice effect in the speed domain (all P<0.02 for logarithmic slope; Table 3). Subjects were getting faster on all of these tests with repeated administrations. Based on standardized estimates, practice effects in the speed domain were most pronounced for the BART, ERT, DSST, and VOLT. Only 6 out of the 10 tests demonstrated a significant practice effect in the accuracy domain (MP, VOLT, F2B, AM, MRT, and PVT). With the exception of MP, subjects were getting more accurate with repeated administrations. Based on standardized estimates, practice effects in the accuracy domain were most pronounced for AM, F2B, and VOLT. As expected, practice effects were more pronounced during the first few administrations but continued well beyond the first few administrations for several tests (Figures 1 and 2).

Practice effects were modified by administration interval for speed outcomes on the MP (p<0.01), AM (p=0.03), and ERT (p=0.01; Figure 3) only. Further analyses showed that the slope differed significantly for all 3 groups (long, short, and ultra-short) on the MP, whereas it did not differ for the short and ultrashort groups on the AM (p=0.76) and the ERT (p=0.16). None of the practice effects observed for accuracy outcomes was modified by administration interval (all p>0.05; Figure 4).

Stimulus Set Effects

Stimulus sets differed significantly in the speed domain for VOLT, ERT, MRT, and BART (Figure 5) and in the accuracy domain for all six *Cognition* tests with version-unique stimulus sets: VOLT, F2B, AM, ERT, MRT, and BART (Figure 6; Table 2). Standardized estimates for the difference of individual test version scores relative to scores across all versions are provided in supplementary Table S3. For those tests that do not have fixed stimulus sets, none of the individual test versions differed significantly from the mean across versions in both the speed and accuracy domain. According to conventional standards (Cohen, 1988), effect sizes were <0.5 for all tests in the speed and accuracy domain, except for BART and ERT accuracy, where individual tests differed by >0.5 SD (BART) and >0.8 SD (ERT) from the overall mean across versions, respectively.

Correction for Practice and Stimulus Set Effects

Information from the regression models was used to generate correction factors for practice and stimulus set effects. Correction factors were generated for tests with significant (p<0.05) administration number or stimulus set effects (see Supplementary Tables S4–S7 for correction factors). We also provide practice effect correction factors for accuracy on the ERT, as this test just missed statistical significance (p=0.07). Of note, the (arbitrary) p<0.05 cut-off for generating correction factors was of little relevance, as the data nicely separated tests requiring or not requiring correction. The tests with the lowest p-value that did not meet the p<0.05 criterion had p-values of 0.44 (practice/speed), 0.37 (practice/accuracy), 0.16 (version/speed), and 0.35 (version/accuracy). For MP, AM, and ERT speed, separate administration number correction factors were produced for the long, short, and ultra-short groups (MP only), and for the long and short/ultra-short groups (AM and ERT). Administration number corrections were expressed relative to administration #15. This typically meant that earlier administrations received a response time and accuracy bonus, i.e., response time was shortened and accuracy was increased according to the regression models.

A model containing administration number and battery version as factors was the basis for correcting for stimulus set differences. We used effect coding for battery version, i.e., the estimate for each battery version reflected the difference from the overall mean across all batteries. These estimates were then directly used for correction purposes. For example, if the estimate for VOLT accuracy for battery version 4 indicated that subjects were on average 3% less accurate, these 3% should be added to accuracy scores for battery version #4 to adjust for stimulus set differences. When administration number and battery version corrections are applied, accuracy needs to be restricted to a range between 0% and 100% (reflecting worst and best possible performance).

Discussion

This study investigated practice, administration interval, and stimulus set effects of the *Cognition* test battery, which was specifically designed for repeated administration in high-performing individuals with 15 unique versions available. The study adds to the small body of literature investigating practice and stimulus set effects beyond the typical 2 administrations (Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010; Falleti et al., 2006; Schlegel et al., 1995; Wilson, Watson, Baddeley, Emslie, & Evans, 2000). That each subject performed all 15 versions of the Cognition battery in a randomized yet balanced fashion allowed for disentangling the otherwise confounded practice and stimulus set effects.

We found significant practice effects for all tests but the PVT (Basner et al., 2018) in the speed domain and for 6 out of the 10 tests in the accuracy domain. Subjects became faster with repeated administration on all tests, and they became more accurate with repeated administration on all tests but the MP. The MP requires subjects to click into a square that is getting progressively smaller and changing positions from trial to trial. Although subjects are instructed to "click each square as quickly and accurately" as they can, they are not explicitly instructed to click in the center of the square. The findings therefore suggest that, with repeated administration, subjects increasingly adopt a strategy to click in the periphery of the square to increase speed on the MP task.

Notably, standardized practice effects were larger for BART, ERT, DSST, and VOLT (ES ≥0.31) in the speed domain compared to the other six tests (ES≤0.19). Likewise, standardized effects were larger for F2B and AM (ES ≥0.30) in the accuracy domain compared to the other eight tests (ES≤0.18). Thus, effect sizes are small to negligible for most accuracy domains while notable for four speed domains. That effects were stronger for speed than accuracy would suggest caution in using speed measures as primary outcome in treatment studies, unless a control group is included with the same number of repeated measures or correction algorithms as the ones derived in this study are available. Concerning accuracy, repeated measurement affects most strongly executive tasks, abstraction and mental flexibility and working memory, which implicate prefrontal and frontal regions and were shown to activate these regions in functional neuroimaging (Ragland et al., 2002; Roalf et al., 2014). This finding is consistent with evidence from functional neuroimaging that frontal regions are fastest to habituate to repeated measurements (Warach et al., 1992). It is noteworthy that the executive system, whose role is to adjust to situational challenges, is the one who benefits the most from repeated exposure to the testing situation.

