

Executive Abilities:
Measures and Instruments for
Neurobehavioral Evaluation and
Research (EXAMINER)

User Manual 3.6

Table of Contents

Acknowledgements	4
Chapter 1. Mission	5
Chapter 2. Project Structure	6
Chapter 3. Background	7
Chapter 4. Framework for Current Battery	9
Sub-Components of Executive Function	9
The Role of Non-Executive Skills	9
Broad-based and Novel Strategies	10
Current Framework.....	10
Chapter 5. Description of Tasks	11
Domain: Working Memory.....	11
Domain: Inhibition.....	13
Domain: Set Shifting.....	14
Domain: Fluency.....	16
Domain: Planning	17
Domain: Insight	18
Domain: Social Cognition and Behavior	18
Chapter 6. Data Collection	20
Diagnostic Groups	20
General Inclusion Criteria	20
General Exclusion Criteria	20
Subject Diagnostic Criteria	21
Chapter 7. Software Installation and Administration	25
Requirements.....	25
Acquiring and Installing Software Dependencies	25
Chapter 8. Psychometric Properties	33
Administration Issues.....	33
Distributions.....	33
Reliability: Individual Tests.....	34
Reliability: Executive Composite and Factor Scores	36
Chapter 10. Validity	37
Executive Composite Score	37
Working Memory Factor Score	39
Cognitive Control Factor Score.....	42
Fluency Factor Score.....	43
Measures not included in factor scores	45
Chapter 11. Variable and Scale Construction	48
Individual Tests.....	48
Factor Scores.....	51

Chapter 12. Administration and Scoring Guidelines	58
Global considerations.....	58
Fluency	58
Unstructured Task	62
Flanker.....	64
Set Shifting	65
Dot Counting	66
Continuous Performance Test.....	67
1-Back.....	68
2-Back.....	69
Social Norms Questionnaire	72
Chapter 13. Procedures for Calculating Composite and Factor Scores	76
Installing the Scoring Program	76
Running the Scoring Program.....	76
Preparing an Input File for the Scoring Program	77
Composite and Factor Score Output Variables	78
Chapter 14. Known Issues	79
References	80

Acknowledgements

The EXAMINER project represents the combined efforts of many individuals to whom we are very grateful.

Our external advisers were generous with their time and wisdom throughout the course of the project. Kimberly Espy, Josette Harris, Steven Hinshaw, Robert Knight, David Knopman, Paul Malloy, Jennifer Manly, Dan Mungas, Ron Ruff, Tim Salthouse, Donatella Scabini, and Phil Zelazo all made immensely valuable contributions.

The team at NIH, Emmeline Edwards, Helene Braun, and Laurie Leonard provided able stewardship, guidance and assistance, helping us navigate through a great many scientific and administrative challenges.

We are extremely grateful to the PIs and coordinators at all of the data collection sites, including Ramon Diaz-Arrestia, Kimberly Espy, Robert Knight, Dan Mungas, Celiane Rey-Casserly, Jeffrey Schatz, Glenn Smith, Gerry Taylor, and Daniel Tranel. We were fortunate to have the opportunity to collaborate with so many gifted, generous, and capable colleagues.

The UCSF team members each contributed their own unique brand of creativity, hard work, and brilliance to make this project happen. Special thanks go to Blaire Benavides, Caroline Becerra, Ashley Berhel, Lauren Gritzer, Caroline Latham, Lena Sinha, Susan Verde, and Mary Widmeyer. Katherine Possin was instrumental in all aspects of our working memory tasks, and Katherine Rankin took the lead on the social cognition measures. Adam Boxer played a key role in our eye movement task. Howard Rosen assisted with the insight measures. Katrin Schenk inspired novel analytic approaches to random number generation. Bruce Miller, Ari Green, and Chad Christine facilitated subject recruitment and offered ongoing advice. Kathleen Drew was the backbone of our administrative team, and John Neuhaus and Alan Bostrom provided valuable biostatistical input at every step of the way. Joe Hesse and Charlie Toohey built the best data management system imaginable and programmed the computer tasks and scoring systems.

Special thanks go to Dan Mungas, who in addition to being an adviser and a site PI, led our efforts to use item response theory for scale construction.

Finally, we are indebted to the patients and control subjects who willingly subjected themselves to hours of testing, and to their families who also assisted in so many ways. Measurement tools are built to help us understand the effects of disease and aging, and our patients remain our ultimate inspiration.

Chapter 1. Mission

In May 2005, the NINDS issued a call for proposals to develop domain specific methods for defining and measuring executive functioning. Our team at the University of California, San Francisco, led by Joel Kramer, PsyD, was fortunate to be awarded the contract. The project, entitled *Executive Abilities; Methods and Instruments for Neurobehavioral Evaluation and Research (EXAMINER)* commenced on December 30, 2005.

Executive function refers to a constellation of cognitive abilities that include the ability to plan, organize, self-monitor, and manage multiple tasks simultaneously. Despite its importance for clinical and neuroscience research, the paucity of valid and reliable tasks that specifically tap domains of executive functioning remains a significant obstacle for research in this area.

An executive function battery should enable clinical investigators to assess executive functions reliably and validly across a variety of ages and disorders in cross-sectional and longitudinal studies. Such a battery should reflect both cognitive and non-cognitive behaviors and bear some relationship to day-to-day executive functioning. In addition to these underlying themes, we believe that a successful executive function battery should have the attributes described below.

Modular. Executive functioning encompasses multiple domains of behavior, such as generation, set-shifting, working memory, inhibition, concept formation, and social behavior. Separate modules offering the ability to quantify each of the relevant domains should be available to researchers to use separately or collectively, depending on the researcher's needs.

Modifiable. The specific needs of individual clinical investigations can vary quite considerably in terms of the types of executive tasks, the age and level of impairment of the study sample, and the overall study design. Any standard battery of executive tasks must be flexible enough to be adapted to a range of experimental and clinical situations.

Efficient. In most research settings, executive functioning will be only one of several data points required. To be maximally useful, an executive function battery will need to be very efficient, providing reliable and valid test data in as brief a period as possible.

Applicable to a broad range of subjects in terms of age and ethnicity. Very few standardized neuropsychological instruments are available that can be used with both pediatric and adult populations. In addition, executive tasks, like most cognitive measures, tend to be culture and language specific, and their utility for administration in other languages has not been established. As a result, the issue of measurement bias is of central importance in cross-cultural applications of neuropsychological tests. Executive measures should detect cognitive change, and an unbiased measure should be equally sensitive to change in individuals from different age and ethnic groups.

Psychometrically robust. Extant executive measures are often criticized for low or unmeasured reliability, questionable relationships with frontal lobe injury or real world behavior (Heflin et al., 2011; Stuss et al., 1995), or skewed distributions. Good psychometric properties are a key element for measures designed for clinical trials, as are the availability of equivalent alternate forms.

Chapter 2. Project Structure

EXAMINER was developed in two general phases. Phase I lasted two years and emphasized battery development. In the first year of Phase I, the UCSF team was built, and a website was created to facilitate communication to NIH and the public (examiner.ucsf.edu). The literature on executive functioning was extensively reviewed by the UCSF team and posted on the website. A team of external advisers was created, with the first meeting taking place in San Francisco April 24, 2006. Finally, the advisers and experts in the field were surveyed on what they felt were the highest priorities for battery development. These results are also summarized on the website. Priorities identified by the NINDS, the external advisers, and the survey of experts were to: 1) have a brief (30–40 minute) battery designed for clinical trials with alternate forms and a single composite score; 2) create a menu of tasks from which investigators can select to meet specific research goals; 3) use non-copyrighted tasks that NIH could distribute freely; and 4) validate the battery by demonstrating a relationship with real-world markers.

During Phase I's second year, attention shifted toward defining the conceptual framework for the EXAMINER battery, selecting extant executive paradigms from the research and clinical literature, and developing novel tasks. Extensive piloting was undertaken at UCSF and UC-Davis (Dan Mungas, PI), and tasks were continually revised. Record forms, test stimuli, software for computerized tasks, and training materials were created. Translation of test materials was carried out by a professional translation service with back translation. Traditional neuropsychological measures like Trail-Making, Stroop Interference, WAIS-III Digit Symbol, D-KEFS Design Fluency, and Wide Range Achievement Test-III Reading subtest were added to the battery as control measures. The Frontal Systems Behavior Scale™ (FrSBe) and the Behavior Rating Inventory of Executive Function® (BRIEF), copyrighted informant-based questionnaires, were added as measures of day-to-day executive functioning and behavior. Concurrently, the information technology team under the direction of Joe Hesse began work on a web-based data management system for use during the data collection phase. Finally, subcontract sites for data collection were identified, and the contracts and grants process was initiated to enable sites to begin data collection in January 2008. A second advisory meeting took place in San Francisco November 1, 2007. Additional advisory meetings took place on June 24, 2009 and February 9, 2011.

Phase II was initially designed for two years of data collection; a third year was later added, and EXAMINER was able to exceed its original recruitment goals. Data collection is described in more detail in Chapter 6. Several different approaches to data reduction were piloted, and item response theory was ultimately selected as the best method for generating a smaller set of meaningful scores with ready application to research and clinical trials settings.

Chapter 3. Background

Executive abilities, which are widely accepted to be a central component of human cognition, reflect a capacity to engage in goal-oriented behavior. To evaluate executive abilities, clinicians have tended to emphasize constructs like fluency, working memory, concept formation, set shifting, and inhibition, using a range of cognitive measures. Test batteries specifically designed to measure executive functioning have been developed, including the Frontal Assessment Battery (FAB) (Dubois, Slachevsky, Litvan, & Pillon, 2000), Executive Interview (EXIT25) (Royall, Mahurin, & Gray, 1992), Behavioral Assessment of Dysexecutive Syndrome (Wilson, 1996), Cambridge Neuropsychological Test Automated Battery (CANTAB) (Robbins et al., 1994), and the Delis-Kaplan Executive Function System (Delis, Kaplan, & Kramer, 2001). Espy (Espy & Cwik, 2004; Espy, Kaufmann, & Glisky, 2001; Espy, Kaufmann, McDiarmid, & Glisky, 1999; Espy et al., 2002) and others (D. C. Delis, et al., 2001; Delis, Kramer, Kaplan, & Holdnack, 2004; Korkman, Kemp, & Kirk, 2001) have thoughtfully extended assessment of executive function to younger children. Methods for assessing frontally-mediated neuropsychiatric symptoms and dysexecutive syndromes have also been developed (Cummings et al., 1994; Malloy & Grace, 2005).

Despite the proliferation of executive measures, there is little agreement about what the primary executive abilities are, how they are organized, what the underlying neuroanatomy is, or how they should best be measured. There are also no widely accepted unified models of executive functions. Nonetheless, several key concepts have been proposed that have had a significant influence on our understanding of executive control. For example, Shallice has proposed that the frontal lobes organize a Supervisory Attention System that distinguishes between routine tasks for which contention scheduling is sufficient and novel problems that require more top-down control (Shallice & Burgess, 1996). Stuss et al (Stuss, Shallice, Alexander, & Picton, 1995) have refined and extended this model to incorporate a range of anterior attentional functions. Miyake and colleagues (2000) reported three different executive functions, shifting, inhibition, and updating, that were modestly correlated but separable. Updating, as viewed by Miyake, overlaps with the broader concept of working memory. Models of working memory have included a “central executive”, mediated largely by dorsolateral prefrontal cortex (Baddeley, 2002; Baddeley & Della Sala, 1996) and sub-processes that include storage/maintenance, rehearsal, interference control, inhibition, and scanning functions (D'Esposito et al., 1995; D'Esposito, Postle, Ballard, & Lease, 1999).

Decision-making, reward processing, self-regulation, and inhibition are additional key components of executive functioning that have been studied experimentally, and their relationships with medial, ventral, and dorsolateral prefrontal structures are being defined (Bechara, Damasio, Tranel, & Anderson, 1998; Levine et al., 2000; McDonald, Ko, & Hong, 2002; Shallice & Burgess, 1991).

It has also become clear that simple paper and pencil tasks will not always capture real-life social and executive deficits. Tasks that capture deficits in executive control in the area of social cognition are needed.

The psychometric properties of executive tasks pose yet another challenge to clinical investigators interested in measuring executive functioning. Construct validity refers to how well an instrument measures what it purports to measure. Importantly, clinical neuropsychological instruments have been criticized for being multifactorial, drawing on several non-executive component skills. In fact, listed among the top 20 “executive tasks” in a survey of neuropsychologists by Rabin et al (Rabin, Barr, &

Burton, 2005) were a memory task (CVLT) and visuospatial tasks such as clock drawing, Rey-Osterrieth Complex Figure, and block design. While no one would argue that executive skills were irrelevant to performing these tests, it is not possible to untangle the various component skills. Not surprisingly, the ability of these tasks to differentiate between frontal and non-frontal patients is fraught with error (Anderson, Damasio, Jones, & Tranel, 1991; Berman et al., 1995; Dunbar & Sussman, 1995; Manchester, Priestley, & Jackson, 2004). Another psychometric issue is test-retest reliability. This has particular importance for clinical trials where researchers must be able to attribute change in cognition to the intervention, and not poor reliability or practice effects (Beglinger et al., 2005; Bowden, Benedikt, & Ritter, 1992).

Most current measures of executive function also have limited cross-cultural application. Often the stimuli require reasonable mastery of English (e.g., Similarities, Stroop, DKEFS Card Sorting), or they are culturally based (e.g., proverb interpretation). Executive tasks also tend to be highly correlated with education, and education levels vary across ethnic groups. Even when tasks appear to be readily translatable, assessment of differential item functioning has revealed item bias (Marshall, Mungas, Weldon, Reed, & Haan, 1997). While progress is being made (Chan, Hoosain, & Lee, 2002; Chan & Manly, 2002; Chan, Robertson, & Crawford, 2003; Mungas, Reed, Crane, Haan, & Gonzalez, 2004; Rodriguez del Alamo, Catalan Alonso, & Carrasco Marin, 2003), the shortage of validated clinical measures that are applicable across ethnic and language groups poses a major obstacle to clinical research.

In sum, despite the wealth of available instruments, there are continued concerns about psychometric properties, validity, applicability to settings and populations other than the ones the tests were developed for, suitability for all ages and non-English speaking subjects, and adaptability for clinical trials. There is also no consensus on what the primary components of executive functioning are or how they are best operationalized. There remains a compelling need to have a battery of tests that can be routinely integrated into neurobehavioral research that will reliably and validly measure constructs that clinical investigators agree are important.

Chapter 4. Framework for Current Battery

This proposal to develop psychometrically robust measures of executive function was guided by three basic premises that influenced the initial stages of task selection. First, the term “executive function” is a broad term that requires breaking down into smaller conceptual units. Second, because executive abilities are measured using tasks that require multiple abilities, methods that parse the executive component from other skills are needed. Finally, executive function encompasses both cognitive and non-cognitive behaviors. A multimodal approach using cognitive and observational methods is necessary to capture the broad range of deficits seen in patients with executive dysfunction.

Sub-Components of Executive Function

The first underlying premise is that the term “executive functioning” is an overarching rubric that encompasses multiple domains that are mediated by different neural structures and networks. Various executive abilities include maintenance and manipulation of information, temporal organization, set shifting, self-monitoring, concept formation, fluency, inhibition, motivation, organization, and planning. There is no single “executive function” that investigators can turn to when conducting clinical studies. In many instances, patients with deficits in executive functioning will perform well on certain domains but poorly on others. For example, patients with dorsolateral prefrontal cortex injury tend to perform least well on working memory measures and set-shifting tasks that require shifts between dimensions (Rogers, Andrews, Grasby, Brooks, & Robbins, 2000), yet do well on reversal learning and inhibition tasks. The opposite pattern can be seen in patients with more ventromedial involvement.

Most factor analytic studies support a fractionated, multi-component view of executive function. Although the number and type of factors depend on the variables that are included, some overlap is evident. Miyake et al. (2000) identified three related but separable functions: mental set shifting, information updating and monitoring, and inhibition of pre-potent responses. Another 3-factor model generated by observational data identified behavior regulation, emotional regulation, and a metacognition factor (Gioia, Isquith, Retzlaff, & Espy, 2002), while a 5-factor model yielded constructs like intentionality, interference management, inhibition, planning, and social regulation (Amieva, Phillips, & Della Sala, 2003). These findings support the need for an executive battery that includes a range of executive abilities that are non-redundant.

The Role of Non-Executive Skills

The second underlying premise is that the measurement of these executive abilities is typically carried out using heterogeneous tasks that require multiple non-executive skills. Most tasks designed to assess some aspect of executive functioning also involve varying degrees of information processing speed, working memory, motor speed, language processing, spatial processing, and fundamental perceptual and motor skills. For example, the Trail-Making Test is a very widely used neuropsychological tool for measuring set shifting or mental flexibility, yet can be failed for reasons more related to problems with visual scanning, hand-eye coordination, ability to count, facility with the alphabet, sustained attention, motor speed, and other skills. Accordingly, some elements of the battery were designed explicitly to enable users to parse out the inhibitory or set shifting components.

Broad-based and Novel Strategies

The third underlying premise of this proposal was that a comprehensive approach to executive functioning requires broad-based and novel assessment strategies to capture cognitive and non-cognitive behaviors. Executive functioning is typically viewed as a cognitive ability assessed by neuropsychologists using achievement scores on tasks like Trail-Making, sorting tasks, Stroop Interference paradigms, and reasoning tasks. These types of tasks evaluate critical cognitive components of executive function, but can miss important executive behaviors and social abilities. It is well known, for example, that patients with very impaired executive functioning in the real world (e.g., impaired interpersonal conduct, poor impulse control) can perform normally on these neuropsychological measures of executive functioning. This pattern may be particularly pronounced in patients whose pathology involves ventromedial prefrontal structures; these frontally damaged patients have very disturbed interpersonal behavior yet might perform normally on cognitive testing (Saver & Damasio, 1991; Tranel, Bechara, & Denburg, 2002). Consequently, battery construction requires a broad-based approach that conceptualizes executive functioning as the full set of cognitive, emotional, and social abilities that enable successful goal-directed behavior (Rosso, Young, Femia, & Yurgelun-Todd, 2004).

Current Framework

Our ultimate approach to developing the EXAMINER battery was to integrate the cognitive literature on executive functioning with the clinical literature on the sequelae of frontal injury. We selected Miyake's model as the core conceptual structure for battery design, and targeted tasks that measured mental set shifting, information updating and monitoring, and inhibition of pre-potent responses. For mental set shifting, we emphasized measuring performance when attention or response set must shift, and contrasting this measure with performance on component tasks that did not require a shift. For information updating, we tapped into the larger construct of working memory, recognizing that working memory tasks range in the degree to which they require updating of information versus manipulation of information in short-term memory. Consistent with Miyake, we targeted tasks that required some degree of updating information. Inhibition of pre-potent responses covered a broad range of tasks that potentially measure cognitive and behavior control in varying degrees. To this core set of constructs we added fluency, a measure with a rich clinical tradition. Fluency refers to the ability to utilize one or more strategies that maximize the production of responses while avoiding response repetition (Ruff, Allen et al. 1994). Word fluency tests are typically considered measures of executive function because they tap organizational strategies required for retrieval and recall, as well as self-monitoring, self-initiation, and inhibition. Planning is another concept that is widely considered to be a component of executive functioning, although very challenging to operationalize and quantify. Unlike the components of Miyake's model, planning is clearly heterogeneous and multifactorial, requiring a number of component processes like sustained attention, abstract thinking, temporal sequencing, and reasoning. Insight is another construct that is often included in discussions of executive functioning and linked to the frontal lobes, so we inserted methods to evaluate subjects' ability to accurately appraise their performance. Finally, how someone actually behaves and functions in real life is an important non-cognitive measure of executive functioning, and includes social cognition as well as behavioral control. While traditionally assessed using collateral report, the need for EXAMINER to provide a stand-alone assessment without reliance on informants led us to develop and incorporate novel self-report and observational methods to capture social cognition and behavior.

In Chapter 5, we describe in more detail our measures of working memory, inhibition, set shifting, fluency, planning, insight, social cognition, and behavior.

Chapter 5. Description of Tasks

This chapter provides a descriptive overview of each EXAMINER measure by domain, along with a brief discussion of administration time. We include all the tasks that we collected data on during the field testing phase, even if a task was not included in the final battery. In addition, after data were collected on the first 800 subjects, several tasks were carefully analyzed to see if they could be shortened without losing any information. Any modifications to the tasks are also described.

Domain: Working Memory

Dot counting

The dot counting task measures verbal working memory. The examinee is asked to look at a screen with a mixed array of green circles, blue circles and blue squares. The examinee is asked to count all of the blue circles on the screen one at a time, out loud and remember the final total. Once the examinee finishes counting the blue circles on one screen, the examiner switches the display to a different mixed array of green circles, blue circles and blue squares. The examinee is instructed to count the blue circles in the new display. The number of different displays presented to the examinee in each trial increases from two to seven over six experimental trials. After counting the blue circles on all of the displays presented within a trial, the examinee is asked to recall the total number of blue circles that were counted in each of the different displays in the order in which they were presented. Partial credit is given based on how many totals the examinee can recall correctly from each trial.

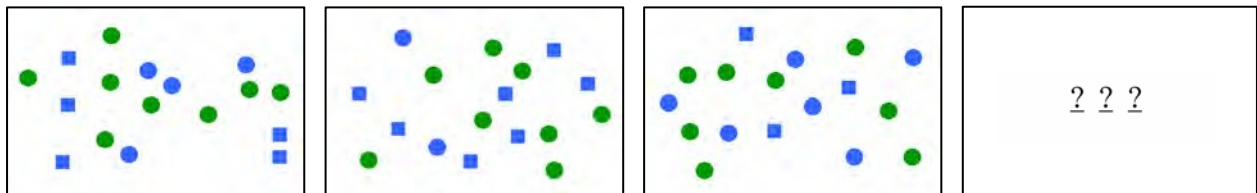


Figure 1. Example of Dot Counting stimuli

In an effort to reduce patient burden and time of administration, the number of trials administered was reduced by half. Initially this task consisted of two trials at each series length, two through seven. After observing that the correlation between the mean percent correct for all 12 trials (out of 54 total points) and the mean percent correct for trials 1,3,5,7 & 9 (out of 27 total points) were highly correlated, we eliminated the second of the two trials for each series length. This resulted in the task being reduced from 12 trials to six trials with possible scores ranging from 0 to 27.

