Journal of the International Neuropsychological Society (2014), **20**, 11–19.
Copyright © INS. Published by Cambridge University Press, 2013.
doi:10.1017/S1355617713001094

SPECIAL SERIES

NIH EXAMINER: Conceptualization and Development of an Executive Function Battery

Joel H. Kramer,[1] Dan Mungas,[2] Katherine L. Possin,[1] Katherine P. Rankin,[1] Adam L. Boxer,[1] Howard J. Rosen,[1] Alan Bostrom,[1] Lena Sinha,[1] Ashley Berhel,[1] AND Mary Widmeyer[3]

[1]Department of Neurology, University of California, San Francisco, California
[2]Department of Neurology, University of California, Davis, California
[3]Rosalind Franklin University of Medicine and Science, Chicago, Illinois

(RECEIVED March 18, 2013; FINAL REVISION August 23, 2013; ACCEPTED August 30, 2013; FIRST PUBLISHED ONLINE October 8, 2013)

Abstract

Executive functioning is widely targeted when human cognition is assessed, but there is little consensus on how it should be operationalized and measured. Recognizing the difficulties associated with establishing standard operational definitions of executive functioning, the National Institute of Neurological Disorders and Stroke entered into a contract with the University of California-San Francisco to develop psychometrically robust executive measurement tools that would be accepted by the neurology clinical trials and clinical research communities. This effort, entitled Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (EXAMINER), resulted in a series of tasks targeting working memory, inhibition, set shifting, fluency, insight, planning, social cognition and behavior. We describe battery conceptualization and development, data collection, scale construction based on item response theory, and lay the foundation for studying the battery's utility and validity for specific assessment and research goals. (*JINS*, 2014, *20*, 11–19)

Keywords: working memory, cognitive control, fluency, planning, social cognition, item response theory

INTRODUCTION

Executive deficits are reported in numerous neurobehavioral conditions, and may be the primary locus of cognitive impairment in attention-deficit/hyperactivity disorder (Barkley, 2010), behavioral variant frontotemporal dementia (Boone et al., 1999; Hutchinson & Mathias, 2007; Slachevsky et al., 2004), subcortical ischemic vascular disease (Moorhouse et al., 2010; Reed et al., 2004), traumatic brain injury (Caeyenberghs et al., 2012; Levin & Hanten, 2005; Stuss, 2011), multiple sclerosis (Arnett et al., 1997; Chiaravalloti & DeLuca, 2003; Foong et al., 1997), Huntington's disease (Aron et al., 2003; Paulsen, 2011), progressive supranuclear palsy (Gerstenecker, Mast, Duff, Ferman, & Litvan, 2013), Parkinson's disease (Ravizza & Ciranni, 2002), and even normal aging (Amieva, Phillips, & Della Sala, 2003; Buckner, 2004).

Neuroscientists and cognitive psychologists have begun to parse executive functioning into subcomponents and identify relevant anatomical regions and networks. Clinical assessment of executive control, however, has fallen behind these basic science advances. This gap is particularly evident in clinical trials, where despite the importance of executive abilities for daily living (Asimakopulos et al., 2012; Cahn-Weiner, Boyle, & Malloy, 2002), measures of executive ability are often omitted or underrepresented in clinical trial batteries. When executive functioning is targeted in research, there is considerable variability in how it is operationally defined. Tasks purportedly measuring fluency, working memory, concept formation, set shifting, inhibition, organization, abstract reasoning, and novel problem solving, either individually or in various combinations, are all used as markers of executive functioning, with the implicit assumption that these tasks measure the same construct.

Recognizing the challenges associated with conceptualizing and measuring executive functioning, the National Institute of Neurological Disorders and Stroke (NINDS) awarded a contract to the University of California-San Francisco (UCSF) to develop psychometrically robust executive measurement tools that would be accepted by the neurology clinical trials and clinical research communities. Initial goals

Correspondence and reprint requests to: Joel H. Kramer, 675 Nelson Rising Lane, Suite 190, MC 1207, San Francisco, CA 94158. E-mail: jkramer@memory.ucsf.edu

for the battery were: (1) multiple domains of executive functioning; (2) modularity (e.g., flexibility in which tasks are administered); (3) portability; (4) replicability across laboratories; (5) suitable across a broad range of ages and neurobehavioral conditions; (6) adaptable for clinical trials; (7) available in the public domain; and (8) English and Spanish versions. An External Advisory Board further recommended: (1) administration time of less than 45-min; (2) multiple alternate forms; (3) usage of computer-administered tasks; and (4) external measures of real-world functioning to validate the battery.