As expected, practice effects were typically most pronounced during the first few administrations and then started leveling off. In contrast to conclusions drawn by Falleti et al. (2006), our data do not support a practice effect restricted to the first re-administration. Rather, and depending both on the test and the outcome domain (speed/accuracy), we observed protracted practice effects sometimes extending to the 15th test administration. This is in line with Schlegel et al. (1995) who found that "performance improved rapidly over the first three to five trials", "the rate of improvement leveled off by the eighth trial", and a minimum of 15 required administrations before stable performance levels were reached. However, in the vast majority of research and clinical contexts, it will not be possible to train subjects this extensively. The correction factors derived in this study can be used to adjust for practice effects and help shorten practice schedules. We typically use the first test administration to familiarize subjects with the requirements of each test, which could be considered a minimal practice requirement. The findings also have clinical implications as they demonstrate that practice effects between the first and second administration of a test are profound and continue well beyond the second administration. Any practice effects observed in this study will likely be even more pronounced in clinical batteries that do not have alternate versions.

Practice effects were significantly affected by the test-retest interval for only three tests (MP, AM, ERT) and exclusively in the speed domain. Practice effects were less pronounced for longer test-retest intervals, and they vanished for intervals ≥10 days for the MP and AM. This suggests that for the majority of the 10 *Cognition* tests and for accuracy outcomes on all tests, test-retest intervals do not relevantly affect practice effects, at least for test-retest intervals of up to 12 days on average.

We saw significant differences in stimulus set difficulty across the 15 versions of *Cognition* for 4 tests in the speed domain and for six tests in the accuracy domain. That the latter 6 tests are those that use unique (in contrast to randomly generated) stimulus sets lends face validity to the design and analytic approach of this study. The correction factors established in this study allow for correcting for these differences in stimulus sets.

Effect sizes for test score differences of individual versions relative to the mean across versions were typically negligible to small, and reached medium to large levels only for the BART and ERT in the accuracy domain. The stimuli of the ERT are based on professional actors who were asked to express the five ERT emotions with different degrees of intensity. Versions with 40 stimuli each were generated to reflect a balance of the five emotions (8 stimuli each), the degree of expression, as well as race and sex. Based on the findings of this study, we performed additional analyses based on IRT to better characterize individual stimuli, and identify those that were too easy or not able to differentiate good from bad performers. The results of these analyses were recently used to generated alternate versions of the ERT with similar properties that only use 20 instead of 40 stimuli. During development of the VOLT, stimulus set difficulty was informed by crowd-sourcing, and, while not perfect, VOLT stimulus sets differed less strongly compared to the ERT. These are two examples of how better characterization of individual stimuli can help in generating test versions that are more comparable.

For the BART, 15 sets of 30 balloons with specific properties (i.e., how soon the balloons pop on average) were generated. Subjects were tasked with adjusting their level of risk taking based on the observed behavior of the balloons. As evidenced by differences in risk taking among BART versions, subjects were not able to fully implement this strategy. The corrections provided in the Supplement can be used to adjust for these differences.

Finally, clinical or research neuropsychologists who use standard neuropsychological tests such as described in (Lezak, Howieson, Bigler, & Tranel, 2012), may wonder whether the present study has any relevance to their present or future work. These neuropsychological tests are well-established and based on solid foundations of normative and clinical data on multiple populations of patients with focal and diffuse brain disorders. Traditional neuropsychological tests, unfortunately, are time-consuming and require highly trained professionals for administration, scoring and interpretation. These features make such assessments not feasible in situations such as exploration-type spaceflight missions, to which Cognition has been designed. The same features also make traditional batteries unsuitable for large-scale population-based studies or clinical settings in Low and Middle Income Countries (LMIC). Increasingly in large scale genomic and clinical studies they are being replaced with computerized batteries with automated scoring algorithms that are not only more efficient but also offer critical diagnostic information and treatment targets. Batteries such as Cognition and its parent battery PennCNB (Gur et al., 2001; Gur et al., 2010) are based on cognitive neuroscience approaches and validated with functional neuroimaging (Roalf et al., 2014) and may eventually replace current batteries (Roalf & Gur, 2017). From this perspective, the current study offers a glimpse into what can be done with this novel approach and how it can be integrated into a routine assessment process that can track deficits and their resolution with treatment.

Strengths and Limitations

Strengths of the study include the relatively large sample size for a study that required repeated testing across several weeks or months, and the study design that allowed for disentangling practice and stimulus set effects. Potential limitations include the following: We did not fix time of day across *Cognition* administrations within subjects. This potentially increased variability across test administrations, although cognitive performance is typically relatively stable across the first 16 hours of the wake period in adults (Basner et al., 2011). Furthermore, we investigated a high performing population. It is unclear whether our findings translate to other populations with lower degrees of educational attainment or lower levels of motivation, or clinical populations with brain disorders or dysfunction. However, the usefulness of the Cognition battery extends beyond astronauts to other high-performing populations (e.g., military personnel, physicians) that typically outperform the general population and may appear normal on standard test batteries even if cognitively impaired relative to their ability level. A third limitation is that subjects were presented with a feedback score at the end of each test, and it is unclear whether this feature influenced our findings. Finally, the study was not powered to investigate age or sex effects on practice or stimulus set effects, potentially resulting in inaccuracies of the correction factors established in this study for specific age or sex groups.