N-back

The n-back paradigm is a widely used measure of working memory that requires flexible updating capabilities. EXAMINER includes a spatial 1-back and 2-back task to assess spatial working memory. The 1-back requires maintaining and updating 1 location at a time, whereas the more difficult 2-back requires maintaining and updating two locations.

During both the 1-back and the 2-back, the examinee is shown a series of 2.4 cm white squares that appear in 15 different locations on a black screen. Each square is presented for 1000 msec. All of the locations are equidistant from the center of the screen. During the 1-back, the examinee is instructed to press the left arrow key whenever the square is presented in the same location as the previous one and the right arrow key if the square is presented in a different location as the previous one. Responses should be given as quickly as possible while trying to maintain accuracy throughout the trials. The next square appears on the screen after each response is given. A number (varying from 1–9, selected randomly) appears in the center of the screen 500 msec after each response and remains on the screen for 1000 msec. The examinee should say this number out loud immediately when it appears on the screen before responding to the next square. The 1-back consists of one block of 30 trials, ten of which match the location of the previous square, and 20 that are in a different location. The stimuli are presented in a fixed order for all participants.

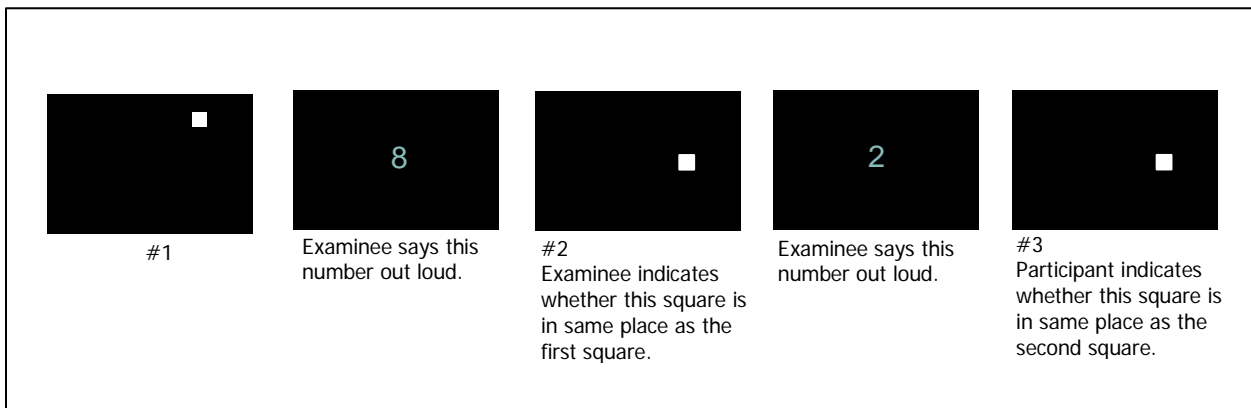


Figure 2. Example of 1-back stimuli

During the 2-back, the examinee is instructed to press the left arrow key whenever the square is presented in the same location as the square two squares before and the right arrow key if the square is presented in a different location as the square two before. The 2-back consists of one block of 90 trials, 30 of which match the location of the square two before, and 60 that are in a different location. The squares are presented in a fixed order for all participants.

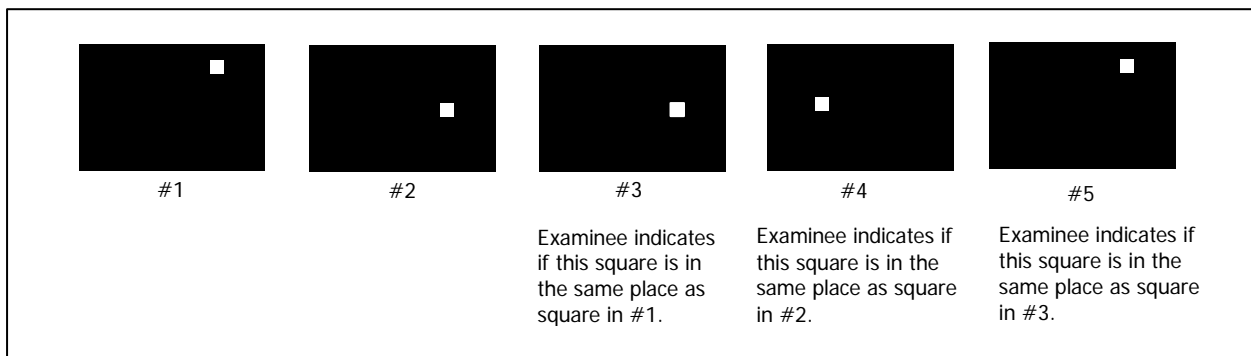


Figure 3. Example of 2-back stimuli

Domain: Inhibition

Flanker

The examinee is instructed to focus on a small cross in the center of the screen. After a short variable duration (1000 msec–3000 msec), a row of five arrows is presented in the center of the screen either above or below the fixation point. The duration of the stimuli presentation for each trial is 1000 msec.

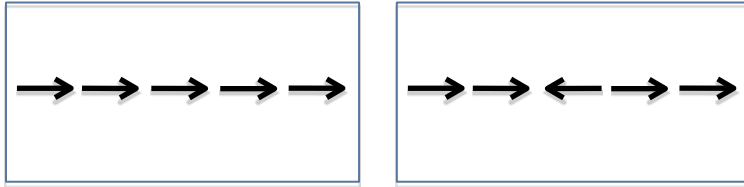


Figure 4. Example of Flanker stimuli

The examinee is then required to indicate whether the centrally presented arrow is pointing either to the left or right by pressing the left or right arrow key. The examinee is presented with two different conditions during the task, incongruent and congruent. In the congruent trials, the non-target arrows point in the same direction as the target arrow and in the incongruent trials they point in the opposite direction. Examinees should respond as quickly and accurately as possible. The stimuli are presented in a random order with each condition being presented 24 times resulting in 48 total trials.

Initially the Flanker task consisted of 64 experimental trials, 32 of which were congruent and 32 of which were incongruent. In order to determine how many trials we would be able to eliminate without compromising precision, we ran correlations between the median RTs for k number of trials (where k is 1 through 32) and the median RTs resulting from all 32 trials. This was done separately for both the congruent and the incongruent conditions. We observed that when k equals 24, the mean correlation between median reaction times for both congruent and incongruent trials showed a correlation greater than .95 with the median RTs for the full 32 trials. In addition, the difference in median RT between a 24-trial task and a 32-trial task was negligible. This led us to eliminate 8 trials from each condition, reducing the total number of trials down from 64 to 48. Figure 5a shows the correlation between k trials and the 32-trial median RT; figure 5b shows the difference in median RT between k trials and 32 trials.

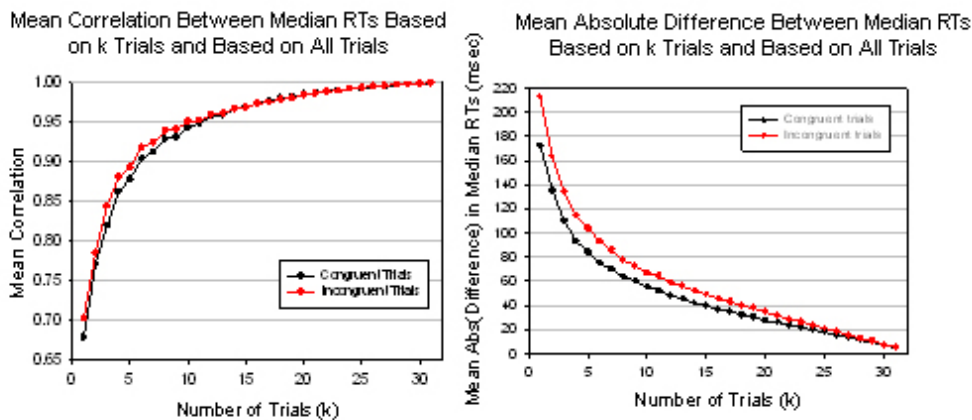


Figure 5a and 5b. Flanker simulation

Continuous Performance Test (CPT)

The continuous performance task is a classic response inhibition task, which requires subjects to respond to a certain type of stimulus and withhold a response to another.

The examinee is presented with different images in the center of the screen instructed to press the left arrow key for only the target image (e.g., a white five-pointed star in Form A), responding as quickly and accurately as possible. The task consists of 100 experimental trials, 80% of which were the target image. The five non-target images that were presented were of a similar shape and of comparable size to the target for each form.

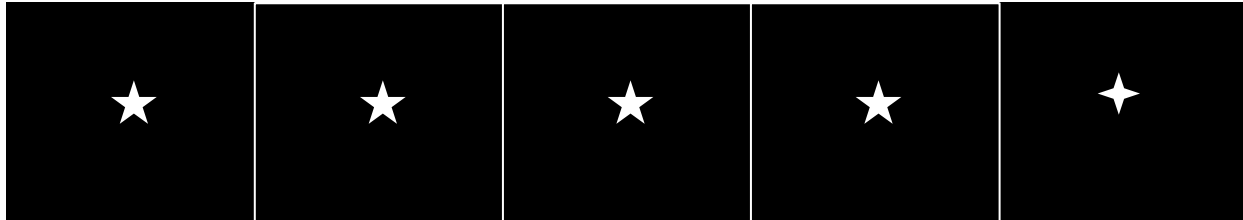


Figure 6. Example of CPT stimuli

Anti-saccades

This is an eye movement task. There are three blocks of trials in which subjects look at a fixation point in the center of a computer screen and move their eyes upon presentation of a laterally presented stimulus. In the first block (pro-saccade), subjects are instructed to move their eyes in the direction of the presented stimulus. In the second and third blocks (anti-saccade), subjects are instructed to move their eyes in the opposite direction of the presented stimulus.

Dysexecutive Errors

An underlying assumption in developing the EXAMINER battery is that executive related deficits can manifest as impulsive errors, failure to shift set, perseverative behavior, and stimulus-boundedness, even when achievement scores on tests are unremarkable. Accordingly, we generated a composite error score using several EXAMINER tasks. This composite includes false alarm responses on the CPT, rule violations on the verbal fluency tasks, the tendency to make errors on Flanker incongruent trials relative to congruent trials, the tendency to make errors on the Set Shifting shift trials relative to the non-shift trials, and the total score on the Behavior Rating Scale.

Domain: Set Shifting

Set Shifting

Participants are required to match a stimulus on the top of the screen to one of two stimuli in the lower corners of the screen. In task-homogeneous blocks, participants perform either Task A (e.g., classifying shapes) or Task B (e.g., classifying colors). In task-heterogeneous blocks, participants alternate between the two tasks pseudo-randomly. The combination of task-homogeneous and task-heterogeneous blocks allows measurement of general switch costs (latency differences between heterogeneous and homogeneous blocks) and specific switch costs (differences between switch and non-switch trials within the heterogeneous block).

The examinee should be positioned in front of the computer screen. The screen on each trial contains a red triangle in the bottom left corner and a blue rectangle in the bottom right corner. These reference objects remain on the screen for the duration of the task. At the beginning of each trial, a cue that reads “shape” or “color” appears at the bottom of the screen followed by a blue triangle stimulus or a red rectangle stimulus in the top-center of the screen. The examinee is then instructed to match the stimulus presented with one of the reference objects based on the cue provided. The right arrow key is used to match the stimulus object with the reference object on the right side of the screen and the left arrow key is used to match the stimulus object with the reference object on the left side of the screen.

The task was organized into three blocks, homogenous block A, homogenous block B and heterogeneous block AB. The two homogenous blocks each consisted of 20 trials for which the same cue is presented, one being all shape and one being all color. The starting homogenous block was counterbalanced between color and shape across participants. The heterogeneous block consisted of 64 trials, 32 of which had a color cue and 32 of which had a shape cue. The 64 trials were sampled randomly within the heterogeneous block requiring participants to shift between color matching and shape matching for a randomly sampled subset of the trials.

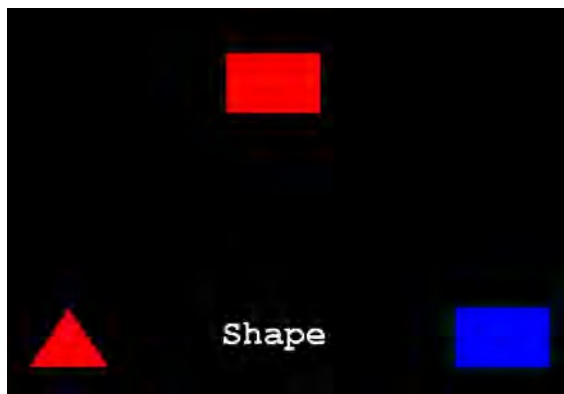


Figure 7. Example of Set Shifting stimulus

In order to determine how many trials we would be able to eliminate from the shifting block without sacrificing precision, we plotted the mean shift effect (absolute difference between shift trials) for k number of trials where k is 1–72. We observed that this curve was relatively flat at the point that k equals 64. Since there was no significant difference in mean shift cost between 64 and 72 trials, we concluded that we could eliminate 8 trials from the shifting block.

In order to determine how many trials could be eliminated from the Shape and Color blocks, we ran correlations similar to those run for the Flanker task. We observed that mean reaction times where k equaled 20 were not significantly different than when k equaled 24. We concluded that we could eliminate 4 trials from each of these two blocks. In total we removed 16 trials bringing the number of total trials from 120 to 104. The absolute difference between median RT of k trials versus all trials is shown in figure 8a. The absolute difference in shift cost (number of milliseconds slower on shift trials relative to non-shift trials) between k and 94 trials is shown in figure 8b. Figures 8c and 8d show the correlation of median RT between k trials and all trials for shift and non-shift blocks.

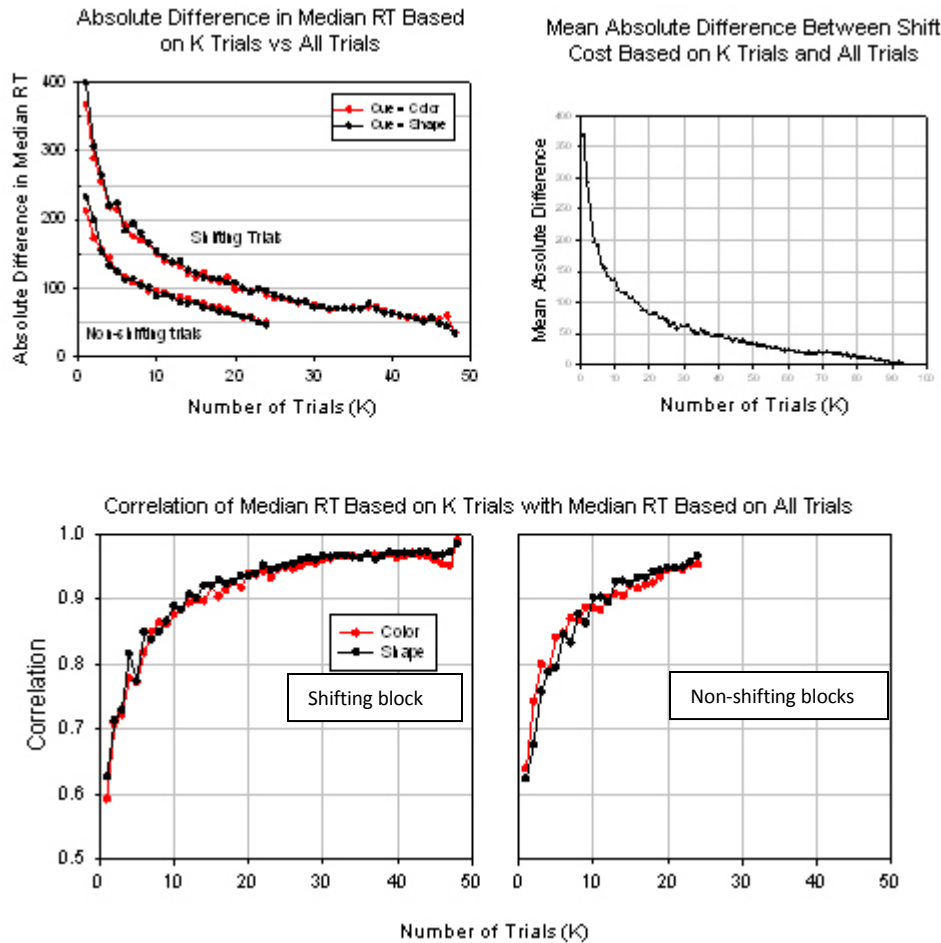


Figure 8. Set Shifting simulation

Domain: Fluency

Phonemic Fluency

For the phonemic fluency task, examinees are instructed to name as many words as they can that begin with a particular letter of the alphabet as quickly as they can. Sixty seconds are allowed for each letter. The examinee is instructed that names of people, places and numbers are not acceptable responses. Grammatical variants of previous responses (plurals, altered tenses, and comparatives) are also not acceptable responses. All responses should be recorded by the examiner. The number of correct responses, repetitions and rule violations are then totaled for each letter.

Category Fluency

For the category fluency task, examinees are asked to generate as many items that they can think of that belong to a particular category as quickly as possible. Sixty seconds are allowed for each category. All responses are recorded by the examiner. The number of correct responses, repetitions and rule violations are totaled for each category.

Domain: Planning

Unstructured Task

This task was modeled after the 6-elements test (Shallice & Burgess, 1991). Subjects are presented with three booklets, each containing five pages of simple puzzles (4 per page). The puzzles were designed to be cognitively simple (e.g., connect the dots; trace the design) but average completion times range from 4 to 60 seconds. Each puzzle has a designated point value, and subjects are given 6 minutes to earn as many points as possible. Irrespective of actual point value, puzzles can have a high cost-benefit ratio (i.e., the time required to complete the puzzle makes it less desirable) or a low cost-benefit ratio (i.e., the time required to complete the puzzle makes it more desirable). In addition, the proportion of low cost-benefit items decreases as subjects proceed through a booklet. Subjects need to plan ahead, avoid items that are strategically poor choices, and be cognizant of when a particular booklet offers diminishing returns.

INSTRUCTIONS:

You can earn points by completing the puzzles in these booklets. Some are easier than others. You will only have 6 minutes to earn as many points as possible so choose your puzzles carefully. Be sure to complete items accurately in order to receive full credit. You do not have to do all of the pages in a book, and you do not have to do all of the puzzles on a page.

GOAL: Earn as many points as possible.

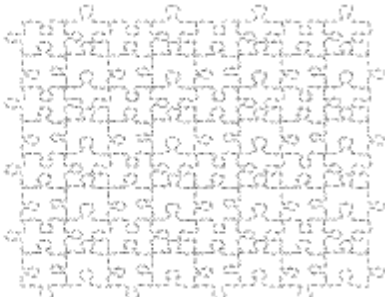
25 pts.

Solve the problems below.

$$\begin{array}{r} 6 \\ -3 \\ \hline \end{array} \quad \begin{array}{r} 8 \\ -2 \\ \hline \end{array} \quad \begin{array}{r} 9 \\ -7 \\ \hline \end{array} \quad \begin{array}{r} 4 \\ -4 \\ \hline \end{array} \quad \begin{array}{r} 7 \\ -5 \\ \hline \end{array}$$

75 pts.

Trace the dotted lines.



75 pts.

Cross out all of the letters J and K.

J	F	T	K	I	O	K	J	N
K	J	N	J	G	B	U	K	I
D	E	K	F	C	K	U	J	Y
J	K	J	B	A	K	N	B	J
I	A	K	D	K	L	J	T	K
V	J	Y	H	J	K	C	J	V
U	K	J	D	R	J	G	K	B
K	N	F	K	J	E	K	R	J
J	L	J	B	O	K	J	U	P

10 pts.

How many shaded squares (■)?

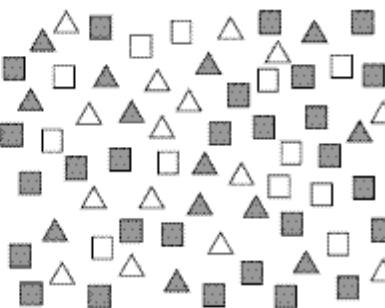


Figure 9. Example of Unstructured Task stimulus page

Domain: Insight

Insight

Examinees are asked to rate themselves on their performance immediately after completing the well-normed verbal fluency tasks. Before the fluency tasks begin, the examinee is informed that after performing the task they will be asked to assess their performance. They are instructed to assess their own performance relative to a hypothetical sample of 100 people of a similar age and level of education. After the fluency task is complete they are shown a picture of a bell curve with corresponding percentile rankings at the bottom of the page (Figure 10 below). They are then reminded that, on a typical task, the majority of healthy age matched peers would perform at the 50th percentile, with smaller numbers performing above or below average (corresponding locations were pointed to by the examiner).

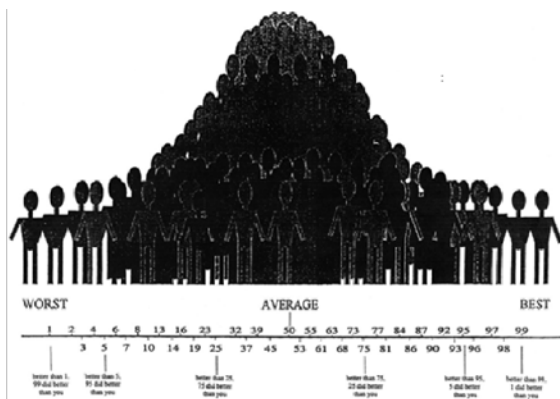


Figure 10. Normative graph for examinee self-rating

Domain: Social Cognition and Behavior

The Social Norms Questionnaire

This task measures subjects' crystallized knowledge of social norms in a linguistically and cognitively simple manner. This yes-no questionnaire is designed to determine the degree to which subjects actually understand and can accurately identify implicit but widely accepted social boundaries in the dominant U.S. culture. The Social Norms Questionnaire includes both socially inappropriate behaviors (e.g., "Cut in line if you are in a hurry," "Pick your nose in public," and "Wear the same shirt every day") and generally acceptable behaviors (e.g., "Tell a coworker your age," "Blow your nose in public," and "Eat ribs with your fingers,"). The subject must decide whether the behavior is socially appropriate or not if it were hypothetically enacted with an acquaintance or coworker. Two subscales are derived that represent a) whether the subject errs in the direction of breaking a social norm, the "Break" score (e.g., indicating that it is permissible to cut in line if one is in a hurry); or b) in the direction of interpreting a social norm too rigidly, the "Overadhere" score (e.g., indicating that it is not permissible to eat ribs with one's fingers). There is a 22-item questionnaire for adults and a 30-item questionnaire for children. This measure also has an alternate form for test-retest purposes.