Battery Development

The UCSF project, entitled Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (EXAMINER) proceeded in two general phases, battery development and data collection. During the development phase, the UCSF team was built, a website (examiner.ucsf.edu) was created to facilitate communication with National Institutes of Health (NIH) and the public, and the literature on executive constructs and instruments was extensively reviewed, including test batteries, attention, set shifting, inhibition, social functioning, and self-monitoring; the complete review was posted on the website. A team of external advisers was convened that included neurology, developmental psychology, neuropsychology, cross-cultural psychology, clinical trials, and experts on executive functioning. Experts in the field were surveyed using SurveyMonkey® to elicit information on what they believed were priorities for battery development. These steps led to defining the conceptual framework for the NIH-EXAMINER battery, selecting existing executive paradigms from the research and clinical literature, developing novel tasks, and carrying out extensive piloting. Record forms, test stimuli, software for computerized tasks, and training materials were created. Translation of test materials into Spanish was carried out by a professional translation service with back translation. The Frontal Systems Behavior Scale® (FrsBe; Malloy & Grace, 2005) and the Behavior Rating Inventory of Executive Function® (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000) were added as informant-based measures of day-to-day executive functioning and behavior. Trail-making (Reitan & Wolfson, 1985) and Stroop interference (Kramer et al., 2003; Lezak, 2004) were added as comparison tasks drawn from the traditional neuropsychological literature, and the Wide Range Achievement Test (WRAT-4) Reading subject was added as a proxy for verbal intelligence (Griffin, Mindt, Rankin, Ritchie, & Scott, 2002). Three alternate forms were created for all measures. Concurrently, we built a web-based data management system for use during the data collection phase. Finally, subcontract sites for data collection were identified.

Battery development was guided by three basic premises. First, the term "executive function" is overly broad, so smaller conceptual units were needed. Second, executive abilities are measured using tasks that require multiple abilities, so methods that parse the executive component from other skills were preferable. Finally, executive function encompasses both cognitive and non-cognitive behaviors. A multimodal approach using cognitive and observational methods was necessary to capture the broad range of deficits seen in patients with executive dysfunction.

We selected Miyake's model (Miyake et al., 2000) as the core conceptual structure for battery design, targeting tasks that measured mental set shifting, information updating and monitoring, and inhibition of pre-potent responses. For mental set shifting, we emphasized comparing performance when attention or response set must shift to 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 versus manipulation of information in short-term memory. Inhibition of pre-potent responses covered a broad range of tasks that potentially measure cognitive and behavior control. To this core set of constructs we added verbal fluency, a measure with a rich clinical tradition as a measure of executive function. Planning is another concept that is widely considered to be a component of executive functioning, although challenging to operationalize because planning tasks require several component processes like attention, abstract thinking, temporal sequencing, and reasoning. Insight is also often included as a possible measure of executive functioning. 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.

During the second (data collection) phase, the NIH-EXAMINER battery was administered at nine collaborating sites across the country. The final dataset included adults and children, Spanish and English speakers, and multiple diagnostic cohorts. Collaborating sites were: University of California-Davis, University of California-Berkeley, Case Western Reserve, University of Texas Southwestern Medical Center, University of Nebraska-Lincoln, Boston Children's Hospital, University of Iowa, University of South Carolina, and Mayo Clinic-Rochester.

Availability of Materials

All NIH-EXAMINER components are in the public domain and freely available to qualified users upon request at http://memory.ucsf.edu/resources/examiner. There are English and Spanish versions, each with a regular and young (pre-literate) children's version. There are three alternate forms for each version. Record forms are in pdf format. An examiner's manual provides detailed instructions for test administration, scoring, and scale construction. Materials also include software for administering the computerized tasks and generating IRT scores. Because NIH-EXAMINER delivers a cross-platform, open-source technology solution, software installation varies as a function of the particular platform and configuration. Technical support is available through the EXAMINER website. Finally, there is a training video that provides an overview of the battery and demonstrates test administration.

https://doi.org/10.1017/S1355617713001094 Published online by Cambridge University Press

METHODS

NIH-EXAMINER Tasks

Tasks are listed by cognitive domain, including measures of working memory, inhibition, set shifting, fluency, planning, insight, and social cognition and behavior. Our initial goal was to have at least two tasks in each domain represented in Miyake's model, shifting, working memory, and updating, plus at least one task in the domains of planning, insight, social cognition, and behavior.