Conclusions

This study investigated practice, test-retest administration interval, and stimulus set effects for the *Cognition* test battery. Protracted practice effects well beyond the second administration were observed for most of the 10 tests both in the accuracy and speed domain, but test-retest administration interval significantly affected practice effects only for three out of the 10 tests and only in the speed domain. Stimulus set effects were observed for the six *Cognition* tests that use unique sets of stimuli. Factors were established that allow for correcting for both practice and stimulus set effects. This is a unique feature of the *Cognition* battery that can help avoid masking of practice effects, address noise generated by differences in stimulus set difficulty, and help interpret results of studies with inadequate controls or short practice schedules.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the National Space Biomedical Research Institute (NSBRI) through grant NBPF00012 (NASA NCC 9-58) and by the National Aeronautics and Space Administration (NASA) through grant NNX14AH98G.

References

Archdeacon DS, Dinitz JH, Stinson DR, & Tillson TW (1980). Some new row-complete Latin squares. Journal of Combinatorial Theory, 29(Series A), 395–398.

Barger LK, Flynn-Evans EE, Kubey A, Walsh L, Ronda JM, Wang W, … Czeisler CA (2014). Prevalence of sleep deficiency and hypnotic use among astronauts before, during and after spaceflight: an observational study. Lancet Neurology, 13(9), 904–912. [PubMed: 25127232]

Bartels C, Wegrzyn M, Wiedl A, Ackermann V, & Ehrenreich H (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. BMC Neuroscience, 11, 118. doi:10.1186/1471-2202-11-118 [PubMed: 20846444]

Basner M, & Dinges DF (2011). Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep, 34(5), 581–591. [PubMed: 21532951]

Basner M, Hermosillo E, Nasrini J, McGuire S, Saxena S, Moore TM, … Dinges DF. (2018). Repeated Administration Effects on Psychomotor Vigilance Test Performance. Sleep, 41(1), zsx187: 181–186 doi:10.1093/sleep/zsx187

Basner M, Mollicone DJ, & Dinges DF (2011). Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. Acta Astronautica, 69, 949–959. [PubMed: 22025811]

Basner M, Nasrini J, Hermosillo E, McGuire S, Dinges DF, Moore TM, … Bershad EM. (2017). Effects of −12° head-down tilt with and without elevated Levels of CO2 on cognitive performance: the SPACECOT study. Journal of Applied Physiology, 124(3), 750–760. [PubMed: 29357516]

Basner M, Savitt A, Moore TM, Port AM, McGuire S, Ecker AJ, … Gur RC. (2015). Development and validation of the Cognition test battery Aerospace Medicine and Human Performance, 86(11), 942–952. [PubMed: 26564759]

Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, … Siemers ER. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. Archives of Clinical Neuropsychology, 20(4), 517–529. doi:10.1016/j.acn.2004.12.003 [PubMed: 15896564]

Benedict RH, & Zgaljardic DJ (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. Journal of Clinical and Experimental Neuropsychology, 20(3), 339–352. doi:10.1076/jcen.20.3.339.822 [PubMed: 9845161]

Benjamini Y, & Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300.

Benton AL, Varney NR, & Hamsher KD (1978). Visuospatial judgment. A clinical test. Archives of Neurology, 35(6), 364–367. [PubMed: 655909]

Bland JM, & Altman DG (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet, 346(8982), 1085–1087. doi:S0140-6736(95)91748-9 [pii] [PubMed: 7564793]

Bozdogan H (1987). Model Selection and Akaike Information Criterion (AIC) - the general-theory and its analytical extensions. Psychometrika, 52(3), 345–370. doi:10.1007/bf02294361

Cohen J (1988). Statistical power analysis for the behavioral sciences (2nd edition ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dinges DF (1995). An overview of sleepiness and accidents. J.Sleep Res, 4(S2), 4–14. [PubMed: 10607205]

Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, … Pack AI. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. Sleep, 20(4), 267–277. [PubMed: 9231952]

Duff K (2012). Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. Archives of Clinical Neuropsychology, 27(3), 248–261. doi:10.1093/arclin/acr120 [PubMed: 22382384]

Falleti MG, Maruff P, Collie A, & Darby DG (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. Journal of Clinical and Experimental Neuropsychology, 28(7), 1095–1112. doi:10.1080/13803390500205718 [PubMed: 16840238]

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

Glahn DC, Gur RC, Ragland JD, Censits DM, & Gur RE (1997). Reliability, performance characteristics, construct validity, and an initial clinical application of a visual object learning test (VOLT). Neuropsychology, 11(4), 602. [PubMed: 9345704]

Gur RC, Ragland JD, Moberg PJ, Turner TH, Bilker WB, Kohler C, … Gur RE. (2001). Computerized neurocognitive scanning: I. Methodology and validation in healthy people. Neuropsychopharmacology, 25(5), 766–776. doi:10.1016/S0893-133X(01)00278-0 [PubMed: 11682260]