Behavior Rating Scale

This rating scale is completed by the examiner after completion of the testing. Examiners restrict their ratings to behaviors that they have observed directly, but include all observed behaviors, regardless of

the context. Thus, although behaviors during the actual assessment will likely provide the bulk of data, examiners should also note behaviors exhibited in all other situations, such as the waiting room and walking to and from the exam room. There are nine behavioral domains to rate, including agitation, stimulus-boundedness, perseverations, decreased initiation, motor stereotypies, distractibility, degree of social/emotional engagement, impulsivity, and social appropriateness.

Chapter 6. Data Collection

A central goal of the EXAMINER project was to develop a battery that could reliably and validly assess executive functions across a wide range of ages and disorders. Data collection targets were established with this goal in mind by the EXAMINER advisory panel, NINDS focus groups, NINDS Project Officer, and the UCSF team. Piloting and ongoing data collection were conducted utilizing the large research infrastructure at UCSF and in collaboration with nine remote sites to represent a full range of geographic regions, ethnic groups, age groups, and diagnostic disorders.

Diagnostic Groups

The diagnostic categories below represent neurological and neuropsychiatric syndromes associated with executive deficits. Data were also collected on healthy subjects across the age span.

Adults and children with the following neurological conditions or neurodegenerative disorders are represented in the EXAMINER battery dataset: attention deficit hyperactive disorder (ADHD), Alzheimer's disease (AD), focal lesions, behavioral variant frontotemporal dementia (bvFTD), Huntington's disease (HD), mild cognitive impairment, amnesic and executive subtypes (MCI-mem, MCI-exec), multiple sclerosis (MS), Parkinson's disease (PD), progressive supranuclear palsy (PSP), sickle-cell anemia, traumatic brain injury (TBI), and very low birth weight (VLBW).

General Inclusion Criteria

Collectively, participants had to be between 3–90 years old and speak fluent English and/or Spanish. Subjects who were unable to consent for themselves required an informant to consent for them.

General Exclusion Criteria

Participants were excluded if they had the following:

- Current alcohol abuse or dependence
- Current drug abuse
- Psychiatric disorder (apart from those specified in diagnostic groups of interest)
- B12 deficiency or other metabolic syndrome
- Hypothyroidism (i.e. TSH>150% of normal)
- Known HIV
- Renal failure
- Respiratory failure (i.e. requiring oxygen)
- Significant systemic medical illnesses (e.g. deteriorating cardiovascular disease)
- Current medication likely to affect CNS functions (e.g. benzodiazepines, antidepressants, lithium, and/or neuroleptics in the phenothiazine and haloperidol families).

Please see EXAMINER Diagnostic Table below for additional inclusion and exclusion criteria.

Subject Diagnostic Criteria

Inclusion and exclusion criteria for the diagnostic categories followed classifications as reflected in the following guidelines. Principal investigators at each site reviewed and confirmed subject diagnoses.

Table 1. EXAMINER Diagnostic Table

<i>Diagnosis</i>	<i>Criteria</i>
Alzheimer’s Disease (AD)	McKhann et al, NINCDS-ADRDA Criteria (1984)
Attention Deficit Hyperactive Disorder (ADHD)	Diagnostic and Statistical Manual (DSM-IV; 1994) Age range of 7–16 yrs and IQ > 90. <i>Additional Exclusion Criteria:</i> Comorbid mental retardation (MR), neurological diagnosis, psychiatric condition, or major learning disability.
Behavioral Variant Frontotemporal Dementia (bvFTD)	Neary et al Criteria (Neary et al., 1998)
Focal Lesion	Focal lesions included: lateral frontal, ventromedial frontal, non-frontal, and basal ganglia lesions secondary to ischemic stroke, tumor, or focal injury.
Huntington’s Disease (HD)	HD mutation, gene positive
Mild Cognitive Impairment (MCI)	Petersen et al (Petersen et al., 2001) In addition, MCI participants were further subdivided into those that met criteria for a primarily amnesic presentation (MCI-mem) or dysexecutive presentation (MCI-exec).
Multiple Sclerosis	McDonald Revised Criteria (Polman et al., 2005)
Parkinson’s Disease	Albanese Criteria (2003)
Progressive Supranuclear Palsy (PSP)	Litvan et al, NINDS-SPSP Criteria (Litvan et al., 1996)
Sickle Cell Anemia	Confirmed diagnosis of sickle cell anemia Age range of 8-17 yrs.
Traumatic Brain Injury (TBI)	Moderate to severe, as defined by a Glasgow Coma Scale (GCS) < 12. Age range of 18-50 yrs. Injury at least 6-months prior to testing
Very Low Birth Weight (VLBW)	Birth weight of <1000 grams and/or <28 weeks gestational age. Age range of 10-12 yrs.

Data Collection Sites

The EXAMINER battery was administered at nine collaborating sites across the country. The final dataset includes adults and children, Spanish and English speaking, across a wide range of diagnostic cohorts. Please see the Data Collaboration Table below for details regarding baseline data collection for each site.

Adults

The Memory and Aging Center at the University of California-San Francisco (PI: Joel Kramer, PsyD) is the largest source of EXAMINER battery data. UCSF administered the battery to a total of 249 people, and a subset of these individuals had longitudinal data collected. Specifically, 44 subjects were administered EXAMINER twice and 3 received the battery a third time. The diagnostic breakdown of subjects from UCSF includes Alzheimer's disease, bvFTD, Huntington's disease, PSP, MCI, MS, and Parkinson's subjects. Additionally, UCSF recruited and tested normal controls.

The University of California-Berkeley's Helen Wills Neuroscience Institute (PI: Robert Knight, MD) administered the battery to a total of 45 people, including frontal-lesion and non-frontal lesion subjects, as well as normal controls.

The Mayo Clinic Alzheimer's Disease Research Center (PI: Glenn Smith, PhD) administered the battery to a total of 79 people, including AD, bvFTD, MCI, and Parkinson's disease subjects, as well as normal controls.

The University of Iowa (PI: Daniel Tranel, Ph.D.) administered the battery to 87 people, including frontal and non-frontal lesion subjects, as well as normal controls. Two subjects were also evaluated longitudinally.

The University of Texas Southwestern Medical Center (PI: Ramon Diaz-Arrastia, MD, PhD) administered the battery to 33 subjects, including TBI subjects as well as normal controls.

Spanish Language

The University of California-Davis, Department of Neurology (PI: Dan Mungas, PhD) recruited Spanish-speaking subjects to ensure psychometric validity between the English and Spanish language instruments. Dr. Mungas and his team administered the battery to 180 older, predominately healthy, Spanish-speaking subjects.

Children

The Developmental Cognitive Neuroscience Laboratory at University of Nebraska-Lincoln (PI: Kimberly Espy, PhD) administered the battery to 207 normal control children. A subset of these children was evaluated longitudinally. Specifically, 131 children returned for a second battery administration approximately 1 year from their initial visit, and 30 children received the battery a 3rd time.

Boston Children's Hospital Department of Neurology (PI: Celiane Rey-Casserly, PhD) administered the battery to 41 children, including ADHD subjects and normal controls.

Case Western Reserve University Rainbow Hospital (PI: H. Gerry Taylor, PhD) administered the battery to 72 children with Very Low Birth Weight.

The University of South Carolina Neuropsychology and Human Development Lab (PI: Jeffrey Schatz, PhD) administered the battery to 117 people, including individuals diagnosed with Sickle Cell Anemia and normal controls.

Table 2. EXAMINER Baseline Data Collaboration Table

	Diagnosis															Total
	Normal control	ADHD	Alzheimer's	Frontal lesion	Non-frontal lesion	FTD	Huntington's	MCI mem	MCI exec	Multiple sclerosis	Parkinson's	PSP	Sickle cell	VLBW	TBI	
Collaborating Site																
Case Western Reserve	0	0	0	0	0	0	0	0	0	0	0	0	0	72	0	72
Boston Children's Hospital	7	34	0	0	0	0	0	0	0	0	0	0	0	0	0	41
Mayo Clinic	30	0	22	0	0	9	0	3	14	0	1	0	0	0	0	79
University of California, Berkeley	17	0	0	21	7	0	0	0	0	0	0	0	0	0	0	45
University of California, Davis	180	0	1	0	0	0	0	0	0	0	0	0	0	0	0	181
University of California, San Francisco	137	0	16	0	0	13	18	7	7	17	21	13	0	0	0	249
University of Iowa	21	0	0	22	44	0	0	0	0	0	0	0	0	0	0	87
University of Nebraska-Lincoln	207	0	0	0	0	0	0	0	0	0	0	0	0	0	0	207
University of South Carolina	83	0	0	0	0	0	0	0	0	0	0	0	34	0	0	117
University of Texas, Southwestern Medical Center	14	0	0	0	0	0	0	0	0	0	0	0	0	0	19	33
Total	698*	34	39	43	51	22	18	10	21	17	22	13	34	72	19	1113*

* Two additional normal control subjects were collected from a site that subsequently withdrew from participation.

Chapter 7. Software Installation and Administration

The EXAMINER battery includes software for administering computer-based tasks and for generating executive composite and factor scores. The EXAMINER battery software is designed to work on multiple operating systems and to use open-source, readily available software. This chapter describes the requirements and dependencies of the EXAMINER software, provides detailed instructions on installing, configuring, and running the software.

Requirements

The EXAMINER software requires the following minimum hardware:

- 2GB RAM (Windows XP/Linux: 1GB RAM)
- 2 GHz Intel Core 2 Duo processor (Windows XP/Linux: 1.6 GHz Pentium M processor)
- At least 14" diagonal display (approximately 30 cm side-to-side minimum)
- Keyboard input device for responses
- 500 MB available hard drive space for software files and
- 0.25 MB available hard drive space for output data files per administration of the battery.

The EXAMINER software requires the following minimum software:

- Windows XP (Service Pack 3), Windows 7, Apple OS X 10.6, or Ubuntu 10.4.
- PsychoPy Version 1.73.05
- R (Statistical Software) Version 2.14 included on EXAMINER distribution CD.

Acquiring and Installing Software Dependencies

The EXAMINER battery is distributed with copies of the installation files for the PsychoPy and R software dependencies. These files are located in the SOFTWARE directory of the EXAMINER battery distribution, and are further organized by the target operating system. Person's responsible for installing and configuring the EXAMINER battery are encouraged to review the web sites for these software dependencies for updated versions of the software and for detailed installation and configuration instructions. The EXAMINER battery has been tested with the versions of the software dependencies available on April 15, 2011. Use of updated versions of software dependencies may require additional testing and verification.

You will complete software installation in five steps:

1. Install PsychoPy and update to version 1.71.05
2. Install EXAMINER computer tasks
3. Configure EXAMINER computer tasks
4. Run EXAMINER computer tasks
5. Install R

1. Install PsychoPy

PsychoPy is an open-source application to allow the presentation of stimuli and collection of data for a wide range of neuroscience, psychology and psychophysics experiments. The EXAMINER battery computer tasks are designed to run within the PsychoPy application, and PsychoPy must be installed on every computer that will run the EXAMINER battery computer tasks.

As of April 15, 2011 the current web site for acquiring the PsychoPy software is www.psychopy.org, and the current released version of PsychoPy is 1.73.05. For Windows and OS X operating systems, the easiest way to install PsychoPy is to download the current “Standalone” version of PsychoPy (1.73.02), and to upgrade this standalone version to 1.73.05 using the built in PsychoPy updater, included in your EXAMINER installation package. For Debian based Linux operating systems (e.g., Ubuntu) you can use the packages located at neuro.debian.net.

Therefore, you will need to install three things: PsychoPy software, EXAMINER tasks to run on PsychoPy, and R. All three of these items are included on the EXAMINER distribution disk and installation instructions are detailed below.

Installing PsychoPy

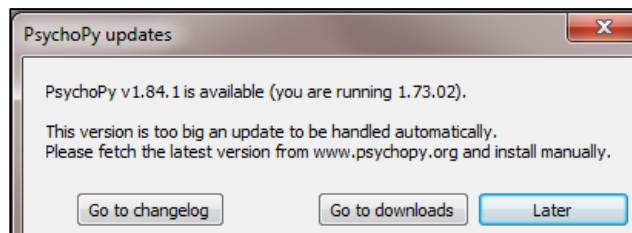
on Windows

1. Download the copy included in the directory of the EXAMINER battery distribution in Examiner 3.6/Software/Windows/Tasks/PsychoPy. Alternatively, you can download the standalone PsychoPy Installer from code.google.com/p/psychopy.
2. Ensure that you have permission on the testing machine to install software (i.e., “Administrative Privileges”).
3. Install PsychoPy by running the downloaded installer called “StandalonePsychopy-1.73.02-win32.exe” and accept License Agreement and all the default settings.
4. Installation will take approximately five minutes.
5. Additional installation instructions for PsychoPy are located at www.psychopy.org/installation.html.
6. Ensure that the installation works by running the PsychoPy2 application.

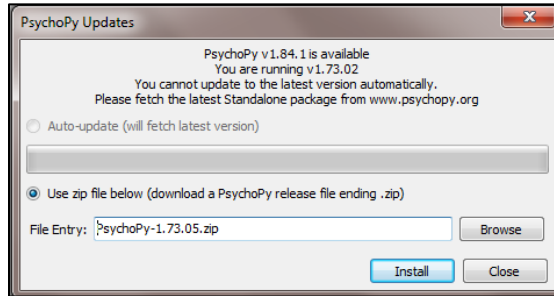
*On some Windows XP machines you may get this error message:
**This application has failed to start because the application configuration is incorrect.
Reinstalling the application may fix this problem.***

*To resolve this error, you can download and install some additional software from Microsoft currently located at:
www.microsoft.com/downloads/en/details.aspx?familyid=A5C84275-3B97-4AB7-A40D-3802B2AF5FC2&displaylang=en*

7. PsychoPy will automatically alert the user to updates that are available (as shown in the screenshot below). Select **Later** and perform a manual upgrade.



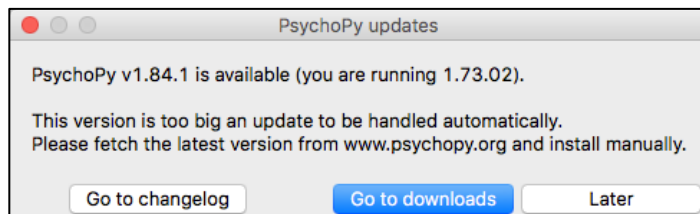
- a. To perform a manual upgrade, select **PsychoPy Updates** from the **Tools** menu.



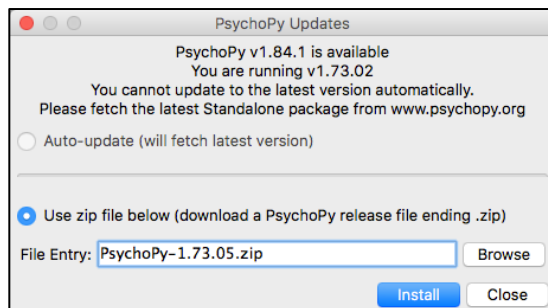
- b. Choose the option to use a zip file, and browse to select the **PsychoPy-1.73.05.zip** located in the **/EXAMINER 3.6/Software/Windows/Tasks/Psychopy** folder of the EXAMINER battery distribution.
- c. Click the **Install** button and restart PsychoPy.

on Mac OS X

1. Download from **/EXAMINER 3.6/Software/OSX/Tasks/Psychopy** directory of the EXAMINER battery distribution.
2. Install PsychoPy by opening the dmg file (disk image) and dragging **PsychoPy2.app** to your **Applications** folder.
3. If needed, additional installation instructions for PsychoPy are located at www.psychopy.org/installation.html.
4. Ensure that the installation works by running the PsychoPy2 application.
5. PsychoPy will automatically alert the user to updates that are available (as shown in the screenshot below). Select **Later** and perform a manual upgrade.



- a. To perform a manual upgrade, select **PsychoPy Updates** from the **Tools** menu.



- b. Choose the option to use a zip file, and browse to select the **PsychoPy-1.73.05.zip** located in the **/EXAMINER 3.6/Software/OSX/Tasks/Psychopy** folder of the EXAMINER battery distribution.
- c. Click the **Install** button and restart PsychoPy.

on Linux

Using a Linux operating system for the EXAMINER computer tasks requires knowledge of how to use the software packaging system for the specific Linux distribution in use to install software. Users of debian-based Linux distributions will find helpful installation instructions at www.psychopy.org/installation.html and will find installation packages for PsychoPy at neuro.debian.net. Copies of the installation packages for PsychoPy 1.73.05 for the Ubuntu 10.4 operating system are included with the EXAMINER battery distribution in the **/Examiner 3.6/Software/Linux/Tasks/Psychopy/ubuntu10.4** directory.

2. Install the EXAMINER Computer Tasks

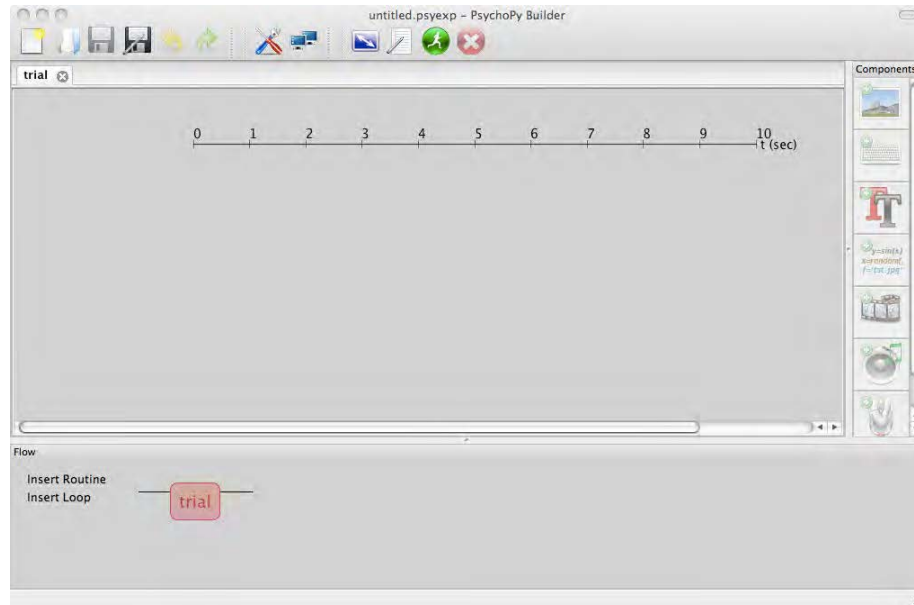
If you have already installed the EXAMINER Computer Tasks on the computer, please make sure to back up the existing **Examiner** folder before proceeding (this will ensure you don't accidentally erase any previously captured data or configuration files).

1. Install PsychoPy following the instructions for your operating system above.
2. Locate and copy the **Examiner** installation file for your operating system to the desktop of the computer where EXAMINER will be installed:
 - a. For Windows the installation file is **Examiner.zip** and is located in the **/EXAMINER 3.6/Software/Windows/Tasks** folder of the EXAMINER battery distribution.
 - b. For OS X the installation file is **Examiner.dmg** and is located in the **/EXAMINER 3.6/Software/OSX/Tasks** folder of the EXAMINER battery distribution.
 - c. For Linux the installation file is **Examiner.tar.gz** and is located in the **/EXAMINER 3.6/Software/Linux/Tasks** folder of the EXAMINER battery distribution.
3. Extract the contents of the **Examiner** installation file to your desktop. To extract, right click on zipped file and choose "extract all." When prompted for destination, choose desktop, and click "extract" button. This should create an **Examiner3_6** folder on your desktop.
*The desktop is the recommended location for the **Examiner3_6** installation folder.*

3. Configure the EXAMINER Computer Tasks

After installing the PsychoPy and Examiner Tasks software you need to configure the software to run on the specific testing machine.

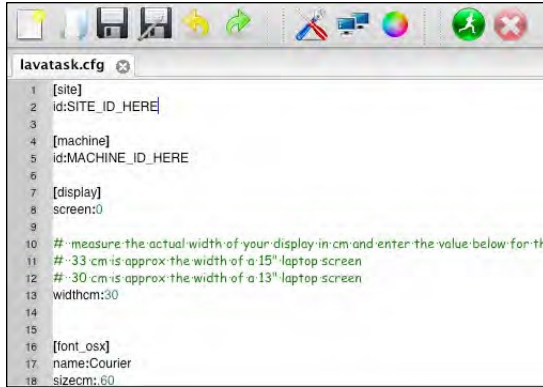
1. Run **PsychoPy** by double clicking on the **PsychoPy2** Icon under Applications (on Mac) or by selecting it from the Start Menu (on Windows).
2. The first time you open **PsychoPy** it will open to the Builder view (shown below).



3. Change to “Coder” view by selecting **Open Coder View** from the **View** menu.



4. Next, from within the **PsychoPy** coder view, open the lavatask.cfg file from the **[Desktop]\Examiner3_6** folder.
5. Enter a unique Site ID (could be the name of your institution, e.g., “UCSF”) in the configuration file, replacing the text that reads SITE_ID_HERE.
6. Enter a unique “Machine ID” in the configuration file, replacing the text that reads MACHINE_ID_HERE. This should be an id that will not change (e.g., the machine name).
7. Next, measure the horizontal width of the primary screen that will be used to display the EXAMINER Tasks to research participants. This is NOT the corner-to-corner width, but rather is the side-to-side width. Measure in centimeters, and enter the value, replacing the default value of 33 centimeters. You can use decimal points or not (e.g., 30, 31.5, 31.75) as needed. Failure to accurately measure and configure the display width of the computer used for the EXAMINER battery will result in an under- or over-sized representation of the tasks on the screen.



Template



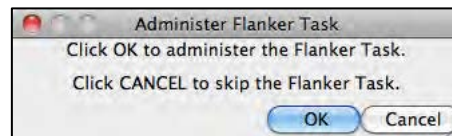
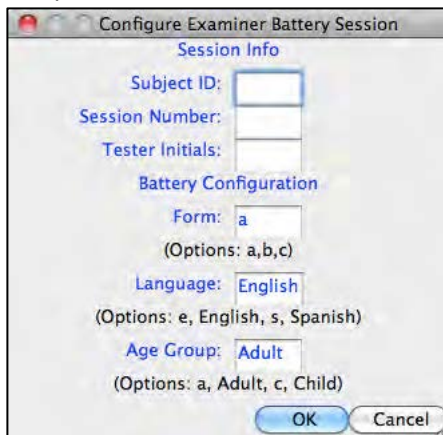
Completed Example

8. Save the configuration file.

4. Run the EXAMINER Computer Tasks

Before running the EXAMINER computer tasks, first configure the tasks as described above (this only needs to be done the first time the tasks are run).