Each NIH-EXAMINER task potentially yields multiple dependent variables. For the purposes of scale construction, however, a single dependent variable was identified for each task based on reliability and psychometric properties.

Testing formats included computerized and paper-and-pencil. During initial data collection, computer tasks were programmed using E-Prime (Psychology Software Tools, 2009), and administered using Dell laptops with 15" screens. To enable NIH-EXAMINER to work on multiple operating systems using open-source, readily available software, all computer tasks were reprogrammed to run in PsychoPy (http://www.psychopy.org/), an open-source application that allows the presentation of stimuli and collection of data for a wide range of neuroscience, psychology, and psychophysics experiments. NIH-EXAMINER battery is distributed with copies of the installation files for PsychoPy and is compatible with Windows, Apple OS, and Ubuntu. Alternate form testing indicated that the E-prime and PsychoPy versions yield equivalent data.

Domain: Working Memory

Dot counting

This verbal working memory task was modeled after the counting span task of Case, Kurland, and Goldberg (1982; Conway et al., 2005). The examinee looks at a computer screen with a mixed array of green circles, blue circles and blue squares, and instructed to count all of the blue circles on the screen one at a time 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 trials. After counting the blue circles on all of the displays presented within a trial, the examinee recalls the total number of blue circles 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 recalls correctly from each trial.

N-back

The n-back paradigm is a widely used measure of working memory that requires flexible updating capabilities (Owen, McMillan, Laird, & Bullmore, 2005). NIH-EXAMINER includes spatial 1-back and 2-back tasks 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 2 locations.

During both the 1-back and 2-back, the examinee is shown a series of 2.4 cm white squares that appear in 15 different locations on a black computer screen. Each square is presented for 1000 ms. 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. Responses should be given as quickly as possible while maintaining accuracy. 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 ms after each response and remains on the screen for 1000 ms. The examinee says this number out loud immediately when it appears on the screen before responding to the next square. This prevents the examinee from visually fixating on the location of the previous square. The 1-back consists of one block of 30 trials, 10 of which match the location of the previous square, and 20 that are in a different location. 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. 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 primary dependent variable for the n-back tasks is a d-prime measure that incorporates both correct hits and correct rejections, although trial information, including accuracy and reaction time, is also recorded and available for analysis.

Tasks not included in battery

A delayed matching-to-sample task was piloted, but dropped from the battery when preliminary analyses indicated ceiling effects.

Domain: Inhibition

Flanker

The flanker is a widely used measure of response inhibition and cognitive control (Krueger et al., 2009). On this computer-administered task, the examinee is instructed to focus on a small cross in the center of the screen. After a variable duration (1000–3000 ms), a row of five arrows is presented in the center of the screen either above or below the fixation point. The duration of stimulus presentation for each trial is 1000 ms. The examinee indicates whether the center arrow is pointing either to the left or right by pressing the left or right arrow key. There are two different conditions during the task. In the congruent condition, the non-target arrows point in the same direction as the target arrow and in the incongruent trials they point in the opposite direction. The stimuli are

https://doi.org/10.1017/S1355617713001094 Published online by Cambridge University Press

presented in a random order with each condition being presented 24 times resulting in 48 total trials.

Software for the flanker provides trial by trial information on accuracy and reaction time, in addition to tabulating the total accuracy and median reaction time for all correct congruent and incongruent trials. For investigators interested in a single score, the software also runs a scoring algorithm combining accuracy and reaction time that is the same as the one applied to the flanker task in the NIH Toolbox (Weintraub et al., 2013). The score ranges from 0–10, and enables combining data from younger or more impaired individuals for whom accuracy is more variable, and data from more intact individuals for whom reaction time is more variable.

Continuous Performance Test (CPT)

The continuous performance task is a classic response inhibition task that requires subjects to respond to a certain type of stimulus and withhold a response to others. The examinee is presented with different images in the center of the computer screen and instructed to press the left arrow key for only the target image (e.g., a white five-pointed star), responding as quickly and accurately as possible. The task consists of 100 experimental trials, 80% of which are the target image. The non-target images are similar in shape and size to the target. The primary dependent measure from this task is the number of false alarms to the non-target images.

Anti-saccades

This task measures control over eye movements (Munoz & Everling, 2004). 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.