Gur RC, Richard J, Calkins ME, Chiavacci R, Hansen JA, Bilker WB, … Gur RE. (2012). Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. Neuropsychology, 26(2), 251–265. doi:10.1037/a0026712 [PubMed: 22251308]

Gur RC, Richard J, Hughett P, Calkins ME, Macy L, Bilker WB, … Gur RE. (2010). A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. Journal of Neuroscience Methods, 187(2), 254–262. doi:10.1016/j.jneumeth.2009.11.017 [PubMed: 19945485]

Heilbronner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, & Hart RP (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: the utility and challenges of repeat test administrations in clinical and forensic contexts. Clinical Neuropsychologist, 24(8), 1267–1278. doi:10.1080/13854046.2010.526785 [PubMed: 21108148]

Jacobson NS, & Truax P (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59(1), 12–19. [PubMed: 2002127]

Lee G, Moore TM, Basner M, Nasrini J, Roalf DR, Ruparel K, … Gur RC. (2020). Age, Sex, and Repeated Measures Effects on NASA’s “Cognition” Test Battery in STEM Educated Adults. Aerosp Med Hum Perform, 91(1), 18–25. doi:10.3357/AMHP.5485.2020 [PubMed: 31852569]

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

Lezak MD, Howieson DB, Bigler ED, & Tranel D (2012). Neuropsychological assessment, 5th ed. New York, NY, US: Oxford University Press.

Maassen GH, Bossema E, & Brand N (2009). Reliable change and practice effects: outcomes of various indices compared. Journal of Clinical and Experimental Neuropsychology, 31(3), 339–352. doi:10.1080/13803390802169059 [PubMed: 18618359]

McSweeny AJ, Naugle RI, Chelune GJ, & Lüders H (1993). “TScores for Change”: An illustration of a regression approach to depicting change in clinical neuropsychology. Clinical Neuropsychologist, 7(3), 300–312. doi:10.1080/13854049308401901

Moore TM, Basner M, Nasrini J, Hermosillo E, Kabadi S, Roalf DR, … Gur RC. (2017). Validation of the Cognition Test Battery for Spaceflight in a Sample of Highly Educated Adults. Aerosp Med Hum Perform, 88(10), 937–946. doi:10.3357/AMHP.4801.2017 [PubMed: 28923143]

Moore TM, Reise SP, Gur RE, Hakonarson H, & Gur RC (2014). Psychometric Properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. doi:10.1037/neu0000093

Owen AM, Hampshire A, Grahn JA, Stenton R, Dajani S, Burns AS, … Ballard CG. (2010). Putting brain training to the test. Nature, 465(7299), 775–778. doi:10.1038/nature09042 [PubMed: 20407435]

Ragland JD, Turetsky BI, Gur RC, Gunning-Dixon F, Turner T, Schroeder L, … Gur RE (2002). Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. Neuropsychology, 16(3), 370–379. [PubMed: 12146684]

Raven JC (1965). Advanced Progressive Matrices: Sets I and II. London: Lewis.

Roalf DR, & Gur RC (2017). Functional brain imaging in neuropsychology over the past 25 years. Neuropsychology, 31(8), 954–971. doi:10.1037/neu0000426 [PubMed: 29376672]

Roalf DR, Ruparel K, Gur RE, Bilker W, Gerraty R, Elliott MA, … Gur RC (2014). Neuroimaging predictors of cognitive performance across a standardized neurocognitive battery. Neuropsychology, 28(2), 161–176. doi:10.1037/neu0000011 [PubMed: 24364396]

Satterthwaite FE (1946). An approximate distribution of estimates of variance components. Biometrics, 2(6), 110–114. [PubMed: 20287815]

Schlegel RE, Shehab RL, Gilliland K, Eddy DR, & Schiflett SG (1995). Microgravity effects on cognitive performance measures: practice schedules to acquire and maintain performance stability (AL/CF-TR-1994-0040). Retrieved from

Usui N, Haji T, Maruyama M, Katsuyama N, Uchida S, Hozawa A, … Taira M (2009). Cortical areas related to performance of WAIS Digit Symbol Test: a functional imaging study. Neuroscience Letters, 463(1), 1–5. doi:10.1016/j.neulet.2009.07.048 [PubMed: 19631255]

Warach S, Gur RC, Gur RE, Skolnick BE, Obrist WD, & Reivich M (1992). Decreases in frontal and parietal lobe regional cerebral blood flow related to habituation. Journal of Cerebral Blood Flow and Metabolism, 12(4), 546–553. doi:10.1038/jcbfm.1992.78 [PubMed: 1618933]

Wilson BA, Watson PC, Baddeley AD, Emslie H, & Evans JJ (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. Journal of the International Neuropsychological Society, 6(4), 469–479. [PubMed: 10902416]