1. Open the PsychoPy application
If PsychoPy opens in builder view, change to coder view (open coder view from the view menu)
2. Open **examiner_adult_english.py** from [Desktop]\Examiner3_6\scripts\examiner\ folder.
Additional scripts that default the configuration settings to child and/or Spanish language settings are also available in the same folder.
3. Click the run icon or selecting **Run** from the **Tools** menu. A black screen will appear while setting up, and the configure examiner battery session screen will appear in minutes.
4. Enter the task administration details and click the OK button to begin the EXAMINER battery computer tasks.



5. Before each task a dialog box will display with the option to run (OK) or skip (Cancel) each task.

EXAMINER Computer Task Output Files

By default, the Examiner Task software will create data files for each administration of an EXAMINER task. These files are stored in the [Desktop]\Examiner3_6\data folder. For tasks that record responses from the research participant, there will be a detail level file with each response recorded, a record in the summary file (one file per task type), and a record in an environment-monitoring file (one file per

testing machine). This [Desktop]\Examiner3_6\data folder should be backed up to protected storage frequently. There are no other copies of the test output data.

The output files are organized into subfolders of the data folder, one per subject id. A summary file for each task administration is stored in the subject id folder and is labeled with the following format:

[task_name]_Summary_[subject_id]_[session_num]_[date/time].csv

The detail level files are stored in the subject id folder and are labeled with the following format:

[task_name]_Summary_[subject_id]_[session_num]_[date/time].csv

A file that combines all the summary records for each task is stored in the data folder and is labeled with the following format:

[task_name]_Summary_Combined_[site_id]_[machine_name].csv

See the task documentation under [Desktop]\Examiner3_6\docs folder for detailed descriptions of each data file format and variable definitions.

5. Install R Software

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS. The EXAMINER battery software for calculating the executive composite and factor scores is programmed in the R language and relies on the ltm module (latent trait models under the Item Response Theory approach). The R software only needs to be installed on computers that will be used to calculate the executive composite and factor scores (e.g., a computer used for data management and analysis).

As of 12/28/2011 the current version of the R software is 2.14.1.

Installing R on Windows

1. Download the R Installer for Windows from www.r-project.org or use the copy included in the /Examiner 3.6/Software/Windows/Scoring/R directory of the EXAMINER battery distribution.
2. Ensure that you have permission on the testing machine to install software (i.e., "Administrative Privileges").
3. Install R by running the downloaded installer and accept all the default settings.
4. Additional installation instructions for R are located at www.r-project.org.
5. Install the "ltm" package:
 - a. Start the R application.
 - b. Select **Install package(s)** from the **Packages** menu.
 - c. Select **USA (CA 1)** from the list of CRAN mirrors. Click OK.
 - d. Select **ltm** from the list of Packages. Click OK. The **ltm** package and all dependencies will download and unpack; the **ltm** package installation is complete.

Installing R on Mac OS X

1. Download the R Installer for OS X from www.r-project.org or use the copy included in the /EXAMINER 3.6/Software/OSX/Scoring/R directory of the EXAMINER battery distribution.
2. Install R by running the downloaded package installer and accept all the default settings.
3. Additional installation instructions for R are located at www.r-project.org.

4. Ensure that the installation works by running the R application and installing the **ltm** package.
 - a. Start the R application.
 - b. Select **Package Installer** from the **Packages & Data** menu.
 - c. Click the **Get List** button.
 - d. Select **ltm** from the list of packages to install.
 - e. Click OK. If prompted, select the option to install all dependencies of the **ltm** package; otherwise R will install all dependencies automatically. The **ltm** package and all dependencies will download and unpack. The **ltm** package installation is complete.

Installing R on Linux

1. Download the R Installer for Linux from www.r-project.org or use the copy included in the **/EXAMINER 3.6/Software/OSX/Scoring/R** directory of the EXAMINER battery distribution.
2. Install R by running the downloaded package installer and accept all the default settings.
3. Additional installation instructions for R are located at www.r-project.org.

Chapter 8. Psychometric Properties

Administration Issues

The entire EXAMINER battery was designed to be administered to all subjects with a few notable exceptions. Phonemic verbal fluency, for example, could not be routinely administered to children under seven or eight because the task required a minimum degree of literacy. In addition, the spatial 2-back was a particularly challenging task that was not administered to subjects who either had considerable difficulty with the 1-back or who were unable to complete the 2-back sample and practice items. The children's version of the Social Norms Questionnaire was not ready until after data collection had started. Also, several subjects were administered short-forms of the battery (e.g., the 2-back was not included for some subjects). Finally, there were instances when a task could not be administered due to situation- or subject-specific issues (e.g., computer issues, sensory-motor deficits, lack of subject cooperation). The completion rates of EXAMINER tests are summarized in the table below. Data quality issues once a test was administered were relatively infrequent.

	3–7 yrs.	8–17 yrs.	18 yrs. and older	Total
Dot counting	84.4	97.5	93.4	94.0
1-back	83.3	95.9	92.0	92.5
2-back	1.0	58.7	79.3	66.3
Flanker	96.9	98.6	95.6	96.6
Set shifting	96.9	98.3	96.1	96.9
CPT	96.9	97.5	93.8	95.3
Anti-saccade	2.1	76.9	66.7	64.5
Verbal fluency	5.2	98.6	99.4	91.4
Category fluency	96.9	100.0	94.6	96.5
Unstructured task	95.8	98.9	97.4	97.8
Social norms questionnaire	46.9	34.7	74.8	59.9

Distributions

As described in Chapter 11, individual EXAMINER variables can be combined into composite and factor scores using item response theory. Procedures to create an Executive Composite and Working Memory, Cognitive Control, and Fluency Factors are laid out in Chapters 11 and 13. The distributions of the Executive Composite score and the Working Memory, Cognitive Control, and Fluency Factor scores were all normal when the sample was viewed in its entirety and when children and adults were viewed separately. Distributions of the Executive Composite score for children and adults are shown in Figure 11. Please note that the Executive Composite and Factor scores are raw scores generated directly from the EXAMINER scoring program and are *not* age-referenced.

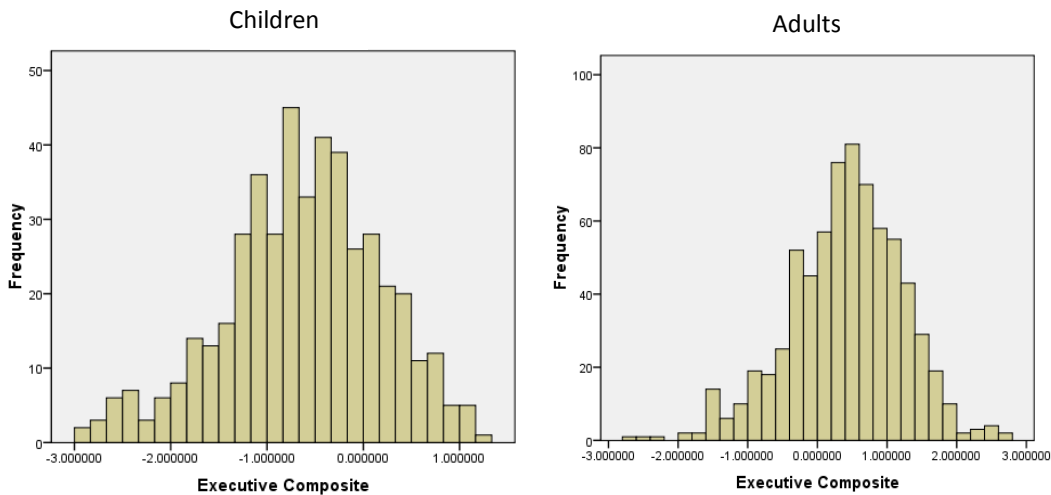


Figure 11. Distributions of Executive Composite score in children and adults

Reliability: Individual Tests

Reliability of individual tests was estimated in several ways depending on the nature of the task. Internal consistency measures were appropriate for all tasks except Unstructured Task and Social Norms; test-retest reliability is reported for these measures. In addition, we report inter-rater reliability estimates for Anti-Saccades, Verbal Fluency, and Behavior Rating.

Internal consistency

Estimates of internal consistency were derived for most individual EXAMINER tests using coefficient alpha.

Dot Counting

The Dot Counting Task originally had 12 trials, two trials for each possible screen length (two through seven). Coefficient alpha for the 12-item version is .81. When only one trial for each screen length is used, the coefficient alpha for the ensuing 6-item version is .69.

Table 4. Dot Counting (6-item)	
Age group	Dot Counting accuracy
<18	0.69
18+	0.65
Total	0.69

Flanker

Coefficient alphas for accuracy were calculated separately for congruent and incongruent trials and for accuracy and reaction time. These results are summarized in Table 5.

Table 5. Flanker reliability				
Age group	Congruent Flanker RT	Congruent Flanker trial accuracy	Incongruent Flanker RT	Incongruent Flanker trial accuracy
<18	0.97	0.80	0.97	0.93
18+	0.97	0.88	0.98	0.93
total	0.97	0.86	0.97	0.93

Continuous Performance Test

The CPT accuracy has an internal consistency reliability coefficient of .64 in children and .78 in adults.

Anti-saccade

There are two blocks of antisaccade trials, with 20 trials within each block. At the block level, the internal consistency reliability is .92.

Set shifting

Coefficient alphas for accuracy and reaction time were calculated separately for color, shape, and shifting blocks. Results are summarized in Table 6.

Age group	Color trials RT	Color trial accuracy	Shape trials RT	Shape trials accuracy	Set shift trials RT	Set shift trial accuracy
<18	0.95	0.86	0.94	0.93	0.97	0.88
18+	0.95	0.92	0.94	0.91	0.98	0.91
Total	0.95	0.89	0.94	0.92	0.97	0.91

Verbal fluency

Reliability estimates for verbal fluency were based on having two 1-minute trials for phonemic fluency and two 1-minute trials for semantic fluency. The reliability of phonemic fluency is .88, and the reliability of category fluency is .78. Reliabilities of phonemic fluency in the two alternate forms of EXAMINER are .91 and .97.

Inter-rater reliability

While most EXAMINER measures are straightforward measures of accuracy or reaction time, scoring of the verbal fluency and anti-saccade tasks often relies on examiner judgment. We report on the inter-rater reliability of these measures below.

Verbal fluency

Twenty trained raters each scored ten written verbal fluency protocols, indicating the number of correct words, rule violations, and repetitions. The intraclass coefficient (ICC) was calculated for correct words, rule violations and repetitions for combined letter fluency (F- and L- words) and combined category fluency (animals and vegetables) measures. Overall, results indicated high ICC values for number correct words in phonemic (ICC=.814) and semantic fluency (ICC=.984). ICC values for repetitions (.830 and .768) were also good, whereas rule violations (.552 and .779) were more variable.

	# Correct	Repetitions	Rule Violations
F words	.989	.811	.701
L words	.974	.730	.625
Animals	.979	.704	.319
Vegetables	.972	.827	.814
F + L words	.814	.830	.552
An. + Veg. words	.984	.768	.779

Anti-saccade

We videotaped the eye movements of six people who were instructed to make mistakes and atypical eye movements during the anti-saccade task. Three raters scored each eye movement. The ICC was .98.

Behavior rating scale

Two raters independently viewed and scored the videotaped testing sessions of 15 patients, including five patients with Alzheimer’s disease, five patients with behavioral variant frontotemporal dementia, and five patients with the semantic variant of primary progressive aphasia. Each patient was rated on the presence and degree of stimulus-boundedness, perseverations, lack of emotional engagement, and initiation. The ICC was 0.61 for a single rater, and 0.76 for two raters.

Test-Retest Reliability

Test-retest reliability was estimated for Unstructured Task and Social Norms Questionnaire on 85 normal subjects who were evaluated twice within 150 days. Results are summarized in Table 8.

	n	Reliability estimate
Unstructured Task	85	.71
Social Norms Questionnaire: Children	15	.93
Social Norms Questionnaire: Adults	52	.69

Reliability: Executive Composite and Factor Scores

Short-term test-retest reliability for the Executive Composite and Factor scores was assessed with 53 normal children (age range 7.2 to 12.5 years) and 32 normal adults (age range 21 to 76 years). The battery was administered twice within a 4-month period (range from 0 to 122 days). Test-retest reliabilities are summarized in Table 9.

The reliability of the Working Memory score was only .39 in the adults. The reliabilities of the individual working memory tests were .85 for 2-back d-prime, .56 for Dot Counting, and -.04 for 1-back d-prime. The poor reliability for 1-back was due to ceiling effects in the adult control sample, a relatively small sample, and restricted age range. Reliability for the full sample is .75, and higher reliability in a patient sample is expected.

Age group	Executive Composite	Cognitive Control	Working Memory	Fluency
<18	0.87	0.83	0.77	0.76
18+	0.78	0.73	0.39	0.86
Total	0.94	0.91	0.75	0.91

One-year test-retest reliability was evaluated in 108 normal children between the ages of 6 and 11. The test-retest interval ranged from 343 to 411 days (mean=361). Reliability of the Executive Composite over the one-year interval was 0.80. The Cognitive Control, Working Memory, and Fluency scores were 0.77, 0.52, and 0.73, respectively.

Chapter 10. Validity

Validity reflects how well a particular task measures what it purports to measure. In this section, we report first on the validity of the executive composite score, followed by validation studies of the individual factor scores. Several different approaches to assessing validity have been incorporated. The primary validation analysis was the relationship between EXAMINER scores and our external measures of day-to-day executive functioning, which were the FrSBe total score in adults, and the BRIEF total scores in children.

Executive Composite Score

Correlations with external measures of executive functioning

Baseline FrSBe scores were obtained on 219 adults who had informants available at the time of the EXAMINER visit and who agreed to complete the informant questionnaire. The correlation between the Composite Score and FrSBe total raw score, after partialling out the effect of age, was $-.57$ ($p < .001$), reflecting a fairly robust association between executive functioning on EXAMINER and estimates of real world executive function. This relationship remained highly significant ($-.48$, $p < .001$) even after controlling for estimates of baseline verbal ability (WRAT-3 Reading) and processing speed (WAIS-III Digit Symbol).

Baseline BRIEF scores were obtained on 404 children. The correlation between the Composite Score and BRIEF total raw score, partialling out the effect of age, was $-.21$ ($p < .001$).

Differences between patients and normal controls

Another basic validation set of analyses addressed how well the Composite Score separated patients from normal controls. Separate analyses were carried out for children and adults with baseline scores using general linear model controlling for age.

In the adult sample, there were 275 patients and 386 controls. The difference between patients and controls was significant ($F=41.6$, $p < .001$, partial $\eta^2=.06$). In the child sample, there were 139 patients and 310 controls. The difference between patients and controls was significant ($F=32.4$, $p < .001$, partial $\eta^2=.07$). Estimated marginal means are summarized in Table 10.

	Patients	Normals
	Mean (SE)	Mean (SE)
< 18	-.81 (.04)	-.50 (.03)
18+	.19 (.05)	.57 (.04)

Correlations with age

Executive function changes with age, so another approach to validating the Composite Score was to measure the correlation with age in normal older adults and children.

The correlation with age in older adults was assessed using baseline data from 212 normal controls over the age of 55. The correlation between the Composite Score and age was $-.30$ ($p < .001$).

The correlation with age in children was assessed using baseline data from 278 normal controls under the age of 15. The correlation between the Composite Score and age was $.83$ ($p < .001$) (see Figure 12).

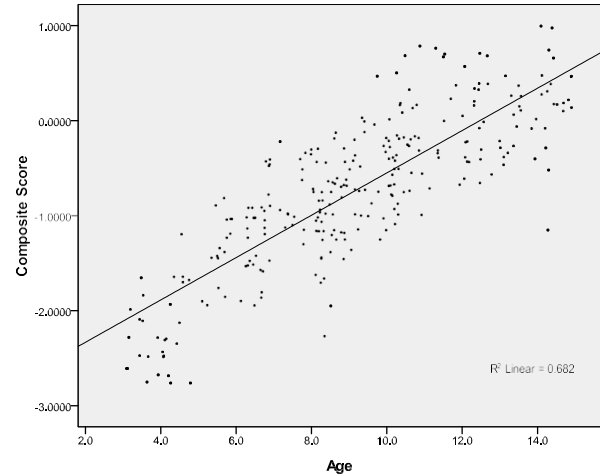


Figure 12. Executive Composite and age in children

Longitudinal Change

We also predicted that executive functioning improves significantly with age in normally developing children. Longitudinal change over a 12-month period was assessed in 108 normal children between the ages of 6 and 11. The rest-retest interval ranged from 343 to 411 days (mean=361). The increase in the Executive Composite score after 12-months was significant ($t = -10.3$, $p < .001$) and approximately one standard deviation in size; data are summarized in Table 11.

Table 11. Longitudinal change in children	
Baseline	12-month follow-up
Mean (sd)	Mean (sd)
-.83 (.55)	-.49 (.59)

Selected group comparisons

Lateral Frontal vs Medial Frontal vs Posterior focal lesion

Several studies have shown that lateral frontal structures are more critical for the cognitive features of executive functioning than medial frontal structures (Krueger et al., 2010). We tested this by evaluating 21 patients with focal lateral frontal lesions (mean age=58.4, sd=10.7), 21 patients with focal medial frontal lesions (mean age=55.7, sd=15.1) and 39 patients with temporal and parietal focal lesions (mean age=58.8, sd=11.3). As predicted, after controlling for age, lateral frontal lesion patients (mean=0.06, se=.14) performed less well on the Executive Composite score than focal posterior lesion patients (mean=.41, se=.10). Medial frontal patients performed more closely to the posterior lesion patients (mean=.40, se=.14).

Alzheimer's vs Mild Cognitive Impairment vs Controls

We predicted a gradient of scores, with AD patients performing least well, age-matched controls performing best, and patients with MCI performing in between. We compared 39 AD patients (mean age=72.4, sd=10.7), 38 MCI patients (mean age=74.0, sd=7.7), and 151 controls (mean age=70.8, sd=9.3) on the Executive Composite score.

Estimated marginal means are summarized in Table 12. Controls performed significantly better than AD and MCI, and MCI patients performed significantly better than AD patients.

Controls	MCI	AD
Mean (SE)	Mean (SE)	Mean (SE)
.70 (.05)	.29 (.11)	-.20 (.10)

Subcortical syndromes

Impairment in executive functioning is widely considered to be a central cognitive deficit in patients with syndromes affecting subcortical-frontal systems. Our subcortical group consisted of 17 patients with multiple sclerosis (mean age=42.5; sd=11.9), 21 patients with idiopathic Parkinson's disease (mean age=67.4; sd=7.0), and 13 patients with progressive supranuclear palsy (mean age=65.9; sd=7.6). When compared to 172 age-matched controls, the subcortical group performed significantly less well; estimated marginal means after controlling for age are summarized in Table 13.

	Subcortical group	Normals
	Mean (SE)	Mean (SE)
Executive Composite score	.15 (.10)	.86 (.05)

Convergent and Divergent Validity

Convergent and divergent validity was assessed in 74 normal adults who were also administered the California Verbal Learning Test-II (D.C. Delis, Kramer, Kaplan, & Ober, 2000) and the Stroop Interference Test. The correlation between the Executive Composite score and 20-minute delayed free recall was .05 ($p > .65$), whereas the correlation between the Executive Composite score and Stroop Interference was .58 ($p < .001$). The correlation between the Executive Composite score and Stroop Interference remained significant even after controlling for Stroop Color Naming ($r = .35$, $p < .005$), further supporting the Executive Composite score as a measure of executive ability.

Working Memory Factor Score

Correlations with external measures of executive functioning

The correlation between the Working Memory Factor Score and FrSBe total raw score, after partialling out the effect of age, was $-.54$ ($p < .001$), reflecting a fairly robust association between working memory on EXAMINER and estimates of real world executive control. This relationship remained highly significant ($r = -.35$, $p < .001$) even after controlling for estimates of baseline verbal ability (WRAT-3 Reading) and processing speed (WAIS-III Digit Symbol).

The Working Memory Factor correlated with the BRIEF total raw score $-.21$ ($p < .001$) and the BRIEF working memory subscore, $-.23$, $p < .001$, partialling out the effect of age.

Differences between patients and normal controls

To determine how well the Working Memory Composites separated patients from controls, separate analyses were carried out for children and adults with baseline scores using general linear model controlling for age.

In the adult sample, there were 262 patients and 241 controls. Estimated marginal means are summarized in Table 14. The difference between patients and controls was significant ($F=50.8$, $p<.001$, partial $\eta^2=.17$).

In the child sample, there were 119 patients and 368 controls. Estimated marginal means are summarized in Table 14. The difference between patients and controls was significant ($F=90.7$, $p<.001$, partial $\eta^2=.27$).

	Patients	Normals
	Mean (SE)	Mean (SE)
< 18	-.76 (.03)	-.23 (.03)
18+	.14 (.05)	.61 (.04)

Correlations with age

The correlation of the Working Memory Factor Score with age in adults was assessed using baseline data from 153 normal controls over the age of 55. The correlation between the Factor Score and age was $-.32$ ($p=.001$).

The correlation with age in children was assessed using baseline data from 341 normal controls under the age of 15. The correlation between the Factor Score and age was $.53$ ($p<.001$).

Longitudinal Change

Longitudinal change in Working Memory Composite over a 12-month period in 95 normal children between the ages of 6 and 11 was assessed. The increase in the Working Memory Factor after 12-months was significant ($t=-2.84$, $p=.006$).

Baseline	12-month follow-up
Mean (sd)	Mean (sd)
-.51 (.79)	-.29 (.77)

Selected analyses

Prodromal Huntington's Disease

We evaluated 13 patients with prodromal Huntington's disease (HD) (mean age=47.1, $sd=12.8$) and compared them to a sample of 50 healthy individuals who were matched on age, education, and sex. The prodromal HD patients were impaired on the Working Memory Factor ($p=.008$). In addition, the Working Memory Factor evidenced a trend in association with disease burden "CAP" scores in the prodromal patients, $r=-.53$,

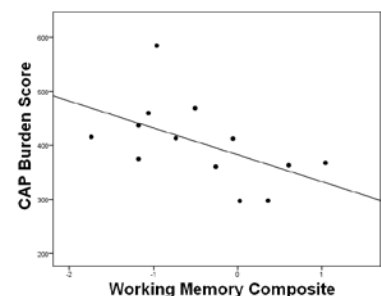


Figure 13. Working Memory and CAP

$p=.064$, indicating that it may be a good candidate for marking early disease progression (see Figure 13).