Tasks not included in battery

Pilot data were collected on random number and random letter generation tasks because of their potential utility as an executive function measure (Peters, Giesbrecht, Jelicic, & Merckelbach, 2007). Preliminary data suggested that behavioral variant frontotemporal dementia (bvFTD) patients were less capable of avoiding overlearned sequencing relative to age-matched controls (Schenk, Berhel, Verde, Widmeyer, & Kramer, 2008). The task was dropped from the battery, however, because of difficulty deriving a reliable outcome measure and poor suitability for younger subjects.

Domain: Set Shifting

Dimensional set shifting

This computer-administered task was modeled after paradigms used in the cognitive neuroscience literature (Kray & Lindenberger, 2000; Monsell, 2003). Participants match a stimulus on the top of the screen to one of two stimuli in the lower corners of the screen. In the beginning of each trial, the dimension on which to match (color *vs.* shape) appears in the bottom of the screen (see Figure 1). The version of the task for pre-literate children uses a voice rather than a written word to instruct examinees. In task-homogeneous blocks, participants match to either color or shape. 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 two homogenous blocks each consist of 20 trials for which the same cue is presented. The heterogeneous block consists of 64 trials, 32 of which have a color cue and 32 of which have a shape cue.

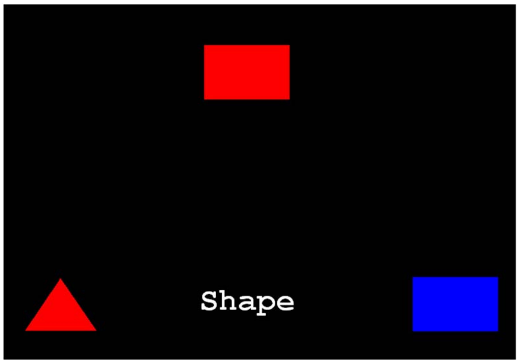

Fig. 1. Example of dimensional set shifting. Examinees are instructed to match by shape, hence the correct response is the blue rectangle.

The software for set shifting provides trial information on accuracy and reaction time, and tabulates total accuracy and median reaction time for all correct trials in the color-only, shape-only, and shifting blocks, plus shift versus non-shift trials within the shifting block. For investigators interested in a single score, the software also runs a scoring algorithm similar to the one used for the flanker task (Weintraub et al., 2013).

Tasks not included in battery

We collected data on a design fluency task that contained a shifting condition, but the task was dropped from the battery because it was not in the public domain.

Domain: Fluency

Phonemic fluency

Examinees are instructed to quickly name as many words as they can that begin with a particular letter of the alphabet. There are two separate phonemic fluency trials. Sixty seconds are allowed for each letter. The examinee is instructed that

https://doi.org/10.1017/S1355617713001094 Published online by Cambridge University Press

names of people and places, numbers, and grammatical variants of previous responses (plurals, altered tenses, and comparatives) are not acceptable responses. All responses are recorded by the examiner. The number of correct responses, repetitions, and rule violations are then totaled for each letter.

Category fluency

Examinees are instructed to quickly generate as many items as possible belonging to a particular category. There are two separate category fluency trials. 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.

Tasks not included in battery

We collected data on a design fluency task that was dropped from the battery because it was not in the public domain.

Domain: Planning

Unstructured task

This task was modeled after the six-elements test (Shallice & Burgess, 1991). Examinees are presented with three booklets, each containing five pages of simple puzzles (four 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 s (see Figure 2). Each puzzle has a designated point value, and subjects are given 6 min 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. The primary dependent measure was the sum of the proportion of puzzles completed that had a low cost-benefit ratio and the log of the total score.

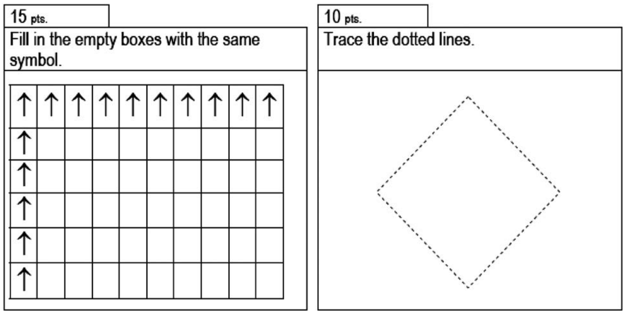

Fig. 2. Example of unstructured task stimuli. Although the puzzle on the left offers more points, the puzzle on the right offers a better benefit:cost ratio because it can be completed much more quickly.