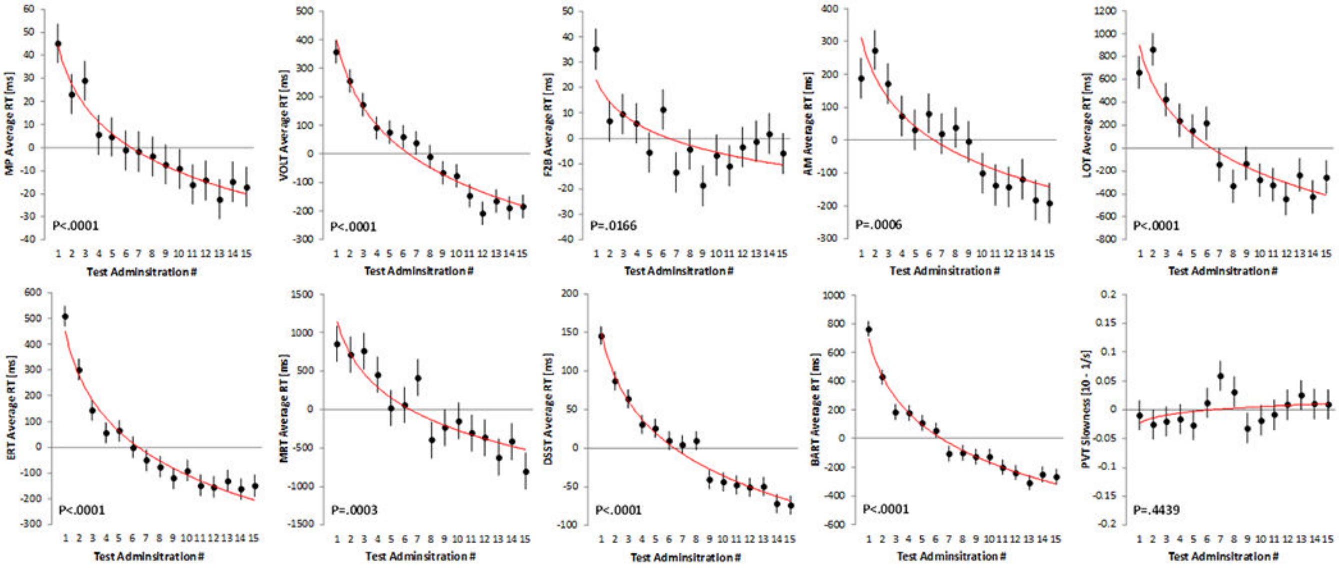

Figure 1: Changes in Cognition speed outcomes with repeated test administration.

Each subject performed the Cognition battery 15 times. A significant logarithmic trend was found for all tests but the Psychomotor Vigilance Test (PVT). Black dots represent estimated means relative to the overall mean across test administrations. Error bars reflect standard errors. The red line reflects a logarithmic trend line fitted to the estimated means. MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test

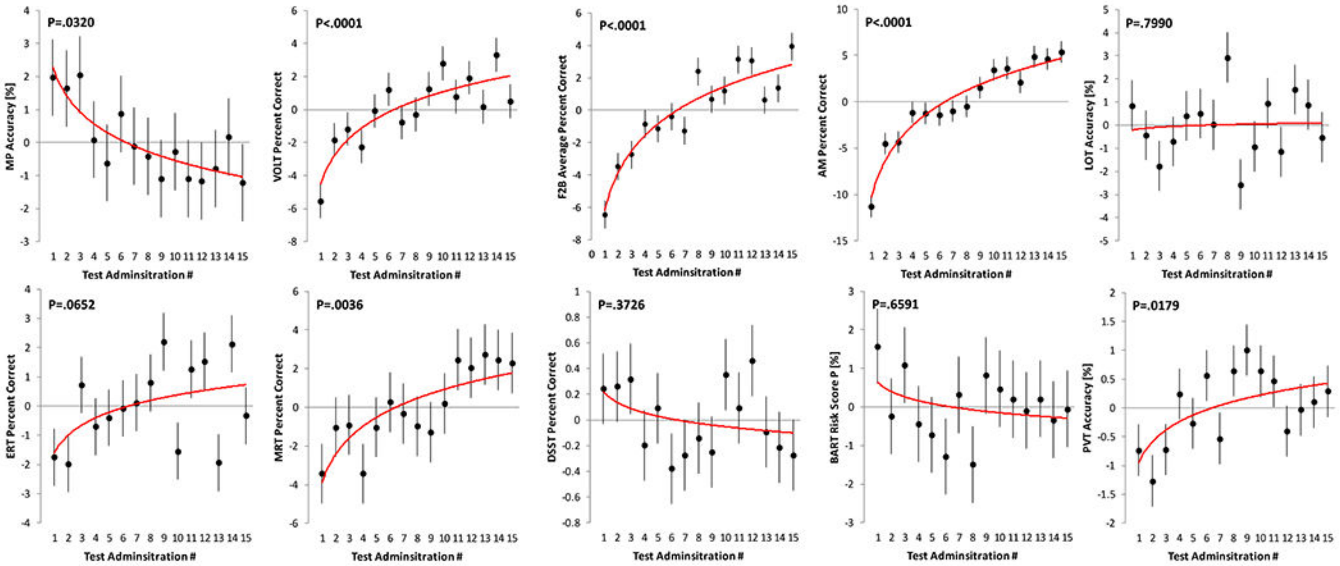

Figure 2: Changes in Cognition accuracy outcomes with repeated test administration.