To determine whether the Working Memory Factor is a sensitive marker of early HD-related changes in striatal volume, volumetric 3-tesla T1 imaging with Freesurfer segmentation of the striatum was performed on 12 patients with prodromal HD. In addition to the EXAMINER, patients were administered measures of language (category fluency), visuospatial skills (Benson Figure copy), and memory (Benson Figure delayed recall) for comparison. The Working Memory Factor correlated with the left striatum, $r=.81$, $p=.001$, and the right striatum, $r=.66$, $p=.019$ (see Figure 14). In contrast, correlations with measures of language, visuospatial skills, and memory were not significantly associated with striatal volumes.

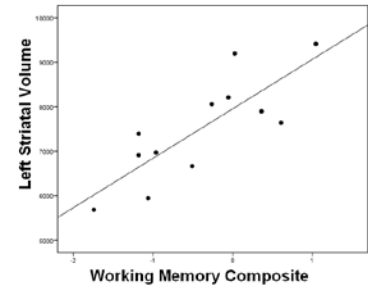


Figure 14. WM and striatum volume

Attention Deficit Disorder

We predicted that children diagnosed with attention deficit hyperactivity disorder (ADHD) would show an impairment on the Working Memory Factor and that their performance would correlate with learning

problems, as measured by the Conners Comprehensive Behavior Rating Scales. We compared 32 ADHD children between the ages of 8 and 16 (mean=11.1, $sd=2.3$) to 164 age-matched healthy controls. The ADHD children performed less well on the Working Memory Composite, $F(1, 193)=6.5$, $p=.01$. Working memory performance correlated with learning problems, $r=-.46$, $p=.009$, controlling for age (Figure 15).

Controls	ADHD
Mean (SE)	Mean (SE)
-.17 (.05)	-.48 (.11)

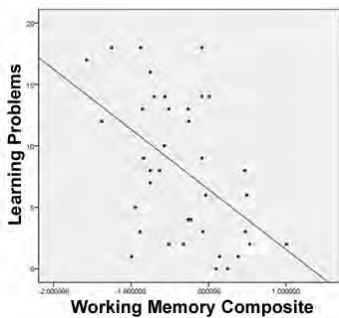


Figure 15. Working Memory and Connors

Parkinson's disease vs Controls

We predicted that nondemented patients with Parkinson's disease (PD) would be impaired on the Working Memory Composite. We compared 20 PD patients (mean age=67.4, $sd=7.5$) to 125 controls (mean age=68.2, $sd=7.1$). Estimated marginal means are summarized in Table 17. Controls performed significantly better than PD patients, $F(1, 142)=9.0$, $p=.003$, controlling for age.

Controls	PD
Mean (SE)	Mean (SE)
.59 (.07)	.06 (.16)

Convergent and Divergent Validity

Convergent and divergent validity was assessed in 96 normal adults who were also administered the California Verbal Learning Test-II (Delis, et al., 2000) to measure verbal episodic memory, the Digit Span Backward test to measure verbal working memory, and the modified Boston Naming Test to measure confrontation naming. Controlling for age, the correlation with Digits Backward, $r=.38$, $p<.001$, was greater than the correlation with long delay free recall, $r=.21$, $p=.04$. The correlation with Digit Span remained significant even after controlling for long delay free recall, $r=.36$, $p<.001$, supporting the Working Memory Factor as a measure of working memory. The Working Memory Factor did not correlate with the Boston Naming Test of confrontation naming, $r=.09$, $p=.40$.

Cognitive Control Factor Score

Correlations with external measures of executive functioning

The correlation between the Cognitive Control Factor Score and FrSBe total raw score, after partialling out the effect of age, was $-.56$ ($p < .001$), reflecting a fairly robust association between Cognitive Control on EXAMINER and estimates of real world executive function.

The Cognitive Control Factor correlated with the BRIEF total score $-.22$ ($p < .001$), Shifting Scale, $-.20$, ($p < .001$), and Inhibition Scale $-.22$, ($p < .001$) partialling out the effect of age.

Differences between patients and normal controls

To determine how well the Cognitive Control Factor separated patients from controls, separate analyses were carried out for children and adults with baseline scores using a general linear model controlling for age.

In the adult sample, there were 262 patients and 241 controls. The difference between patients and controls was significant ($F=16.9$, $p < .001$, partial $\eta^2=.03$). In the child sample, there were 119 patients and 368 controls. The difference between patients and controls was significant ($F=37.0$, $p < .001$, partial $\eta^2=.08$). Estimated marginal means are summarized in Table 18.

	Patients	Normals
	Mean (SE)	Mean (SE)
< 18	$-.76$ (.03)	$-.23$ (.03)
18+	$.14$ (.05)	$.61$ (.04)

Correlations with age

The correlation of the Cognitive Control Factor with age in adults was assessed using baseline data from 153 normal controls over the age of 55. The correlation between the Factor Score and age was $-.41$ ($p=.001$). The correlation with age in children was assessed using baseline data from 341 normal controls under the age of 15. The correlation between the Cognitive Control Factor Score and age was $.78$ ($p < .001$).

Longitudinal Change

Longitudinal change in the Cognitive Control Factor Score over a 12-month period in 95 normal children between the ages of 6 and 11 was assessed. The increase in the Cognitive Control Factor after 12 months was significant and approximately a half a standard deviation in size ($t=-2.84$, $p=.006$).

Baseline	12-month follow-up
Mean (sd)	Mean (sd)
$-.72$ (.62)	$-.32$ (.66)

Convergent and Divergent Validity

Convergent and divergent validity was assessed in 75 normal adults who were also administered the California Verbal Learning Test-II (Delis, et al., 2000) to measure verbal episodic memory, Boston Naming Test to measure naming, and the Stroop Interference Test as a marker of cognitive control. Controlling for age, the correlation with Stroop Interference was .55 ($p < .001$), whereas the correlations with CVLT-II 20-minute delayed free recall ($r = -.08$) and Boston Naming Test ($r = .07$) were not significant.

Fluency Factor Score

Correlations with external measures of executive functioning

The correlation between the Fluency Factor Score and FrSBe total raw score for adults ($n = 214$), after parceling out the effect of age, was $-.43$ ($p < .001$). This relationship remained significant ($-.28$, $p < .001$) even after controlling for estimates of baseline verbal ability (WRAT-3 Reading) and processing speed (WAIS-III Digit Symbol).

Baseline BRIEF scores were obtained on 398 children. The correlation between the Factor Score and BRIEF total raw score, partialling out the effect of age, was $-.09$ ($p = .06$).

Differences between patients and normal controls

In the adult sample, there were 275 patients and 385 controls. The difference between patients and controls was significant ($F = 19.9$, $p < .001$, partial $\eta^2 = .03$). In the child sample, there were 139 patients and 309 controls. The difference between patients and controls was significant ($F = 9.39$, $p = .002$, partial $\eta^2 = .02$). Estimated marginal means are summarized in Table 20.

	Patients	Normals
	Mean (SE)	Mean (SE)
< 18	-.76 (.05)	-.58 (.03)
18+	.21 (.05)	.51 (.04)

Correlations with age

The correlation of the Fluency Factor with age in adults was assessed using baseline data from 211 normal controls over the age of 55. The correlation between the Factor Score and age was $-.13$ ($p = .07$). The correlation with age in children was assessed using baseline data from 277 normal controls under the age of 15. The correlation between the Factor Score and age was $.70$ ($p < .001$).

Correlations with brain MRI

Several studies have implicated left frontal regions as the primary underlying neuroanatomy for verbal fluency. We examined the neuroanatomical correlates of the Fluency Factor in 152 subjects who underwent a research brain MRI within 90 days of fluency data collection. The T1 MPRAGE structural MR images were analyzed using Freesurfer, a freely available and semi-automated parcellation system for calculating the volumes of specified regions of interest. We calculated cortical volumes for the left frontal, temporal, and parietal lobes. In a multiple regression analysis, with the Fluency Factor Score as the dependent variable, the three brain regions together explained an additional 8% of the variance

above that of age, MMSE, and intracranial volume. Only frontal volumes remained in the model, however ($\beta=.31, p<.001$). In a follow-up regression, with parietal and temporal volumes forced into the model first, left frontal volumes uniquely explained an additional 4.7% of the variance.

Longitudinal Change

Longitudinal change in the Fluency Factor Score over a 12-month period in 115 normal children between the ages of 6 and 11 was assessed. The increase in the Fluency Factor after 12 months was significant and approximately a half a standard deviation in size ($t=-5.91, p.001$).

Baseline	12-month follow-up
Mean (sd)	Mean (sd)
-.77 (.53)	-.54 (.57)

Selected group comparisons

Lateral Frontal vs Medial Frontal vs Posterior focal lesion

We evaluated 21 patients with focal lateral frontal lesions (mean age=58.4, sd=10.7), 21 patients with focal medial frontal lesions (mean age=55.7, sd=15.1) and 39 patients with temporal and parietal focal lesions (mean age=58.8, sd=11.3). As predicted, lateral frontal lesion patients (estimated marginal mean=-.296, se=.17) performed less well on the Fluency Factor Score than focal posterior lesion patients (mean=-.48, se=.12; $p<.001$). Medial frontal patients performed more closely to the posterior lesion patient (mean=.41, se=.17).

bvFTD vs Alzheimer's vs Controls

We predicted that patients with bvFTD would perform worse on the Fluency Factor relative to patients with Alzheimer's disease, who in turn would perform worse than controls. We compared 21 bvFTD patients (mean age=66.22, sd=7.7), 39 AD patients (mean age=72.4, sd=10.7), and 231 controls (mean age=69.5, sd=9.9) on the Fluency Factor Score, controlling for age. Estimated marginal means are summarized in Table 22. Controls performed significantly better than bvFTD and AD, and bvFTD patients demonstrated the most impoverished fluency profile.

bvFTD	AD	Controls
Mean (SE)	Mean (SE)	Mean (SE)
-.39 (.18)	.07 (.13)	.61 (.05)

Convergent and Divergent Validity

Fluency Factor Scores were calculated on a group of 302 older subjects with a mix of neurodegenerative diagnoses (e.g., AD, bvFTD, MCI, movement disorders, progressive aphasia) and normal controls. We predicted that the Fluency Factor would correlate most strongly with design fluency (Delis et al., 2001), moderately with other measures of working memory and executive functioning (e.g., backward digit span, Stroop Interference), and least strongly with episodic memory (10-minute delayed recall of the CVLT-SF), naming (Boston Naming Test) and spatial ability (Benson figure copy). The sample had a mean age of 65.1 years (sd=8.5) and a mean MMSE of 26.6 (sd=3.2). Age and MMSE were included as covariates. Results were consistent with expectations and support the convergent and divergent validity of the Fluency Factor (see Table 23).

Table 23. Correlations between Fluency Factor and other cognitive tasks					
	Design fluency	Digits backward	Stroop	Delay recall	Benson copy
Fluency Factor	.36	.29	.30	.11	.04
p value	<.001	<.001	<.001	.048	.496

Measures not included in factor scores

Unstructured Task

The weighted composite score on the Unstructured Task correlates well with FrSBe ($r=-.29$), separates adult patients from controls ($F=11.2$, $p<.005$), and correlates well with age in normal children ($r=.70$) and older adults ($r=-.35$).

Insight

Krueger et al. (2011) studied 91 children (mean age 11.9 years, $sd=1.7$), including 47 typically developing children and 44 patients. Per standard EXAMINER procedures, after completion of each fluency task, each child assessed his or her performance relative to a hypothetical sample of children of the same age, sex, and education. The relationship between self-appraisal accuracy and parents' rating of executive behavior (BRIEF Shifting score) was significant when examining typically developing children alone with ($r = .40$, $p = .006$) and without ($r = .41$, $p = .006$) controlling for actual performance.

Social Norms Questionnaire

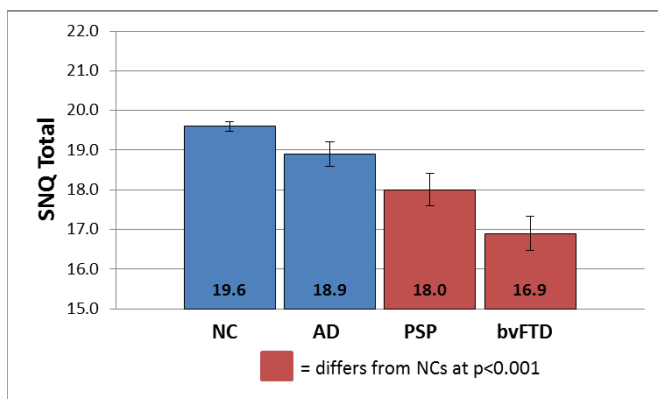
Content Validity

The initial phase of validity testing for the SNQ was to determine whether healthy control subjects agreed about the implicit social rules represented in each item. The preliminary adult version included 24 items that were administered to a validation sample of 35 healthy controls to determine agreement. Two items were rejected because agreement among controls was low, resulting in a 22-item version (see Table 20). Final level of agreement for each item was determined based on a larger sample of healthy controls ($N=203$) ranging in age from 18–94. There were no significant age or sex differences in SNQ performance.

Item	“Would it be socially acceptable to...”	Subscale	Initial Validation Sample (NC=35)	Final Validation Sample (NC=203)
1	Tell a stranger you don’t like their hairstyle?	Break	97.0	95.6
2	Spit on the floor?	Break	100.0	97.5
3	Blow your nose in public?	Over	74.0	65.5
4	Ask a coworker their age?	Break	80.0	76.8
5	Cry during a movie at the theater?	Over	97.0	95.6
6	Cut in line if you are in a hurry?	Break	100.0	97.5
7	Laugh when you yourself trip and fall?	Over	89.0	90.5
8	Eat pasta with your fingers?	Break	89.0	93.5
--	Hug an acquaintance without asking first?	Break	46.0	--
9	Tell a coworker your age?	Over	80.0	87.2
10	Tell someone your opinion of a movie they haven’t seen?	Over	86.0	72.8
11	Laugh when someone else trips and falls?	Break	94.0	91.6
12	Wear the same shirt every day?	Break	94.0	91.6
13	Keep money you find on the sidewalk?	Over	77.0	78.6
14	Pick your nose in public?	Break	100.0	97.0
15	Tell a coworker you think they are overweight?	Break	97.0	96.1
--	Drive fast if you are in a hurry?	Break	51.0	--
16	Eat ribs with your fingers?	Over	100.0	95.0
17	Tell a stranger you like their hairstyle?	Over	83.0	85.7
18	Wear the same shirt twice in two weeks?	Over	97.0	95.6
19	Tell someone the ending of a movie they haven’t seen?	Break	91.0	96.1
20	Hug a stranger without asking first?	Break	100.0	92.6
21	Talk out loud during a movie at the theater?	Break	97.0	96.1
22	Tell a coworker you think they have lost weight?	Over	83.0	83.3

Differences between patients and normal controls

To determine how well the Social Norms Questionnaire separated patients from controls, separate analyses were carried out for children and adults with baseline scores using general linear model controlling for age.



In the adult sample, there were 242 patients and 202 controls. The difference between patients and controls was significant (F=11.5, p<.001, partial eta²=.03).

In additional selected comparisons between healthy controls and patients with neurodegenerative disease, patients performed significantly worse than controls overall (F=16.74, p<0.001), but this effect was driven

Figure 16. SNQ performance in controls and patients with neurodegenerative disease

particularly by patients with diseases known to cause social deficits (i.e., behavioral variant frontotemporal dementia and progressive supranuclear palsy), while patients with Alzheimer's disease performed normally despite their cognitive deficits (see Figure 16).

Convergent and Divergent Validity

Convergent and divergent validity for the adult form of the SNQ were assessed in a large sample of diagnostically mixed adults. Correlation with tests of dysexecutive behavior: Controlling for age and sex, the correlation with FrSBe total score in 178 subjects was $r = -0.39$ ($p < .001$), indicating that poorer ability to evaluate social norms on the SNQ predicts higher levels of dysexecutive behavior in daily life according to informant ratings, including apathy, disinhibition, and executive dysfunction. Correlation with tests of social behavior/social sensitivity: SNQ total score also correlated at $r = 0.22$ ($p < .001$) with informant ratings of 212 subjects' empathy using the Interpersonal Reactivity Index (IRI), indicating that individuals with lower SNQ scores had decreased ability to take another person's perspective in an empathic manner in daily life. It also correlated at $r = 0.32$ ($p < .001$) with ability to detect subtle social cues according to Revised Self-Monitoring Scale (RSMS) score in 204 subjects. The SNQ correlated at $r = 0.33$ ($p < .001$) with ability to read emotions in 168 subjects who also were tested using the Emotion Evaluation Test from The Awareness of Social Inference Task (TASIT-EET). It also showed a weak but significant negative association ($r = -0.16$, $p < .01$) with the total score for the Behavior Rating Scale in 433 subjects, indicating that subjects who performed worse in the SNQ were rated by examiners as having displayed more behavioral abnormalities during the Examiner testing session. Thus, the psychometric construct measured by the SNQ overlaps with both executive functioning and socio-emotional sensitivity, and predicts real-life behavior abnormalities observed by clinicians and informants.

Chapter 11. Variable and Scale Construction

The tasks contained in the EXAMINER battery are designed to be used either individually or combined into scales. In this chapter, we review the dependent measures that can be derived from individual tests, and describe the methods underlying scale construction. For more detailed descriptions of variables produced by the computer-based tasks, refer to the task documentation under the [Desktop]\Examiner3_6\docs folder for descriptions of all variables included in the data output files.

Individual Tests

Dot Counting

In Dot Counting, each of the six trials is scored as per the procedures outlined in Chapter 12, and the primary dependent variable is the total score summed across the six trials.

Spatial 1 and 2-back

Total Correct and *D-prime* are the primary indices for characterizing accuracy on the n-back tasks. In the 1-back, there are 30 trials (10 “yes” trials and 20 “no” trials), and in the 2-back, there are 90 trials (30 “yes” and 60 “no”). The total number of correct responses for each task is tabulated. Because there is a 50% chance of making a correct response by chance, total scores less than 10 on the 1-back and less than 30 on the 2-back should be carefully scrutinized. The EXAMINER scoring program also generates measures of discriminability (d-prime) using signal detection parameters, where 1-back d-prime is the difference between the z-transforms of the hit rate ($[\text{number of hits} + .05]/11$) and the false positive rate ($[\text{false positives} + .5]/21$). D-prime for 2-back is based on a hit rate $= (\text{number of hits} + .5)/31$ and a false positive rate $= (\text{number of false positives} + .5)/61$. The primary dependent variables that contribute to the Working Memory Factor and Executive Composite scores are the 1-back and 2-back d-prime scores.

There are additional process variables that may be of interest to some users.

Response bias is the mean of the hit rate and false positive rate z-transforms. Higher scores represent a “yes” bias. Note that d-prime but not total correct is independent of response bias.

Interference Trials. On the 2-back, there are 8 trials on which the current square does not match the comparison square 2 before, but it does match the square 1 before. Elevated errors on these interference trials may represent difficulties with inhibitory control.

Similarity Effect. Each trial consists of the display of a “probe” square in one of 15 possible locations equidistant from the center of the screen. On “no” trials, the probe can differ from the location of the comparison “target” square by 1, 2, 3, or 4 positions. Accuracy can be evaluated in terms of this similarity effect. Although all examinees are expected to have more difficulty on similar trials than dissimilar trials, elevated errors on the similar trials may reflect impairment in accurately representing spatial information.

Visual Hemifield. Performance can be evaluated separately for left hemifield targets or right hemifield targets. Elevated errors for one hemifield may reflect difficulty attending to or processing information in that hemifield (e.g., neglect).

Flanker

In addition to storing individual trial data, the EXAMINER scoring program generates several Flanker summary scores. The first set of scores consists of the accuracy and median reaction times for the congruent and incongruent trials. Researchers often use the relationship between the incongruent and congruent trials (e.g., incongruent RT regressed over the congruent RT) as a marker of cognitive control, although careful attention should be paid to the accuracy rates to ensure that they are high enough to render the reaction time data meaningful.

Scores that contrast congruent and incongruent RT data often do not have sufficient reliability to make them suitable for clinical trial use. Accordingly, using an approach similar to that taken by the NIH Toolbox, EXAMINER also generates a more reliable variable that combines reaction time and accuracy data from the incongruent trials. This scoring method generates accuracy and reaction time sub-scales that range in value between 0 and 5, and the ultimate composite score ranges in value between 0 and 10. The accuracy score is the proportion of correct responses (out of 24 trials), multiplied by 5 to create a range from 0 to 5.

Like the accuracy score, the reaction time score ranges from 0 to 5 points. Because RT data is often positively skewed, a log (Base 10) transform is applied to the median RT score. To further reduce skewing, we set the minimum RT to 500 msec and the maximum reaction time to 3,000 msec; scores that fall outside that range are truncated (e.g., an RT of 4000 msec is set equal to 3000 msec). The range of 500-3000 msec was based on analyses of our full sample. Log values are algebraically rescaled from a $\log(500) - \log(3,000)$ range to a 0-5 range. Faster RTs result in a higher score. The formula for rescaling is:

$$\text{Reaction Time Score} = 5 - \left(5 * \left[\frac{\log RT - \log(500)}{\log(3000) - \log(500)} \right] \right)$$

The total Flanker score is the sum of the accuracy and reaction time scores. This Flanker score is the variable that contributes to the Executive Composite and the Cognitive Control factor score.

Continuous Performance Test

This version of the CPT contains 80% target trials, and is designed to elicit false alarm errors. The primary dependent measure from the task is the total number of false alarm errors, and this variable contributes to the total error score, which in turn is a donor scale to the Executive Composite and Cognitive Control scores.

Investigators can derive several other measures in addition to the primary dependent measure. Accuracy and reaction time are recorded for each trial, so response speed and omission data are available. In addition, the 100 trials of the CPT are divided into four blocks of 25, so rates of change in response speed and accuracy can be derived.