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, examinees are informed that after performing the task they will be asked to assess their 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.

Domain: Social Cognition and Behavior

The Social Norms Questionnaire (SNQ)

This task measures crystallized knowledge of social norms in a linguistically and cognitively simple manner, and is designed to determine the degree to which subjects understand and accurately identify implicit but widely accepted social boundaries in the dominant United States culture. Because social norms vary across cultures and subcultures, this version is considered valid only when administered to individuals expected to be well-acculturated to the dominant United States culture. To confirm that these rules were normative, items survived the test development process only if they demonstrated a high level of agreement across healthy, racially heterogeneous individuals who had lived in the US for decades.

The SNQ was initially piloted with 12 situationally matched pairs of items, each with one item in which the behavior described would be socially appropriate (e.g., “Eat ribs with your fingers”) and one inappropriate (e.g., “Eat pasta with your fingers”). This initial 24-item questionnaire was tested with 38 healthy controls between the ages of 45 and 87 who were fluent in English and well-acculturated to the dominant U.S. culture (i.e., had lived in this country >20 years), after which two items were removed due to lack of agreement (i.e., <80%) among respondents. The final 22-item SNQ was then performed with 84 additional healthy controls, for a total of 122 individuals in the normative sample (ages, 46–92 years; 57 males/65 females). No significant gender differences in SNQ performance were found.

The final version of the SNQ includes both socially inappropriate behaviors (e.g., “Cut in line if you are in a hurry,” “Pick your nose in public”) and generally acceptable behaviors (e.g., “Tell a coworker your age,” “Blow your nose in public”). Examinees decide whether the behavior is socially appropriate 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.

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. All behaviors are rated on a 4-point Likert scale.

Tasks not included in battery

We piloted tasks that involved video presentation of social interactions and asking subjects to identify emotions. These were dropped because they were not in the public domain and were potentially prone to age and cultural differences.

Tabulated Scores

An underlying assumption in developing NIH-EXAMINER 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 (Kramer et al., 2003; Stuss, Floden, Alexander, Levine, & Katz, 2001). Accordingly, we developed a summary error score across several NIH-EXAMINER tasks, including 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.

PROCEDURES

Participants

A total of 1248 participants were administered the NIH-EXAMINER battery. All data collection was obtained in compliance with the Institutional Review Board at each participating institution.

The sample included 485 participants who were below the age of 18 years (range, 3–17 years) and 763 participants 18 years and older (range, 18–94 years). These cohorts are summarized in Tables 1 and 2.

In addition, 122 healthy subjects were tested a second time within 125 days of their initial testing (mean interval = 19.7 days; *SD* = 23.6), and 110 healthy subjects were tested a second time between 125 and 411 days after their initial testing (mean interval = 360.8 days; *SD* = 14.2). A total of 31 healthy subjects were tested on a third occasion. Nineteen patients (Alzheimer's disease, bvFTD, Huntington's disease, and progressive supranuclear palsy) also returned for a second evaluation after their initial visit (mean interval = 274.7 days; *SD* = 228.0).

Table 1. Diagnostic cohorts: Participants under age 18

Cohort	*n*	Age	% Female
Normal	337	9.9 (3.5)	50.1
ADHD	34	11.9 (2.3)	29.4
Sickle cell anemia	34	12.8 (3.1)	45.5
Very low birth weight	72	11.6 (1.0)	47.1

RESULTS

Scale Construction

We evaluated the psychometric properties and utility of each test by looking at distributions, correlations with age, and group differences (see Table 3). We divided the sample by age group (adults *vs.* children) and by diagnostic group (healthy subjects *vs.* patients) and examined the distributions of each primary variable plus the tabulated error score in each subgroup. Most test scores were normally distributed, with some exceptions including the flanker (negative skewed in all subgroups), anti-saccade (negatively skewed in adults), social norms (negatively skewed in adult controls), and the error score (positively skewed in adults). With the exception of insight, correlations with age in healthy children were all significant, and ranged from .36 to .67. General linear models covarying for age were used to test for group differences; effect sizes are summarized in Table 3.

We also explored ways to reduce the data into a smaller number of scales that could be used by researchers and clinical trials. Insight was excluded from these analyses because it did not discriminate well between healthy subjects and patients. Several exploratory factor analyses were carried out that suggested that a single factor model might explain the majority of the variance, and that a three-factor model was also a good fit for the data, with the three working memory scores comprising one factor, the four fluency measures comprising a second factor, and flanker, set shifting, errors, and anti-saccades forming a third. The unstructured task and Social Norms did not load clearly on any of the three factors.