Each subject performed the Cognition battery 15 times. A significant logarithmic trend was found for 6 out of the 10 tests. Black dots represent estimated means relative to the overall mean across test administrations. Error bars reflect standard errors. The red line reflects a logarithmic trend line fitted to the estimated means. MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test

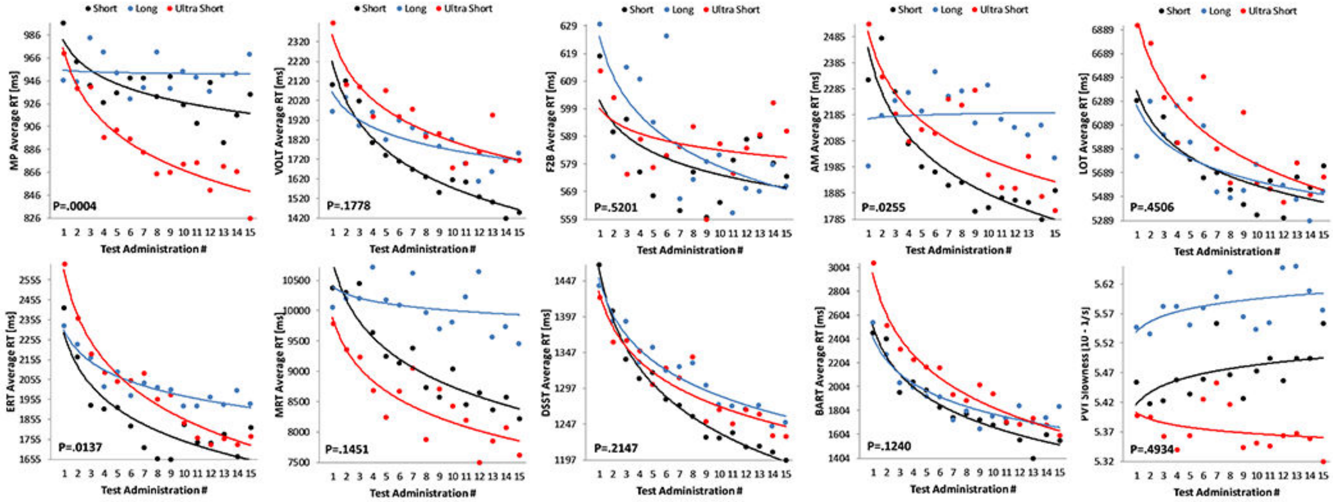

Figure 3: Effect of test-retest administration interval on practice effects in the speed domain.

Each subject performed the Cognition battery 15 times with different test-retest intervals (Long (blue): ≥10 days; Short (black): ≤5 days; Ultrashort (red): four times per day with 15 minute breaks between tests). MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test

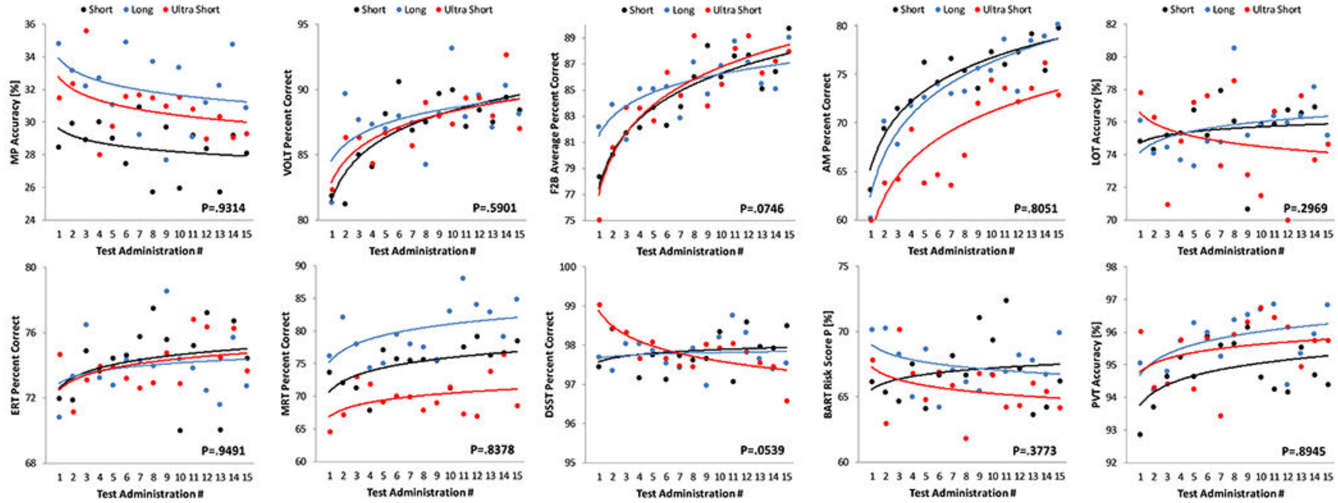

Figure 4: Effect of test-retest administration interval on practice effects in the accuracy domain.