Anti-Saccade Test

Individual trial data as well as separate total scores for the prosaccade trial and two anti-saccade trials can be derived. The primary dependent measure that contributes to the Executive Composite and Cognitive Control scores is the total number of correct responses on the two anti-saccade trials.

Set Shifting

The EXAMINER scoring program generates several summary scores. There are three blocks of trials, one in which the matching criterion is only color, one in which the matching criterion is only shape, and one in which the matching criterion shifts pseudo-randomly. Accuracy and reaction time are recorded for each trial. In addition, total accuracy and median RT are generated for each block, and within the shifting block, total accuracy and median RT are generated for trials in which the matching criterion is the same as the previous trial, and for trials in which the matching criterion is different from the previous trial. Investigators can use the block data to measure general shift costs (i.e., accuracy and RT on the shifting block relative to the non-shift blocks) and specific shift costs (i.e., within the shifting block, accuracy and RT on the shift trials relative to accuracy and RT on the non-shift trials). When using the data in this way, investigators should be certain that the accuracy rates are high enough to render the reaction time data meaningful.

EXAMINER's approach to combining reaction time and accuracy data for Set Shifting is similar to the approach taken for the Flanker. This scoring method generates accuracy and reaction time sub-scales that range in value between 0 and 5, and the ultimate composite score ranges in value between 0 and 10. The accuracy score is the proportion of correct responses in the shifting block (out of 64 trials), multiplied by 5 to create a range from 0 to 5.

Like the accuracy score, the reaction time score ranges from 0 to 5 points. Because RT data is often positively skewed, a log (Base 10) transform is applied to the median RT score. After careful analysis of existing data, we further reduced skewing by setting the minimum RT to 400 msec and the maximum reaction time to 2800 msec; scores that fall outside that range are truncated (e.g., an RT of 4000 msec is set equal to 2800 msec). Log values are algebraically rescaled from a $\log(400) - \log(2800)$ range to a 0-5 range. Faster RTs result in a higher score. The formula for rescaling is:

$$\text{Reaction Time Score} = 5 - \left(5 * \left[\frac{\log RT - \log(400)}{\log(2800) - \log(400)} \right] \right)$$

The total Set Shifting score is the sum of the accuracy and reaction time scores. This Set Shifting score is the variable that contributes to the Executive Composite and the Cognitive Control factor score.

Verbal fluency

There are four verbal fluency tasks available: two phonemic (letter) fluency and two semantic (category) fluency tasks. Each of these tasks generates three measures: total correct, total repetitions, and total rule violations. The total correct scores contribute to the Executive Composite and Fluency Factor scores. The repetition and rule violation scores contribute to the Dysexecutive Errors variable, which in turn contributes to the Executive Composite and Cognitive Control Factor scores.

Unstructured Task

The Unstructured Task generates two variables, the total number of points earned, and the percentage of completed puzzles that were considered high value items. Both of these variables are normally distributed, correlated with FrSBe Total scores, and modestly correlated with one another ($r=.46$).

One goal of the Unstructured Task is to measure strategic planning. A possible disadvantage of relying solely on the total number of points earned is that examinees who are fast can still earn a high number of points, even if they make many poor choices. On the other hand, subjects who solve only one or two puzzles can potentially have inflated values for the percentage of high value puzzles. Accordingly, a weighted composite combining the two values is generated using the following formula: $\text{Weighted composite} = \text{UTpct} * \log_{10}(\text{UTTtotal} + 1)$, where UTpct is the percentage of completed puzzles that were considered high value items, and UTTtotal is the total number of points earned. The weighted composite is also normally distributed (see Figure 17). The Unstructured Task did not fit well with either the one-factor or three-factor model, and thus does not contribute to any of the IRT measures.

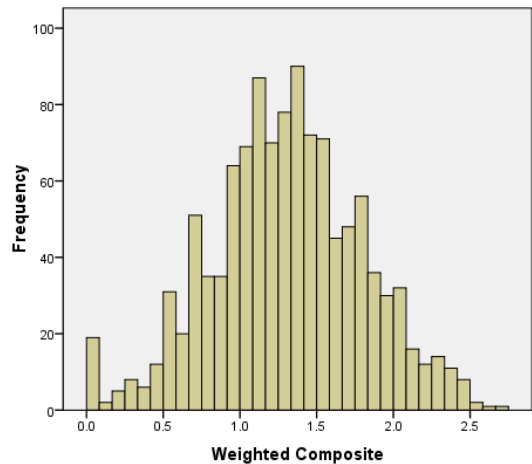


Figure 17. Distribution of Unstructured Task score

Social Norms Questionnaire

The Social Norms Questionnaire provides a total score reflecting the number of items endorsed in the direction of mainstream American culture.

Factor Scores

Confirmatory factor analysis methods were used to identify homogeneous dimensions underlying the executive function measures that could be used to generate composite scores. An iterative process involving exploratory and confirmatory factor analyses was carried out to identify potential donor variables from all EXAMINER tasks. The process identified 11 core variables for further analyses: Dot Counting total, d-prime measures from the 1-back and 2-back, total Flanker score, total Set Shifting score, anti-saccade total, total correct responses from each of the verbal fluency tasks, and total dysexecutive errors. Confirmatory factor analyses were then applied to these 11 variables. Two models were tested. The first was a one-factor model with all 11 donor variables. This unidimensional model was compared with a three-factor model guided by *a priori* conceptual considerations, corresponding to cognitive control (total Flanker score, total Set Shifting score, anti-saccade total, total dysexecutive errors), working memory (Dot Counting total, d-prime measures from the 1-back and 2-back), and fluency (total correct responses from each of the four verbal fluency tasks). The three-factor model provided excellent fit and fit clearly was superior to that of the one-factor model. This suggested that composite scores related to cognitive control, fluency, and working memory were appropriate. An additional analysis was performed to evaluate whether the one-factor model was sufficiently unidimensional to support a global executive function composite score based on all eleven variables. A bifactor model was fit in which all eleven variables loaded on a global factor, and in addition, each variable also loaded on a specific factor corresponding to the three-factor model. The global and specific factors were all constrained to be uncorrelated so ensure identifiability of the model. This model showed excellent fit. Loadings of all 11 variables on the global variable were strong and were as strong as loadings on the specific factors, which was interpreted as providing support for a global executive function composite score.

Item response theory (IRT) methods were used to generate scores for four variables: global executive function, cognitive control, fluency, and working memory. IRT has important invariance properties, and of particular relevance to EXAMINER, examinee scores generated by IRT analysis are invariant to specific items used. Consequently, an IRT score should provide an unbiased estimate of the examinee's ability even if different variables are used to generate that score. Each of the 11 continuous variables was recoded into an ordinal score with up to 20 response categories (see Tables 25–28). This process yielded ordinal scores that roughly matched the distributions of the raw continuous variables but had at least 10 observations in each response category. These ordinal scores were then entered into an IRT analyses corresponding to the four scores of interest. The R ltm module was used to fit a two parameter graded response model. Item parameters were calibrated and saved, and examinee scores and standard errors were calculated using Empirical Bayes scoring. This initially was performed with 18–64 year old English speakers, but subsequent analyses tested for differential item function related to age group (3–10, 11–17, 18–64, 65+) and language (English, Spanish). This was an iterative process in which the baseline model used the same parameters for all groups, and then in each subsequent step one additional variable was split so that parameters were freely estimated in the groups of interest. This process was empirically guided such that the variable selected for differential parameter estimation in any given step yielded a greater improvement in model fit than any of the other variables that had not been differentially estimated. This process was continued until further improvement in model fit was not obtained or there was only one variable left that shared the same parameters across groups. The score from this last iteration was considered a DIF free standard, and was compared with scores from previous iterations using the root mean square error. Results showed that accounting for DIF related to age group did not substantially change resulting IRT scores, but indicated that it was important to account for language for the global executive function score. Consequently, the same item parameters were used across all age groups, but parameters for the global executive function score differed across language groups for seven of the 11 variables.

The item parameters from the final IRT calibrations for each score were saved and a program was created in R to apply these parameters to generate scores for the global factor and the three specific factors. This program accepts raw continuous scores as input, and applies IRT Empirical Bayes scoring based on the saved parameters to generate scores.

Ordinal value	Dot Counting		1-back d-prime		2-back d-prime		Flanker score		Dysexecutive errors		Anti-saccades	
1	1	3.5	-1.110	-0.175	-1.941	-0.181	1.043	3.992	0	0.5	0	14.5
2	3.5	5.5	-0.175	0.067	-0.181	0.096	3.992	4.462	0.5	2.5	14.5	16.5
3	5.5	6.5	0.067	0.334	0.096	0.386	4.462	4.926	2.5	4.5	16.5	18.5
4	6.5	7.5	0.334	0.577	0.386	0.679	4.926	5.355	4.5	6.5	18.5	20.5
5	7.5	9.5	0.577	0.815	0.679	0.970	5.355	5.791	6.5	8.5	20.5	22.5
6	9.5	10.5	0.815	1.059	0.970	1.260	5.791	6.260	8.5	10.5	22.5	24.5
7	10.5	11.5	1.059	1.301	1.260	1.559	6.260	6.690	10.5	12.5	24.5	26.5
8	11.5	13.5	1.301	1.524	1.559	1.843	6.690	7.132	12.5	15.5	26.5	28.5
9	13.5	14.5	1.524	1.811	1.843	2.128	7.132	7.576	15.5	17.5	28.5	30.5
10	14.5	15.5	1.811	2.019	2.128	2.435	7.576	8.028	17.5	19.5	30.5	32.5
11	15.5	16.5	2.019	2.245	2.435	2.707	8.028	8.472	19.5	21.5	32.5	34.5
12	16.5	18.5	2.245	2.468	2.707	3.881	8.472	8.918	21.5	25.5	34.5	36.5
13	18.5	19.5	2.468	2.693			8.918	9.363	25.5	43	36.5	38.5
14	19.5	20.5	2.693	2.974			9.363	9.809			38.5	40
15	20.5	22.5	2.974	3.414			9.809	9.961				
16	22.5	23.5	3.414	3.671								
17	23.5	24.5										
18	24.5	27										
19												
20												

Table 26. Binning values (minimum and maximum) for English speakers										
Ordinal value	Verbal fluency 1		Verbal fluency 2		Category fluency 1		Category fluency 2		Shift Score	
1	0	3.5	0	1.5	0	4.5	0	1.5	1.429	2.879
2	3.5	5.5	1.5	3.5	4.5	6.5	1.5	2.5	2.879	3.618
3	5.5	7.5	3.5	4.5	6.5	8.5	2.5	4.5	3.618	4.069
4	7.5	9.5	4.5	6.5	8.5	10.5	4.5	5.5	4.069	4.477
5	9.5	10.5	6.5	7.5	10.5	12.5	5.5	7.5	4.477	4.894
6	10.5	12.5	7.5	9.5	12.5	14.5	7.5	8.5	4.894	5.304
7	12.5	14.5	9.5	10.5	14.5	16.5	8.5	10.5	5.304	5.699
8	14.5	16.5	10.5	12.5	16.5	18.5	10.5	11.5	5.699	6.109
9	16.5	18.5	12.5	13.5	18.5	20.5	11.5	13.5	6.109	6.510
10	18.5	19.5	13.5	15.5	20.5	22.5	13.5	14.5	6.510	6.921
11	19.5	21.5	15.5	16.5	22.5	24.5	14.5	15.5	6.921	7.329
12	21.5	36	16.5	18.5	24.5	26.5	15.5	17.5	7.329	7.733
13			18.5	19.5	26.5	28.5	17.5	18.5	7.733	8.142
14			19.5	22.5	28.5	40	18.5	21.5	8.142	8.548
15			22.5	30			21.5	29	8.548	8.956
16									8.956	9.358
17									9.358	9.861
18										
19										
20										

Table 27. Binning values (minimum and maximum) for Spanish speakers												
Ordinal value	Dot Counting		1-back d-prime		2-back d-prime		Flanker score		Dysexecutive errors		Anti-saccades	
1	5	8.5	-0.251	1.705	-1.130	0.496	3.932	7.710	0	4.5	0	38.5
2	8.5	9.5	1.705	1.910	0.496	0.686	7.710	8.020	4.5	5.5	38.5	40
3	9.5	10.5	1.910	2.126	0.686	2.485	8.020	8.319	5.5	9.5		
4	10.5	11.5	2.126	2.302			8.319	8.602	9.5	28		
5	11.5	12.5	2.302	2.522			8.602	9.869				
6	12.5	23	2.522	2.693								
7			2.693	3.671								
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												

Table 28. Binning values (minimum and maximum) for Spanish speakers										
Ordinal value	Verbal fluency 1		Verbal fluency 2		Category fluency 1		Category fluency 2		Shift Score	
1	1	4.5	0	4.5	6	13.5	4	8.5	2.100	5.721
2	4.5	7.5	4.5	6.5	13.5	16.5	8.5	9.5	5.721	6.110
3	7.5	8.5	6.5	7.5	16.5	17.5	9.5	10.5	6.110	6.526
4	8.5	9.5	7.5	8.5	17.5	18.5	10.5	11.5	6.526	6.867
5	9.5	10.5	8.5	9.5	18.5	20.5	11.5	12.5	6.867	7.238
6	10.5	12.5	9.5	10.5	20.5	26	12.5	13.5	7.238	9.733
7	12.5	25	10.5	21			13.5	14.5		
8							14.5	20		
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										

Table 29. Correlations between Factor Scores (adult sample in bold)				
	Executive Comp	Cognitive Cont	Working Mem	Fluency
Executive Comp	----	.79	.71	.87
Cognitive Cont	.87	----	.58	.48
Working Mem	.74	.60	----	.45
Fluency	.87	.60	.46	----

We highlight here two key points about the IRT measures. First, the Composite and Factor scores are presented in the original metric produced by the IRT algorithm used by the EXAMINER scoring program. We intentionally chose not to convert the scores to standard scores, scale scores, or anything else that could potentially be misconstrued as a norm-referenced measure. The scores do not contain any type of age adjustment, so it is likely that high functioning young children will have lower scores than impaired middle-age adults. Researchers can use these scores like any other raw score where balanced samples, within-subject designs, and careful attention to covariates are typical. The second key point concerns the standard error of these IRT scores. The fewer donor scales there are contributing to the Composite and Factor scores, the less reliable they will be, and the more cautious researchers must be in interpreting them. Each subject for whom an IRT score is generated will also have an associated standard error. We recommend not using scores that are associated with a standard error greater than .75.

Chapter 12. Administration and Scoring Guidelines

Global considerations

Before beginning a testing session, organize your materials so you can easily move from one task to another. When using alternate forms, ensure that you have the appropriate version.

The optimal seating arrangement is to be across the testing table from the examinee. Do your best to minimize factors that can affect the examinee's performance (e.g., extraneous noises, poor lighting, anxiety, low motivation). Take time to establish rapport with the participant at the beginning of the test session. Give the examinee a brief and very general overview of what the testing will entail. Family members should not be present in the room during an evaluation. Please administer tests in the order in which they are provided on the record form, unless doing so would compromise rapport with the participant.

Instructions should be read verbatim. Do not paraphrase unless necessary. Should the examinee have difficulties understanding the instructions, it is permissible to explain the task to them in another fashion (while maintaining the basic concept of the instructions); however, you should always read the verbatim instructions first. Always try to elicit the examinee's best performance. Encouraging remarks are important, but do not provide specific feedback regarding whether individual responses are correct or incorrect, except during practice trials. If the examinee seems discouraged or anxious about his or her performance, encouraging remarks like "I can tell you are trying your best," "No one gets all of these correct," or "That was a tough one for you, but you did it" are often helpful. If the examinee's motivation wanes, it is okay to explain in general terms what the task is trying to measure (e.g., "This is a measure of a particular kind of attention").

Provide breaks when needed. Examiners should pay attention to the examinee's motivation, emotional state, physical discomfort, and other factors that potentially compromise test validity. Many of the tasks have discontinuation rules. For those that do not, please attempt to complete all tasks whenever possible.

Examinees can use pen or pencil for carrying out the Unstructured Test and the Social Norms Questionnaire. However, if pencil is used, please provide a pencil without an eraser. Examinees should be instructed to cross out incorrect responses rather than attempt to erase them.

Fluency

MATERIALS & SET-UP: Timer- 1 minute, record form to record responses

The fluency task has two conditions, phonemic based word generation and category based word generation. The examiner should recite the instructions for each fluency task verbatim. Should the participant have difficulties understanding the instructions, it is permissible to explain the task to them further or answer any questions they may have. Start the stopwatch once it is clear that the examinee understands the instructions. Write the actual responses as legibly as possible. Record all responses, including repeated words and rule violations. When a rule violation occurs on three consecutive

responses, remind the participant of the correct rule. Each rule can be repeated only once per trial. Stop the procedure at 1 minute. If the examinee gives a response that is unclear during the task, make a note of it and query about it after the 1 minute has elapsed.

Fluency-Phonemic

The examiner asks the examinee to list as many words as they can that begin with a particular letter in one minute. The examinee is asked not to give names of people, names of places, or numbers as responses.

I'm going to say a letter of the alphabet. When I ask you to start, tell me as many words as you can that begin with that letter. You will have one minute before I tell you to stop. None of the words can be *numbers or names of people, or places.*

For example, if I gave you the letter *B*, you could say *brown, bottle or bake*, but you wouldn't say *Barbara, Boston or billion*. Also, don't give me the same word with different endings, so if you said *bake*, you wouldn't also say *baked or bakes*, and if you said *big*, you wouldn't also say *bigger and biggest*.

**Let's begin. Tell me all the words you can, as quickly as you can, that begin with the letter "F."
Ready? Begin.**

If the participant pauses for 15 seconds, prompt them to continue by saying, "**Keep going**" or "**What other words beginning with 'F' can you think of?**" If the participant gives three consecutive responses that do not start with the designated letter prompt them by saying, "**Remember, we are using the letter 'F'.**"

Fluency-Category

The examiner will ask the examinee to list as many words as they can that belong to a certain category in one minute.

Now I am going to give you a category, and I want you to name, as fast as you can, all of the things that belong to that category. For example, if I say "articles of clothing," you would say *shirt, tie or hat*. It doesn't matter what letter the word starts with.

Now I want you to name things that belong to the category: *Animals*. You will have one minute. I want you to tell me all the animals that you can think of in one minute. Ready? Begin.

If the participant pauses for 15 seconds say, **Keep going. What other animals can you think of?** If participant gives 3 consecutive words that do not fit the category say, **The category we are now using is *animals***. If the participant only names animals that begin with the letter L or F, remind the participant, **It doesn't matter what letter the words start with**. Any instruction can be repeated during a trial by participant request.

Phonemic Fluency Scoring Guidelines:

Record all responses, including repeated words and rule violations. When a rule violation (e.g., proper nouns, words beginning with the wrong letter) occurs on three consecutive responses, examiners should remind the participant of the correct rule. Each rule can be repeated only once per trial.

Correct Responses:

Any word that begins with the specified letter, can be found in a dictionary, is not a proper noun, number, or certain grammatical variant (see below), and is not a repetition within that trial, should be scored as a correct response.

Although scoring of most responses is straightforward, many responses are ambiguous. For example, “frank” can refer to a man’s name, a food item, or an adjective. The scoring principle with these sorts of responses is to give the benefit of the doubt and score the item as correct for the first instance of the response in a trial. In some instances, the context in which the response is given can provide clues as to the participant’s meaning. For example, the sound “fôr” is ambiguous, and could be a preposition (for), golf term (fore), or number (four). If the word is given along with other numbers (e.g., “four, five”), the response can be interpreted as a number and be scored as a rule violation. If the response is at all ambiguous, however, apply the general principle of giving the benefit of the doubt. If a person self-corrects a rule violation or repetition during the trial, the response should be scored as correct.

Other types of responses that should be scored as correct include:

- Contractions
- Compound words or conjoined words that convey a single meaning (e.g., ferris wheel)
- Slang words if they can be found in a dictionary
- Proper nouns that are not the names of people or places (e.g., days of the week, months of the year, brand names)

Repetitions:

Any response that is repeated verbatim within the 60-second trial should be scored as a repetition.

If a repeated word has more than one meaning (e.g., “still” can be an adjective and a noun) or is a homophone (e.g., “flue” and “flew”), score the second response as a repetition error unless the participant explicitly or implicitly (e.g., with intonation or gesture) indicates that the second response has a different meaning or spelling, or if the context strongly suggests that it is a different word. For example, for the string of responses “felt, feeling, fresh, fabric, felt,” the second occurrence of “felt” can be scored as correct since the context implies a different meaning than the first occurrence of “felt”.

Grammatical variants should be scored as rule violations, not repetitions (see below).

Rule Violations:

Any response that reflects a deviation from the rules provided to the participant should be scored as a rule violation.

Several types of responses are potentially rule violations and include:

- Words beginning with letters other than the designated letter. This includes words that have the same initial sound but begin with a different letter (e.g., “phone” for f-words).
- Non-words
- Proper nouns that are names of people or places
- Numbers
- Grammatical variants of a previous response. These include words that are exactly the same as a previous response but with a different ending that represents a plural, altered tense, or other grammatical variant (e.g., present participle; comparatives). It is important to note that the examples in the instructions only explicitly prohibit plurals (e.g., bake, bakes), alteration in tense (bake, baking), and comparatives (big, bigger), and thus only these types of variants should be scored as rule violations.
- Responses that are at all ambiguous should not be scored as rule violations. This particularly applies to responses that use the same root word as a previous response, but the addition is not a plural or change in tense. For example, give credit for “bakery,” even if “bake” was a previous response.

Repeated rule violations count as repetitions, not rule violations.

Category Fluency Scoring Guidelines:

The animal total score is the number of correct unique animals produced within the one-minute time limit. If a person self-corrects a rule violation or repetition during the trial, no error is counted.

Correct responses (animals):

Breeds (e.g., terriers)

Male and female

Infant names of a species (e.g., bull, cow, calf)

Both superordinate and subordinate examples of a species (e.g., both dog and terrier are credited)

Birds

Fish

Reptiles

Insects

Repetitions:

Any response that is repeated verbatim within the 60-second trial should be scored as a repetition.

When responses are plurals (e.g., mouse, mice) or different terms for the same animal (e.g., dog, canine), score the second response as a repetition.

Rule Violations:

Mythical animals (e.g., unicorn)

Wrong category

Correct responses (vegetables):

- Both superordinate and subordinate responses (e.g., peppers and jalapeños are credited)

- Less specific names (e.g., greens)
- Names of vegetables found in other cultures but perhaps unfamiliar to you (e.g., jicama) are acceptable only if they can be verified in the dictionary. After completion of the task, ask the participant to spell the word if you are unsure of the correct spelling.
- Grains (e.g., rice, wheat, oats, etc.), gourds, sugarcane, herbs, legumes, nuts, and seaweed
- Tomato, avocado, and pumpkin.