Each subject performed the Cognition battery 15 times with different test-retest intervals (Long (blue): ≥10 days; Short (black): ≤5 days; Ultrashort (red): four times per day with 15 minute breaks between tests). MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test

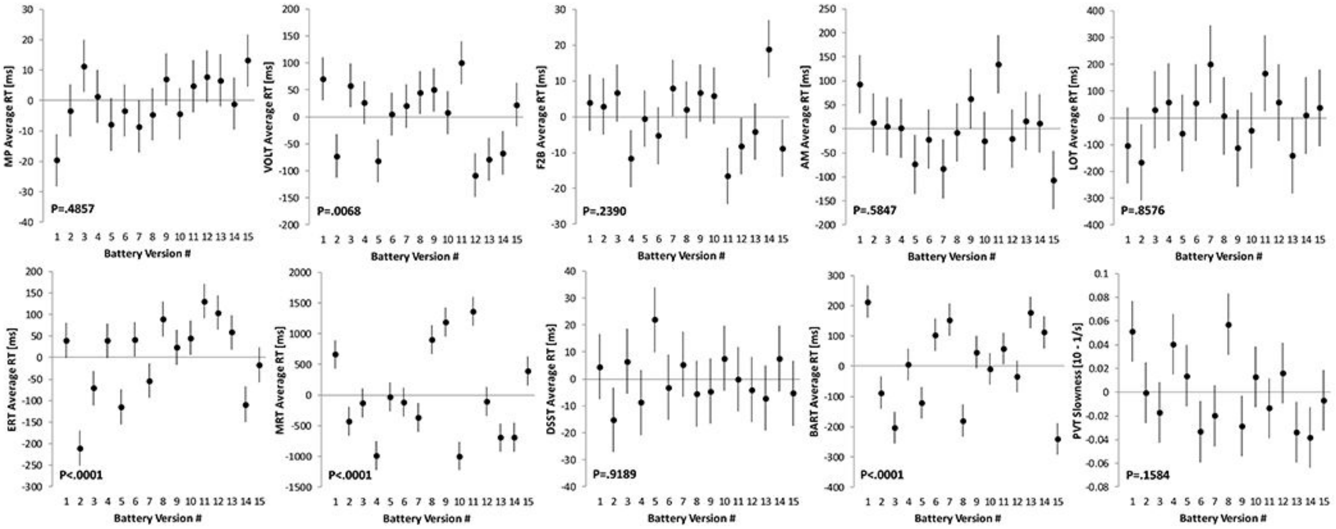

Figure 5: Differences between stimulus sets in the speed domain.

Significant differences in Cognition speed performance across the 15 test versions were found for four out of the 10 Cognition tests. Black dots represent estimated means relative to the overall mean across test administrations. Error bars reflect standard errors. MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test

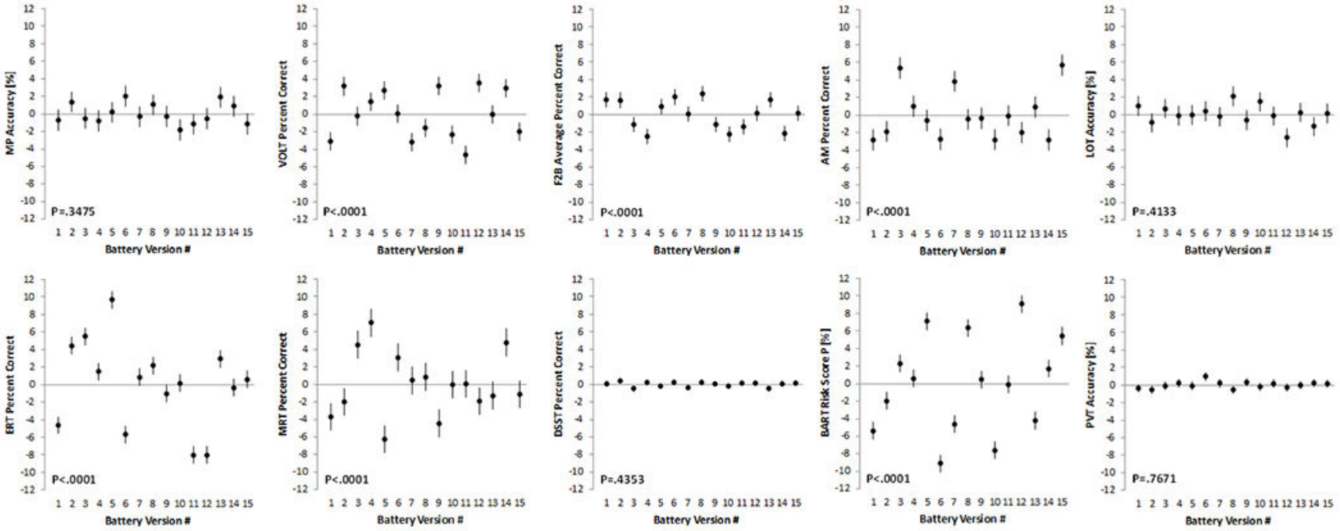

Figure 6: Differences between stimulus sets in the accuracy domain.

Significant differences in Cognition accuracy performance across the 15 test versions were found for those six out of the 10 Cognition tests that have unique stimulus sets. Black dots represent estimated means relative to the overall mean across test administrations. Error bars reflect standard errors. MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test

Table 1:

Study matrix showing the administration order for each of the 15 versions of the Cognition test battery for each of the 15 subjects of one administration interval group. Each cell in the matrix reflects the version of the Cognition test that was administered. The order of the battery versions was randomized but balanced (i.e., across the 15 subjects, each test version was administered in each position exactly once).

	Test Administration Number														
Subject #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	12	6	5	15	9	7	14	11	2	13	4	8	10	1	3
2	6	8	1	4	11	3	2	5	12	14	15	10	13	7	9
3	15	7	6	3	14	12	10	13	8	9	1	2	11	5	4
4	5	14	3	12	13	2	9	6	7	10	8	4	15	11	1
5	13	2	10	8	15	1	11	12	9	4	7	5	3	14	6
6	9	13	11	14	4	5	6	8	1	15	12	3	7	2	10
7	14	12	9	7	3	4	15	1	5	11	2	13	6	10	8
8	8	4	15	13	12	11	5	10	14	1	6	9	2	3	7
9	10	11	12	2	1	13	7	4	3	8	14	6	9	15	5
10	11	10	2	9	8	6	1	3	4	12	5	7	14	13	15
11	7	15	4	6	2	8	13	14	10	5	3	12	1	9	11
12	3	9	14	5	10	15	8	7	13	6	11	1	12	4	2
13	2	1	8	10	7	14	4	9	11	3	13	15	5	6	12
14	1	3	7	11	5	10	12	15	6	2	9	14	4	8	13
15	4	5	13	1	6	9	3	2	15	7	10	11	8	12	14

Table 2:

Subject Characteristics and Administration Interval Statistics[*]

	Long Group ≥ 10 days	Short Group ≤ 5 days	Ultrashort Group 4x per day	All	Difference Between Groups p-value
Number of Subjects	15	16	15	46	N/A
Mean (Range) Administration Interval [days]	12.0 (10-28)	3.1 (2-5)	3.5 (2-7)	6.9 (2-28)	N/A
Mean (SD) Subject Age [years]	32.1 (7.3)	34.4 (6.9)	30.9 (7.5)	32.5 (7.2)	0.38[#]
Male	53.3%	50.0%	46.7%	50.0%	1.00[†]
Ethnicity					
White	66.7%	37.5%	60.0%	54.3%	
Asian	26.7%	37.5%	6.7%	23.9%	0.13[†]
Black	0.0%	12.5%	26.7%	13.0%	
Other/Unknown	6.7%	12.5%	6.7%	8.7%	
Degree MD or PhD	13.3%	12.5%	20.0%	15.2%	0.88[†]

SD: Standard Deviation;

[*] consecutive test days had to be separated by at least one test-free day;

[†] Fisher’s Exact test;

[#] Type-III test for fixed-effects; N/A not applicable

Table 3:

Practice, administration interval, and stimulus set difficulty effects

		Log$_e$(Admin) Slope			Log$_e$(Admin) by Administration Interval Interaction	Stimulus Set Effect
	Outcome	**β (SE)**	**Std. β (SE)**	**P-value**	**P-value**	**P-value**
SPEED	MP Average RT [ms]	−23.8 (4.8)	−0.14 (0.03)	<.01*	<.01*	0.49
	VOLT Average RT [ms]	−216.4 (32.5)	−0.31 (0.05)	<.01*	0.18	<.01*
	F2B Average RT [ms]	−12.8 (5.1)	−0.10 (0.04)	0.017*	0.52	0.24
	AM Average RT [ms]	−153.1 (41.4)	−0.19 (0.05)	<0.01*	0.03	0.58
	LOT Average RT [ms]	−383.7 (87.3)	−0.19 (0.04)	<.01*	0.45	0.86
	ERT Average RT [ms]	−237.5 (24.5)	−0.37 (0.04)	<.01*	0.01	<.01*
	MRT Average RT [ms]	−613.7 (156.4)	−0.15 (0.04)	<.01*	0.15	<.01*
	DSST Average RT [ms]	−80.2 (7.9)	−0.32 (0.03)	<.01*	0.21	0.92
	BART Average RT [ms]	−378.1 (40.3)	−0.38 (0.04)	<.01*	0.12	<.01*
	PVT Slowness [10 - 1/s]	0.0126 (0.0163)	0.02 (0.03)	0.44	0.49	0.16
ACCURACY	MP Accuracy [%]	−0.86 (0.40)	−0.06 (0.03)	0.03*	0.93	0.35
	VOLT Accuracy [%]	2.42 (0.44)	0.18 (0.03)	<.01*	0.59	<.01*
	F2B Accuracy [%]	3.34 (0.41)	0.30 (0.04)	<.01*	0.07	<.01*
	AM Accuracy [%]	5.51 (0.62)	0.34 (0.04)	<.01*	0.81	<.01*
	LOT Accuracy [%]	0.12 (0.46)	0.01 (0.03)	0.80	0.30	0.41
	ERT Accuracy [%]	0.87 (0.46)	0.07 (0.04)	0.07	0.95	<.01*
	MRT Accuracy [%]	2.11 (0.68)	0.10 (0.03)	<.01*	0.84	<.01*
	DSST Accuracy [%]	−0.12 (0.13)	−0.04 (0.04)	0.04*	0.05	0.44
	BART Risk Taking [%]	−0.24 (0.55)	−0.01 (0.03)	0.66	0.38	<.01*
	PVT Accuracy [%]	0.50 (0.20)	0.09 (0.04)	0.02*	0.89	0.77

[*] P<.05 after Benjamini-Hochberg (Benjamini & Hochberg, 1995) adjustment for multiple testing; Admin: Administration Number; Std.: Standardized; SE: Standard Error; RT: Response Time; MP: Motor Praxis; VOLT: Visual Object Learning Test; F2B: Fractal 2-Back; AM: Abstract Matching; LOT: Line Orientation Test; ERT: Emotion Recognition Test; MRT: Matrix Reasoning Test; DSST: Digit Symbol Substitution Test; BART: Balloon Analog Risk Test; PVT: Psychomotor Vigilance Test; ms: milliseconds