Responses not given credit (but not scored as errors)

Prepared vegetable products (e.g., pickles, tomato sauce, ketchup)

Spices (e.g., cinnamon, nutmeg, paprika, cumin, salt)

Repetitions:

Any response that is repeated verbatim within the 60-second trial should be scored as a repetition.

Rule Violations:

Wrong category

Unstructured Task

MATERIALS & SET-UP: Computer Timer - 6 minutes, Practice Page, 3 Stimulus Booklets

The examinee will be provided with three test booklets full of puzzles of varying point values and varying average times of completion. They will be asked to choose and complete the puzzles that will allow them to earn as many points as possible in six minutes.

Position the practice form in front of the examinee. Have participant complete the page.

On this practice page there are six puzzles for you to try. Each puzzle has a different instruction, and some puzzles are easier than others. Go ahead and complete this page so you can see the kinds of puzzles you will do.

Answer any questions the examinee may have regarding the puzzles on the practice form. Once the practice form has been completed, look over it to make sure all of the puzzles have been completed correctly. If an error has been made on the practice form, point this out to the examinee and explain why it is incorrect. Once the practice form has been completed and any errors have been explained move on to the task.

Position the three stimulus books in front of participant. Display the computer timer so that it is in clear view of the examinee. Set the timer for 6 minutes.

Here are three booklets. Each of the booklets has different puzzles you can do. In these booklets, there are four puzzles on each page. Each puzzle has a number of points that you will earn when you complete the puzzle. Some puzzles have higher points than others. Your goal is to earn as many points as possible.

You do not have to complete all of the pages in a book, and you do not have to complete all the puzzles on each page. You can go in any order you want through the puzzles. Each book is

worth the same amount of points. Be sure to read the instructions for each puzzle you do, and complete the puzzles accurately to receive full credit.

You will only have 6 minutes to earn as many points as possible, so choose your puzzles carefully. A timer will be displayed to help you manage your time.

Start computer timer after completing the instructions and it is clear that the examinee understands the instructions. Stop at 6 minutes. Do NOT allow participant to complete any item in progress when the time limit is reached. Collect the booklets from the examinee once the task is complete.

A timer for this task must be displayed on the computer for the participant. A timer program can be downloaded as freeware from the Internet using the following instructions:

1. Go to www.harmonyhollow.net/download.shtml.
2. In the "Product Download" table, find CoolTimer 2.2, freeware license, 657 KB file size. Click on "Harmony Hollow" link for downloading.
3. A prompt will be shown for file "ctimer.exe" Follow prompts to save this file onto the computer and run the installation program.
4. The program will download onto the desktop, and a clock icon will appear.

The default colors of the timer can be altered using the "Options" button. Please choose colors that make the numbers very visible. Do not make the window of the timer smaller. It is important that every participant is able to read the numbers on the computer screen.

The timer can be set to count down by clicking the up arrow next to the Minutes display until it reads 6 minutes. Press the large "Start" button to begin.

Scoring:

Performance on this task is based on the number of puzzles completed, the value of each puzzle, and the total number of points earned. Use the Unstructured Task Answer Key, located in Appendix A, for scoring. It has correct responses highlighted and high value puzzles marked.

A puzzle is counted as complete if at least 75% of the puzzle has been finished correctly. Items can also be counted as complete if there is a systematic error (e.g., circling elements instead of crossing out them out) that would not significantly shorten the time it takes to complete the puzzle. Use the Answer Key to determine completion.

Each puzzle has been designated as high value or low value. Value is determined by the amount of time a puzzle takes related to its point value. High value puzzles have been highlighted on the Answer Key.

Five scores are determined:

1. Count number of high value items completed. High value puzzles are marked on the Answer Key beneath each puzzle. The score includes all high value puzzles that have 75% finished correctly or have been completed with systematic errors that do not affect the puzzle significantly.
2. Count number of low value items completed. This includes puzzles that have 75% finished correctly, and puzzles completed with minor systematic errors.

3. Count number of high value items attempted in all three booklets. Attempted high value puzzles include every puzzle that has been started but is not complete. This includes puzzles with less than 75% completed correctly, and puzzles with consistent errors that would shorten time (e.g., crossing out every item instead of only given ones), and puzzles started but not completed when time is called.
4. Count number of low value puzzles attempted in all three booklets. Include all low value puzzles started that are incomplete or incorrect.
5. Add the points for all items completed. This value will be the sum of the point values for both low and high value puzzles completed.

Flanker

MATERIALS & SET-UP: Use Left and Right arrow keys. Participants should be 30–40 inches from the screen.

The examinee will be shown a row of five arrows on a computer screen and asked to indicate whether the center arrow is pointing to the right or to the left. The right and left arrow keys will be used to indicate direction. The examinee should respond to the stimuli as quickly as possible while trying not to make any mistakes.

You will be shown a row of five arrows on the screen, pointing to the left or right.

Point out the central arrow on the instruction screen so that it is clear to the examinee which arrow to focus on.

Press the LEFT button if the CENTRAL arrow is pointing to the left. Press the RIGHT button if the CENTRAL arrow is pointing to the right.

Point to left and right arrow keys as indicated.

Try to respond as quickly and accurately as you can. Try to keep your attention focused on the cross (“+”) at the center of the screen.

Check that participant’s fingers are appropriately placed on left and right arrow keys. Encourage participant to keep fingers in place until task is complete.

First we’ll do a practice trial. Press the SPACEBAR to begin.

Run practice trial. Provide feedback as needed. If the examinee finishes the practice trial and was not able to respond to 75% of the trials correctly, another practice trial will begin. If an examinee is not able to advance past the third practice trial, the task will discontinue. After the practice trial(s) is finished and it is clear that the examinee understands the directions, go on to the test.

Now we’ll move on to the task, the instructions are the same except you will no longer receive feedback after your responses. Press the LEFT button if the CENTRAL arrow is pointing to the

left. Press the RIGHT button if the CENTRAL arrow is pointing to the right. Remember to keep your focus on the center cross and try to respond as quickly as possible without making mistakes.

Press the SPACEBAR when you are ready to begin.

If the examinee stops responding or appears to be looking away from the screen, redirect the examinee to focus on the task.

Set Shifting

MATERIALS & SET-UP: Use Left and Right arrow keys. Participants should be 30–40 inches from the screen.

The examinee is asked to match objects by color or by shape depending on the verbal cue presented at the bottom of the screen. The examinee should respond to the stimuli as quickly as possible while trying not to make any mistakes.

This is a matching game. You will see a picture in the middle of the screen, and a word at the bottom of the screen. The word at the bottom of the screen will tell you how to match the picture in the middle to one of the objects in the corners.

You can match the picture by SHAPE or COLOR, and the word at the bottom of the screen will tell you which way to match. If you forget how you're supposed to match, look at the word to remind yourself.

When you have to match by COLOR, you should push the LEFT button for RED and the RIGHT button for BLUE. When you have to match by shape, you should push the LEFT button for TRIANGLE and the RIGHT button for RECTANGLE.

Check that the participant's fingers are appropriately placed on the left and right arrow keys. Encourage the examinee to keep their fingers in place until the task is complete.

Try to respond quickly and accurately, but if you make a mistake, just keep going. We'll try some practice trials first. Press the SPACEBAR to begin.

Run practice trial. Provide feedback as needed. If the examinee finishes the practice trial and was not able to respond to 75% of the trials correctly, another practice trial will begin. If an examinee is not able to advance past the third practice trial, the task will discontinue. After the practice trial(s) is finished and it is clear that the examinee understands the directions, go on to the test.

Now let's move on to the task, the instructions are the same but you will no longer receive feedback after your responses. When you have to match by COLOR, you should push the LEFT button for RED and the RIGHT button for BLUE.

When you have to match by shape, you should push the LEFT button for the TRIANGLE and the RIGHT button for the RECTANGLE.

Try to respond quickly as you can without making mistakes, but if you make a mistake just keep going. Press the SPACEBAR when you are ready to begin.

If the examinee stops responding completely or looks away from the screen, redirect their focus to the task.

Dot Counting

MATERIALS & SET-UP: Record form to record responses. Participants should be 30–40 inches from the screen.

The examinee will count and remember the total number of dots on consecutive display screens. After they have finished counting the dots on all of the screens presented, they will be asked to recall the total number of dots they counted on each screen in the order that the screens were presented. The six trials progress in difficulty with the number of display screen ranging in span from 2 to 7.

Position the computer so that both the examinee and the examiner have a clear view of the screen. Read the instructions aloud as they appear on each screen. Have score sheet ready to record responses.

You will be shown a series of images containing blue circles, green circles, and blue squares. You will count and remember the number of BLUE CIRCLES you see on each screen.

Hit the Spacebar to continue.

Count the BLUE circles aloud, one at a time, and then repeat the final total aloud IMMEDIATELY. This will indicate to the examiner that you have finished counting.

Hit the Spacebar to continue. A screen with blue and green circles and blue squares will appear. Instruct the examinee to count the number of blue circles that they see on the screen. After participant counts each blue circle aloud, hit the spacebar to continue to the next screen.

How many BLUE circles did you count?

After the examinee gives a response hit the spacebar to continue.

Now, you will count the BLUE circles on one screen, and then on another screen. Please begin counting the blue circles aloud as soon as they appear on the screen. When you finish counting the circles, repeat the total aloud. As soon as you repeat the final number you'll see a new screen.

Hit the Spacebar to continue. Screen with blue and green circles and blue squares will appear. After participant counts each blue circle aloud, hit the spacebar to continue to the next screen.

After a number of displays you will see question marks on the screen. This will be your cue to repeat the final numbers you counted. Let's do some practice first. Press the SPACEBAR to begin.

Run practice trials. Hit the SPACEBAR after the participant has finished counting the dots on a screen and repeated the final total. Instruct participant to begin counting immediately after a new screen appears.

You have completed the practice trials. Let’s continue with the task. The instructions are the same. Count and remember the number of BLUE CIRCLES you see on each screen. Count the blue circles aloud, one at a time, and then repeat the final total aloud. Repeat the final numbers you counted when you see the question marks appear on the screen.

Press the spacebar to begin the task. Record the totals that the examinee counts on the left side of the recording form. Record the totals that the examinee recalls on the right side of the recording form.

If the participant does not repeat the total aloud after counting, prompt only during the practice trials by saying, “Remember to repeat the total aloud when you are finished counting on the screen.”

Scoring:

A participant’s response is scored correct if they are able to recall the number of dots they have counted even if this is not the number of dots on the screen. No penalty is given for counting the dots on the screen incorrectly. Scoring is solely based on the participant’s ability to recall the numbers they have verbalized. Record verbatim the numbers recalled by the participant in the order in which they are given. Give one point for each correct number that is in the correct place.

See examples below:

Number of dots on screen	Number of dots counted by participant	Numbers recalled	Total score
3-5	3-5	3-5	2
4-6-2	4-6-2	4-2-6	1
8-2-1-7	8-2-1-7	8-2-1-6	3
2-5-4-8-7-5	2-5-4-7-6-5	2-5-4-7-6-5	6

Continuous Performance Test

MATERIALS & SET-UP: Use Left arrow key only. Participants should be 30–40 inches from the screen.

The examinee is asked to press the left arrow key every time the target image is presented. If anything other than the target image is presented they will not press any key. (Please note that the instructions below are for Form A).

You will be presented with different shapes on the screen. If a 5-pointed star is presented on the screen, press the LEFT ARROW key. If any shape other than a 5-pointed star is presented, do not press any button. Respond as quickly as you can without making mistakes. If you do make a mistake, just keep going.

Check that examinee's finger is appropriately placed on the left arrow key. Encourage participant to keep finger in place until task is complete.

Let's start with some practice trials. Press the SPACEBAR to begin.

Run practice trial. Provide feedback as needed. If the examinee finishes the practice trial and was not able to respond to 80% of the trials correctly, another practice trial will begin.

Let's try another practice trial.

If an examinee is not able to advance past the third practice trial, the task will discontinue. After the practice trial(s) is finished and it is clear that the examinee understands the directions, go on to the test.

Now let's move on to the actual test. Press the SPACEBAR when you are ready to begin.

If the examinee stops responding completely or looks away from the screen, redirect their focus to the task.

1-Back

MATERIALS & SET-UP: Use Left and Right arrow keys. Participants should be 30–40 inches from the screen.

The examinee is asked to remember the location of a square on the screen. They are then instructed to indicate whether the next square is in the same location as the previous one using the left arrow key for "yes" and the right arrow key for "no."

Read instructions aloud as they appear on each screen.

Remember the location of this square, so you can compare it to the location of the next square.

Press SPACEBAR to continue. A number will appear in the middle of the screen.

Say this number aloud.

The next square will appear on the screen.

Is this location the same as the one just before? If YES, press the LEFT key. If NO, press the RIGHT key. Now, remember the location of this square so you can compare it to the location of the next.

Let's try some more squares. Remember the location of each square so you can compare it to the next one. When you see a number in the middle of the screen, say the number aloud. Press the SPACEBAR to begin.

Read instructions for first three squares and continue reading if examinee shows difficulty understanding task. Repeat instructions and provide guidance as necessary to help examinee understand task.

Let's try a few more squares. This time you won't receive any directions or feedback. Compare the location of each square to the one just before. Please say the number aloud when you see it. Press the SPACEBAR to begin.

Verbalize instructions and provide guidance as necessary to help participant understand the instructions. After the examinee has successfully completed the practice trials they will go on to the actual task.

It is now time to begin the test. Please respond as quickly as possible without making mistakes. Press the SPACEBAR when you are ready to begin.

Provide guidance if needed on how to perform task, but without revealing correct responses. If participant indicates that he/she does not remember if a current square matches the one before, say, **"Take your best guess and try to get the next one correct."**

2-Back

MATERIALS & SET-UP: Use Left and Right arrow keys. The 2-back should always be administered immediately after the 1-back. Participants should be 30–40 inches from the screen.

Read instructions out loud as they appear on each screen.

Remember the location of this square, so you can compare it to the location of the square after the next one. Press SPACEBAR to continue. Also remember the location of this square so you can compare it to the location of the square after the next one. Press SPACEBAR to continue. Does this location match the one TWO before? If YES, press the LEFT key. If NO, press the RIGHT key.

Read Feedback. If Correct: **Correct! Now, remember the location of this square so you can compare it to the location of the one after next.** If Incorrect: **Try to get the next one right. Remember the location of this square, so you can compare it to the one after next.**

Repeat instructions and provide guidance as necessary to help participant understand task.

Good. Let's start over with some more squares. Press the SPACEBAR to begin.

Read instructions for the first 4 squares, as follows, and continue reading if participant demonstrates any difficulty understanding task. Provide additional guidance as necessary.

Remember this location. Screen advances. Also remember this location. Screen advances. Does this location match the location 2 before? Does this location match the location 2 before?

Let's try a few more squares. This time you won't receive any directions or feedback. Compare each square to the one 2 before. Start responding with the 3rd square. Please respond as quickly as possible without making mistakes. Press the SPACEBAR to begin.

Verbalize instructions and provide guidance as necessary to help participant understand task.

PRACTICE TRIAL 1:

Let's try a few more practice squares. Now, each square will appear for a shorter time. Remember to compare each square to the one 2 before. Start responding with the 3rd square. Please respond as quickly as possible without making mistakes. Press the SPACEBAR to begin.

Run practice trial. Verbalize instructions and provide guidance as necessary to help participant understand task. Do not reveal correct responses. If participant performs well, the test will begin. If they do not respond correctly, Practice Trial 2 will begin.

PRACTICE TRIAL 2:

Let's try a few more practice squares. Remember each square and compare it to the one 2 before. Please respond as quickly as possible without making mistakes. Press the SPACEBAR to begin.

Run practice trial. Verbalize instructions and provide guidance as necessary to help participant understand task, Do not reveal correct responses. If participant performs well, the test will begin. If they do not respond correctly, Practice Trial 3 will begin.

PRACTICE TRIAL 3:

Let's try a few more practice squares. Remember each square and compare it to the one 2 before. Please respond as quickly as possible without making mistakes. Press the SPACEBAR to begin.

Run practice trial. Verbalize instructions and provide guidance as necessary to help participant understand task, Do not reveal correct responses. If participant performs well, the test trials will begin. Otherwise, the test will END.

TEST:

It is now time to begin the test. Please respond as quickly as possible without making mistakes. Press the spacebar when you are ready to begin.

Provide guidance if participant expresses confusion about how to perform task, but without revealing correct responses. If participant indicates that he/she does not remember if a current square matches the one 2 before, say, **"Take your best guess and try to get the next one correct."**

After test is complete:

The test is complete. This was a challenging test and we want to make sure you understood the instructions. Please explain the instructions to the examiner.

On the record form, if the response displays correct understanding of the task, check "Yes." Otherwise, check "No."

Anti-Saccades

MATERIALS & SET-UP: Record form to record responses. Audio setting on the computer should be adjusted to an audible level. For both Pro-Trials and Anti-Trials, sit across from the participant so that you are facing him/her. Be sure you can see the examinee's eyes.

Set computer centered in front of participant. Computer screen should be approximately 31 inches (80 cm) away from participant and as close to eye level as possible.

For both Pro-Trials and Anti-Trials, sit across from the participant so that you are facing him/her. You will not be able to see the computer display. Be sure you can see the participant's eyes.

PRO-SACCADE TRIAL:

Today we are going to be doing a task to watch your eye movements. You will see a dot in the center of the screen. Point to the center of the computer screen. I would like you to move your eyes in the direction the dot moves, which will be either to the left or to the right of the screen. Then follow the dot back to the center. Do not move your head, just your eyes.

Any questions? Ok, ready?

Press the SPACEBAR when you are ready to begin the task. When the test begins, a recorded voice on the computer program will announce the number of each trial. Listen to the number announced to ensure you are recording each eye movement for the correct trial. Record initial direction of eye movement.

There are 10 trials for Pro-Saccades.

After the Pro-Saccades test, the instructions to the Anti-Saccades test will appear on the screen:

ANTI-SACCADE TRIAL:

Now we are going to be doing a second eye movement task. Again, you will see a dot in the center of the screen. Point to the center of the computer screen. The dot will move either to the left or to the right of the screen. This time, I would like you to use only your eyes to look at the opposite direction of where the dot moves. Do not move your head, just your eyes. After looking at the opposite side of where the dot is, return your eyes to look at the dot at the center of the screen. Here is an example. Press the SPACEBAR to display example of task.

Any questions? Ok, ready?

Press the SPACEBAR when you are ready to begin the task. A recorded voice will announce the number of each trial. Listen to the number announced to ensure you are recording each eye movement for the correct trial.

After the first set of 20 trials, press the spacebar to continue on to the next 20 trials.

Scoring:

Correct responses are scored using the recording boxes on the forms. Each response is marked based on the direction of initial eye movement.

Count the number of marked boxes that are not grayed out as the number of correct responses, then add the number of correct responses and record at the bottom of the Saccades score sheet.

Social Norms Questionnaire

MATERIALS & SET-UP: Social Interactions Questionnaire.

The examinee is asked to indicate whether or not a certain behavior is appropriate in mainstream culture of the United States.

Place the questionnaire in front of the participant and read the instruction.

The following is a list of behaviors that a person might do. Please decide whether or not it would be socially acceptable and appropriate to do these things in the mainstream culture of the United States, and answer *yes* or *no* to each. Think about these questions as they would apply to interactions with a stranger or acquaintance, NOT with a close friend or family member.

Collect the form when the examinee has finished. If the participant has difficulty answering an item or asks for clarification, provide guidance without revealing correct responses. Encourage the examinee to complete all of the items.

Behavioral Rating Scale

The Behavioral Rating Scale is completed by the examiner after completion of the testing. Examiners restrict their ratings to behaviors that they have observed directly, but include all observed behaviors, regardless of the context. Thus, although behaviors during the actual assessment will likely provide the bulk of data, examiners should also note behaviors exhibited in all other situations, such as the waiting room and walking to and from the exam room. There are nine behavioral domains rates, including agitation, stimulus-boundedness, perseverations, decreased initiation, motor stereotypies, distractibility, degree of social/emotional engagement, impulsivity, and social appropriateness.

Following the administration of the battery to the study participant, the examiner should indicate the presence of the following behavioral features observed during the examiners' time with the study participant. None indicates the absence of the feature. When a feature is present, it should be rated as mild, moderate or severe depending on the extent to which it disrupts the testing or interpersonal exchanges, or the extent to which it deviates from accepted norms.

ITEM DESCRIPTORS AND BEHAVIOR CHECKLIST:

1. Agitated	None	Mild	Moderate	Severe
2. Stimulus-bound	None	Mild	Moderate	Severe
3. Perseverative	None	Mild	Moderate	Severe
4. Decreased initiation	None	Mild	Moderate	Severe
5. Motor stereotypies	None	Mild	Moderate	Severe
6. Distractible	None	Mild	Moderate	Severe
7. Lacks social/emotional engagement	None	Mild	Moderate	Severe
8. Impulsive	None	Mild	Moderate	Severe
9. Socially inappropriate	None	Mild	Moderate	Severe

GENERAL GUIDELINES:

Examiners should restrict their ratings to behaviors that they have observed, and not use this rating scale to reflect behaviors that are described by caregivers, informants, or other health care professionals but not directly observed by the examiner. Examiners should include all observed behaviors, regardless of the context. Thus, although behaviors during the actual assessment will likely provide the bulk of data, examiners should also note behaviors exhibited in all other situations, such as the waiting room and walking to and from the exam room.

There will be some instances when raters will have to decide which of several potential categories to rate a particular behavior. For example, repeatedly pickup up a pencil from the table and scribbling on test forms could be potentially viewed as perseverative, stimulus-bound, or motor stereotypy. It is important for raters to select **only one** category for an observed behavior. Typically, raters will have to use their best clinical judgment as to the most appropriate category. Some suggestions for determining which category to select are mentioned below. Use the space in the bottom half of the rating page to describe the behavior. This is particular important step for subjects being evaluated longitudinally in the event that a different rater will be seeing the subject at the next visit.

DETERMINING SEVERITY:

As a general rule, the severity of the observed behavior should reflect the extent to which it disrupts the testing or interpersonal exchanges, or the extent to which it deviates from generally accepted norms. Mild refers to an infrequent occurrence of the behavior or if the behaviors observed are present but relatively insignificant. Subjects are easily redirected and there is no or minimal impact on the quality of the testing. A rating of moderate would indicate that the occurrence of behaviors begin to infringe on the quality of the data and neuropsychological test performance. Subjects are less easily redirected. For example, if a subject cannot take social cues from the examiner and continues discussing topics that are inappropriate for the clinical setting, "Socially Inappropriate Behavior" would be rated as "moderate." A rating of "severe" indicates that the behavior occurs very

frequently throughout the testing situation. For example, Distractibility would be rated as “severe” when the subject is very difficult to redirect to the task at hand, to the point that test validity is questionable. Additional examples are provided below.

SPECIFIC BEHAVIORS:

1. Agitation: Agitation can involve inappropriate verbal (screaming, cursing) or physical (repetitive body movements, hitting or throwing objects) behaviors and can be aggressive or non-aggressive in nature (pacing versus kicking). Mild agitation might manifest itself as being anxious to complete the evaluation, argumentative, or complaining about testing. Moderate levels might include disruptive but not harmful behaviors that cannot be easily redirected. Severe agitation would include physical behaviors that put others or self in danger (e.g., hitting, pushing, scratching, throwing things, biting or kicking) or extreme verbal aggression (e.g., screaming or cursing loudly).

2. Stimulus-boundedness: Stimulus-boundedness is an inappropriate response to a salient environmental stimulus. Such behaviors could include unsolicited reading of nearby text, environmental dependency (e.g., picking up a pen from the table and writing; eating food from someone else’s plate), excessive attention to irrelevant objects (e.g., picking up objects from floor), echolalia, or, during cognitive testing, writing or drawing on a model or adjacent stimulus. The occurrence of even one instance of stimulus-boundedness warrants a mild rating. Examples of “moderate” include closing in on a stimulus or being distracted by adjacent stimuli and these behaviors result in clearly impaired performance. Environmentally dependent behaviors occur often and subject cannot be easily redirected. Subjects whose environmental dependency makes task completion very difficult would be rated as severe.

3. Perseverative: Perseveration is the inappropriate and unintentional persistence of a behavior, and can be observed behaviorally, on testing, and in conversational speech. As observed on testing, the persisting behavior may be (1) a repetition of a previously generated response within a task, or (2) a repetition of a response appropriate to an earlier task or condition. As observed in conversation speech, a participant appears stuck on an idea or persistently returns to a previously voiced idea or story. (A few repeated responses during testing are not uncommon in normal individuals, so this behavior should only be scored as present when the frequency deviates from normative expectations). Repetitions that slightly exceed normative expectations would be rated as mild. A moderate rating would be given any time the subject makes a perseverate response without any intervening responses (e.g., drawing the exact same design two or more times in a row on Design Fluency) or reverts to previously established response set (e.g., on Category Fluency, gives responses only beginning with the letter L as required by the previous task, or makes numerous perseverations of any type. A pervasive tendency to perseverate that interrupts the testing would be rated as severe.

4. Decreased initiation: Behaviors suggesting decreased initiation include delayed time to start verbal and/or motor response after being given task instructions, need for additional prompting to initiate a response, and low motivation to perform well. This would also include instances when a subject stops in the middle of a task and requires additional prompts to continue. Note: If decreased initiation occurs only in the context of lack of social engagement or distractibility, code as those domains rather than decreased initiation.

5. Motor stereotypies: These are persistent, repetitive behaviors without a clear purpose. Examples include pacing without apparent purpose, rummaging through pockets or drawers, picking

at skin or clothes, handling buttons, repeatedly putting on and removing jewelry or clothing, repeated lip smacking or other sounds, or repeating phrases that have no communicative value. Note: If a behavior is coded as a motor stereotypy, that same behavior should not also be coded as a perseveration.

6. Distractibility: Attention is easily diverted away by external stimuli (people, objects, etc.), outside noise, or internal thought. This is observed when the subject loses his/her train of thought, needs multiple reminders about instructions, needs to have their attention redirected to the task at hand, or engages in tangential speech or thought. This can be rated as mild if the distractibility is occasional and does not significantly interfere with collecting valid test data, moderate if the subject needs several reminders and redirection, but still able to complete evaluation and severe if the distractibility significantly disrupts testing to the point that validity is questionable.

7. Lack of Social / Emotional Engagement: This refers to the examiner's impression that the subject displays signs of diminished social interest, interrelatedness, or personal warmth. Examples of behaviors might include lack of eye contact, excessive eye contact (e.g., staring), lack of smiling or reduced range of facial affect (unrelated to Parkinson's disease), lack of awareness of how subject's behavior might affect others, lack of interest in others. Additional behaviors include lack of spontaneity and not initiating conversations. The main distinction between lack of social engagement and social inappropriateness is that lack of social engagement typically reflects an absence of behavior (e.g., empathy, warmth) whereas social inappropriateness typically reflects the distinct presence of an inappropriate behavior (e.g., inappropriate touching or remarks).

8. Impulsivity: This refers to acting with insufficient forethought or patience. Potential behaviors include beginning a task before examiner completes instructions, responding to only part of the instructions (e.g., the first part of multi-step instructions), a careless approach to work, finishing too quickly, and interrupting the examiner. A mild rating would refer to when the examinee works quickly without checking work, appears bored when solving complex problems, or starts one test item before the examiner completes the instructions or without adequate planning time. For a moderate rating, the examinee might respond carelessly at the expense of accuracy, start test items before the examiner completes the instructions or without adequate planning time on more than one occasion, or otherwise tend to do or say things rapidly and without planning. An example of severe impulsivity might be when a participant's tendency to act without forethought disrupts the testing or interpersonal exchanges, or markedly deviates from normal behavior.

9. Socially inappropriate: Refers to conduct which may not be suitable in professional settings. Examples may include talking to strangers on the elevator, questioning the examiner/clinician about personal information or their credentials, disregard for or violation of personal space (e.g., excessive or inappropriate touching of examiner), or lack of response to social cues (e.g., when others attempt to end a conversation). Out-of-place topics, crude or sexually oriented comments, jokes or opinions that may be offensive to others, or poor hygiene/grooming (malodorous, stained, torn, or inappropriate clothing) can also fall under this category. Physical behaviors such as flatulence, touching private body parts, belching, and spitting can also be considered.

Chapter 13. Procedures for Calculating Composite and Factor Scores

Installing the Scoring Program

The EXAMINER battery includes software for generating the executive composite and factor scores. The software for calculating the executive composite and factor scores is programmed in the R language and relies on ltm module (latent trait models under the Item Response Theory approach). The R software only needs to be installed on computers that will be used to calculate the executive composite and factor scores (e.g. a computer used for data management and analysis).

1. Install the R software (following instructions in Chapter 7 of this manual).
2. Locate and copy the **ExaminerScoring** installation file for your operating system to the computer where the scoring program will be installed:
 - a. For Windows the installation file is **ExaminerScoring.zip** and is located in the **/Battery and Stimulus/Software/Windows/Scoring** folder of the EXAMINER battery distribution.
 - b. For OS X the installation file is **ExaminerScoring.dmg** and is located in the **/Battery and Stimulus/Software/OSX/Scoring** folder of the EXAMINER battery distribution.
 - c. For Linux the installation file is **ExaminerScoring.tar.gz** and is located in the **/Battery and Stimulus/Software/Linux/Scoring** folder of the EXAMINER battery distribution.
3. Extract the contents of the **ExaminerScoring** installation file to a location of your choice on the computer where the scoring program will be installed.

Running the Scoring Program

The scoring program will read in a data file in CSV (comma-separated values) format, will validate that all required data elements are present, will score the records, and generate an output file with the data from the input file and the composite and factor scores. This section describes how to run the scoring program. See the following section for information in preparing an input file for the scoring program.

1. Start the R program.
The R Console will appear.
2. *Move the Examiner Software folder to your desktop. This folder contains a Scoring and Tasks folder.*
3. Set the working directory in the R Console to the folder that contains the EXAMINER “scoring folder.” Every time you run R you will need to change the working directory to this directory. This directory is the one that contains the “test” and “lookup” folder.
 - a. On Windows, select **Change Dir** from the **File** menu.
 - b. On OSX, select **Change Working Directory** from the **Misc** menu.
4. Type `source("examiner_scoring.R")` in the R Console and hit the RETURN key.
5. Type `score_file("./test/ExampleScoringInput.csv", "./test/TestOutput1.csv")` in the R Console and hit the RETURN key.
When using the scoring program replace the file names in the score_file() function call with the name of the prepared input file and the desired name of the output file.
6. The R program will run the scoring routine and write the output file with the calculated composite and factor scores.

Preparing an Input File for the Scoring Program

The scoring program requires a file in CSV (comma-separated-values) format with 11 “donor” variables and a variable that indicates the language of the battery administration. There is an example input file in the **test** subdirectory of the scoring program located in **EXAMINER**

3.6\Software\Windows\Scoring\scoring.zip\scoring\test called **ExampleScoringInput.csv**.

Data procedures should be established for projects or trials using the EXAMINER battery, which will generate the scoring input file from the data files generated by the EXAMINER computer tasks and the Battery Record Forms used to record the paper and pencil tasks.

The input file for the scoring program requires the variables in the table below. Additional optional variables (e.g., subject_id, session_num) can be included in the input file to help identify the rows of data. **Any ‘-5’ values (which are defined as “could not calculate” in the Summary Data Output Files) or any missing data that have assigned values (e.g., -999, 7777, NA, etc.) should be removed from the input file and replaced with a blank cell before scoring.**

Variable	Source	Definitions/Notes
language	Battery Record Form	The language of battery administration (enter ‘1’ for English and enter ‘2’ for Spanish or any other language).
dot_total	Battery Record Form	Dot Counting Total Correct (sum of all 6 trials, out of 27).
nb1_score	NBack Summary Data Output File	1-back score (d-prime).
nb2_score	NBack Summary Data Output File	2-back score (d-prime).
flanker_score	Flanker Summary Data Output File	Flanker Score.
error_score	Calculated	See calculation instructions below.
antisacc	Battery Form	Antisaccade Total Correct (Trial 1 + Trial 2).
shift_score	Set Shifting Summary Data Output File	Shift Score.
vf1_corr	Battery Record Form	Verbal Fluency: Phonemic Trial 1 Total Correct.
vf2_corr	Battery Record Form	Verbal Fluency: Phonemic Trial 2 Total Correct.
cf1_corr	Battery Record Form	Category Fluency: Semantic Trial 1 Total Correct.
cf2_corr	Battery Record Form	Category Fluency: Semantic Trial 2 Total Correct.

Calculating the error_score Variable

If the administration of the battery is for an adult or child and the CPT, Flanker, Set Shifting, Verbal Fluency 1+2, Category Fluency 1+2 and the Behavior Rating Scale have all been completed, the error_score can be calculated by summing the 12 variables in the table below. If the administration of the battery is for a child less than 7 years old, and the verbal fluency task was not administered, the error_score can be calculated simply by omitting the verbal fluency variables. Otherwise, if any of the tasks contributing to the error_score were not completed during the administration of the battery, do not provide an error_score to the scoring calculation. **Any negative values for flanker_error_diff and/or shift_error_diff should be converted to ‘0’s before calculating error_score.**

Variable	Source	Definitions/Notes
nontarget_errors	CPT Summary Data Output File	The number of non-target trials where the subject provided a response (False Alarms).
flanker_error_diff	Flanker Summary Data Output File	The difference between the number of errors on incongruent trials and congruent trials. <i>Any negative values for this variable should be converted to '0' before calculating error_score.</i>
shift_error_diff	Set Shifting Summary Data Output File	The difference between the number of errors on the shift block and the combined errors on the color and shape blocks. <i>Any negative values for this variable should be converted to '0' before calculating error_score.</i>
vf1_rep	Battery Record Form	Verbal Fluency: Phonemic Trial 1 Repetitions.
vf1_rv	Battery Record Form	Verbal Fluency: Phonemic Trial 1 Rule Violations.
vf2_rep	Battery Record Form	Verbal Fluency: Phonemic Trial 2 Repetitions.
vf2_rv	Battery Record Form	Verbal Fluency: Phonemic Trial 2 Rule Violations.
cf1_rep	Battery Record Form	Category Fluency: Semantic Trial 1 Repetitions.
cf1_rv	Battery Record Form	Category Fluency: Semantic Trial 1 Rule Violations.
cf2_rep	Battery Record Form	Category Fluency: Semantic Trial 2 Repetitions.
cf2_rv	Battery Record Form	Category Fluency: Semantic Trial 2 Rule Violations.
behav_total	Battery Record Form	Behavior Rating Scale Total Score.

Composite and Factor Score Output Variables

The output variables generated by the scoring program are listed in the table below.

Variable	Description
executive_composite	The executive composite score.
executive_se	The executive composite score standard error.
fluency_factor	The fluency factor score.
fluency_se	The fluency factor score standard error.
cog_control_factor	The cognitive control factor score.
cog_control_se	The cognitive control factor score standard error.
working_memory_factor	The working memory factor score.
working_memory_se	The working memory factor score standard error.

Chapter 14. Known Issues

- If you are administering the tasks on a Linux machine, the centering and alignment of the textual instructions may be off.

References

- Amieva, H., Phillips, L., & Della Sala, S. (2003). Behavioral dysexecutive symptoms in normal aging. *Brain Cogn*, 53(2), 129-132.
- Anderson, S. W., Damasio, H., Jones, R. D., & Tranel, D. (1991). Winconsin Card Sorting Test performance as a measure of frontal lobe damage. *J. Clin. Exp. Neuropsychol.*, 13(6), 909-922.
- Baddeley, A. (2002). Fractionating the central executive. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 246-260). New York: Oxford University Press.
- Baddeley, A., & Della Sala, S. (1996). Working memory and executive control. *Philos Trans R Soc Lond B Biol Sci*, 351(1346), 1397-1403; discussion 1403-1394.
- Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation Of working memory from decision making within the human prefrontal cortex. *J Neurosci*, 18(1), 428-437.
- Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., . . . Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Arch Clin Neuropsychol*, 20(4), 517-529.
- Berman, K. F., Ostrem, J. L., Randolph, C., Gold, J., Goldberg, T. E., Coppola, R., . . . Weinberger, D. R. (1995). Physiological activation of a cortical network during performance of the Wisconsin Card Sorting Test: a positron emission tomography study. *Neuropsychologia*, 33(8), 1027-1046.
- Bowden, S. C., Benedikt, R., & Ritter, A. J. (1992). Delayed matching to sample and concurrent learning in nonamnesic humans with alcohol dependence. *Neuropsychologia*, 30(5), 427-435.
- Chan, R. C., Hoosain, R., & Lee, T. M. (2002). Reliability and validity of the Cantonese version of the Test of Everyday Attention among normal Hong Kong Chinese: a preliminary report. *Clin Rehabil*, 16(8), 900-909.
- Chan, R. C., & Manly, T. (2002). The application of "dysexecutive syndrome" measures across cultures: performance and checklist assessment in neurologically healthy and traumatically brain-injured Hong Kong Chinese volunteers. *J Int Neuropsychol Soc*, 8(6), 771-780.
- Chan, R. C., Robertson, I. H., & Crawford, J. R. (2003). An application of individual subtest scores calculation in the Cantonese version of the Test of Everyday Attention. *Psychol Rep*, 93(3 Pt 2), 1275-1282.
- Cummings, J. L., Mega, M., Gray, K., Thompson-Rosenberg, J., Carusi, D. A., & Gornbein, J. (1994). The Neuropsychiatric Inventory: Comprehensive assessment of psychopathology in dementia. *Neurology*, 44, 2308-2314.
- D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, 378(6554), 279-281.
- D'Esposito, M., Postle, B. R., Ballard, D., & Lease, J. (1999). Maintenance versus manipulation of information held in working memory: an event-related fMRI study. *Brain Cogn*, 41(1), 66-86.
- Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System*. San Antonio: The Psychological Corporation.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Holdnack, J. (2004). Reliability and validity of the Delis-Kaplan Executive Function System: an update. *J Int Neuropsychol Soc*, 10(2), 301-303.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test, 2nd Edition*. San Antonio, TX: The Psychological Corporation.
- Dubois, B., Slachevsky, A., Litvan, I., & Pillon, B. (2000). The FAB: a Frontal Assessment Battery at bedside. *Neurology*, 55(11), 1621-1626.
- Dunbar, K., & Sussman, D. (1995). Toward a cognitive account of frontal lobe function: simulating frontal lobe deficits in normal subjects. *Ann N Y Acad Sci*, 769, 289-304.

- Espy, K. A., & Cwik, M. F. (2004). The development of a trial making test in young children: the TRAILS-P. *Clin Neuropsychol*, *18*(3), 411-422.
- Espy, K. A., Kaufmann, P. M., & Glisky, M. L. (2001). New procedures to assess executive functions in preschool children. *Clin Neuropsychol*, *15*(1), 46-58.
- Espy, K. A., Kaufmann, P. M., McDiarmid, M. D., & Glisky, M. L. (1999). Executive functioning in preschool children: performance on A-not-B and other delayed response format tasks. *Brain Cogn*, *41*(2), 178-199.
- Espy, K. A., Stalets, M. M., McDiarmid, M. M., Senn, T. E., Cwik, M. F., & Hamby, A. (2002). Executive functions in preschool children born preterm: application of cognitive neuroscience paradigms. *Neuropsychol Dev Cogn C Child Neuropsychol*, *8*(2), 83-92.
- Gioia, G. A., Isquith, P. K., Retzlaff, P. D., & Espy, K. A. (2002). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. *Neuropsychol Dev Cogn C Child Neuropsychol*, *8*(4), 249-257.
- Heflin, L.H., Laluz, V. Jang, J., Ketelle, R., Miller, B.L., & Kramer, J.H. (2011). Let's inhibit our excitement: The relationship between Stroop, behavioral inhibition, and the frontal lobes. *Neuropsychology*, *25*(5):655-65.
- Korkman, M., Kemp, S. L., & Kirk, U. (2001). Effects of age on neurocognitive measures of children ages 5 to 12: a cross-sectional study on 800 children from the United States. *Dev Neuropsychol*, *20*(1), 331-354.
- Krueger, C. E., Rosen, H. J., Taylor, H. G., Espy, K. A., Schatz, J., Rey-Casserly, C., & Kramer, J. H. (2011). Know thyself: real-world behavioral correlates of self-appraisal accuracy. *Clin Neuropsychol*, *25*(5), 741-756.
- Levine, B., Robertson, I. H., Clare, L., Carter, G., Hong, J., Wilson, B. A., . . . Stuss, D. T. (2000). Rehabilitation of executive functioning: an experimental-clinical validation of goal management training. *J Int Neuropsychol Soc*, *6*(3), 299-312.
- Litvan, I., Agid, Y., Calne, D., Campbell, G., Dubois, B., Duvoisin, R. C., . . . Zee, D. S. (1996). Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): report of the NINDS-SPSP international workshop. *Neurology*, *47*(1), 1-9.
- Malloy, P., & Grace, J. (2005). A review of rating scales for measuring behavior change due to frontal systems damage. *Cogn Behav Neurol*, *18*(1), 18-27.
- Manchester, D., Priestley, N., & Jackson, H. (2004). The assessment of executive functions: coming out of the office. *Brain Inj*, *18*(11), 1067-1081.
- Marshall, S. C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. *Psychol Aging*, *12*(4), 718-725.
- McDonald, R. J., Ko, C. H., & Hong, N. S. (2002). Attenuation of context-specific inhibition on reversal learning of a stimulus-response task in rats with neurotoxic hippocampal damage. *Behav Brain Res*, *136*(1), 113-126.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognit Psychol*, *41*(1), 49-100.
- Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & Gonzalez, H. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): further development and psychometric characteristics. *Psychol Assess*, *16*(4), 347-359.
- Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S., . . . Benson, D. F. (1998). Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. *Neurology*, *51*(6), 1546-1554.

- Petersen, R. C., Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., . . . Winblad, B. (2001). Current concepts in mild cognitive impairment. *Arch Neurol*, *58*(12), 1985-1992.
- Polman, C. H., Reingold, S. C., Edan, G., Filippi, M., Hartung, H. P., Kappos, L., . . . Wolinsky, J. S. (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Ann Neurol*, *58*(6), 840-846. doi: 10.1002/ana.20703
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA Division 40 members. *Arch Clin Neuropsychol*, *20*(1), 33-65.
- Robbins, T. W., James, M., Owen, A. M., Sahakian, B. J., McInnes, L., & Rabbitt, P. (1994). Cambridge Neuropsychological Test Automated Battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers. *Dementia*, *5*(5), 266-281.
- Rodriguez del Alamo, A., Catalan Alonso, M. J., & Carrasco Marin, L. (2003). [FAB: a preliminary Spanish application of the frontal assessment battery to 11 groups of patients]. *Rev Neurol*, *36*(7), 605-608.
- Rogers, R. D., Andrews, T. C., Grasby, P. M., Brooks, D. J., & Robbins, T. W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *J Cogn Neurosci*, *12*(1), 142-162.
- Rosso, I. M., Young, A. D., Femia, L. A., & Yurgelun-Todd, D. A. (2004). Cognitive and emotional components of frontal lobe functioning in childhood and adolescence. *Ann N Y Acad Sci*, *1021*, 355-362.
- Royall, D. R., Mahurin, R. K., & Gray, K. F. (1992). Bedside assessment of executive cognitive impairment: the executive interview. *J Am Geriatr Soc*, *40*(12), 1221-1226.
- Saver, J. L., & Damasio, A. R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, *29*(12), 1241-1249.
- Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in men. *Brain*, *114*, 727-741.
- Shallice, T., & Burgess, P. W. (1996). The domain of supervisory processes and temporal organisation of behaviour. *Phil Trans R Soc Lond B*, *351*, 1405-1412.
- Stuss, D. T., Shallice, T., Alexander, M. P., & Picton, T. W. (1995). A multidisciplinary approach to anterior attentional functions. *Ann N Y Acad Sci*, *769*, 191-211.
- Tranel, D., Bechara, A., & Denburg, N. L. (2002). Asymmetric functional roles of right and left ventromedial prefrontal cortices in social conduct, decision-making, and emotional processing. *Cortex*, *38*(4), 589-612.
- Wilson, B. A., Alderman, N., Burgess, P.W., Emslie, H. & Evans, J. (1996). *Behavioral Assessment of the Dysexecutive Syndrome*: Thames Valley Test Company.