Confirmatory factor analysis (CFA) informed by conceptual models from previous literature and by results of exploratory factor analysis was used to further evaluate dimensional structure and identify homogenous groups of variables that could be used to generate composite scores. These analyses are described in detail in an online supplement. Two alternative models were tested: (1) a one-factor (unidimensional) model in which 11 primary variables defined a single factor, and (2) a three-factor model with factors representing cognitive control (Flanker, Set Shifting, anti-saccade, dysexecutive errors), working memory (Dot Counting, 1-back, 2-back), and fluency

https://doi.org/10.1017/S1355617713001094 Published online by Cambridge University Press

Table 2. Diagnostic cohorts: Participants 18 years and older

Cohort	*n*	age	Educ	% female
Normal	408	55.4 (19.9)	13.6 (4.3)	60.4
Alzheimer's disease	41	72.2 (10.5)	15.9 (3.0)	31.7
Focal lesion	98	58.1 (11.5)	14.6 (2.6)	38.8
bvFTD	40	63.9 (7.6)	16.2 (2.3)	32.5
Huntington's disease	20	47.6 (11.6)	15.8 (2.8)	60.0
Mild cognitive impairment	51	72.1 (9.9)	15.4 (3.0)	45.1
Multiple sclerosis	17	42.5 (11.9)	16.6 (2.6)	58.8
Parkinson's disease	22	68.0 (7.4)	17.0 (2.3)	22.7
Progressive supranuclear palsy	17	66.3 (7.1)	15.6 (3.4)	58.8
Traumatic brain injury	19	29.9 (10.0)	13.4 (2.3)	44.4

(phonemic and semantic). CFA initially was conducted using the adult sample (age 18+) and invariance across age groups was subsequently tested. Results of the CFA analyses indicated that: (1) the NIH-EXAMINER tests can be well characterized by measures of working memory, fluency, and control, but in addition, (2) a global measure of executive function was also supported by psychometric results.

We used item response theory (IRT) methods (Hambleton, Swaminathan, & Rogers, 1991; Mungas, Reed, & Kramer, 2003) to generate scores corresponding to these four variables: global executive function, cognitive control, fluency, and working memory. These methods are also described in more detail in the online supplement. IRT has important invariance properties, and of particular relevance to NIH-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. Software included in the NIH-EXAMINER materials uses the R ltm module to generate four scores corresponding to global executive function, cognitive control, fluency, and working memory; a standard error of measurement for each score is also included.

DISCUSSION

This study describes the conceptualization and development of NIH-EXAMINER, an NINDS-initiated project to develop an executive function battery that is modular, modifiable, efficient, appropriate for a broad range of ages and ability levels, psychometrically robust, and suitable for clinical trials and clinical research. There are English and Spanish versions, and multiple alternate forms. The core conceptual model for the battery encompassed inhibition, set shifting, working memory, fluency, planning, error monitoring, insight, and social function. The final battery includes both computer-administered and paper-and-pencil tasks, and measures inhibition, set shifting, working memory, fluency, planning, error monitoring, insight, and social function. Confirmatory factor analysis supports both a one-factor model and a three-factor model, and these models formed the basis for an IRT-generated Executive Composite score and smaller scales quantifying Working Memory, Cognitive Control, and Fluency. Test–retest reliabilities range from .78 to .93.

NIH-EXAMINER was designed to be applied in multiple ways. Individual components have been used to study conflict monitoring (Krueger et al., 2009), attention networks

Table 3. Distributions, correlations with age, and effect sizes for group differences (Cohen's d)

Test	Distribution	Correlation with age in healthy children	Healthy controls *vs.* patients in children	Healthy controls *vs.* patients in adults
Dot Counting	Normal	.55***	5.87***	4.93***
1-back	Normal	.42***	5.75***	4.41***
2-back	Normal	.39***	6.25***	1.19
Anti-saccade	Negative skew in adult patients and controls;	.42***	6.00***	7.44***
Set shifting	Normal	.67***	5.54***	3.50***
Flanker	Negative skew	.67***	6.49***	5.31***
Verbal fluency	Normal	.37***	4.57***	4.29***
Category fluency	Normal	.65***	2.03*	7.60***
Unstructured task	Normal	.65***	2.71**	5.27***
Social norms	Negative skew in adult controls	n/a	n/a	4.19***
Error score	Positive skew in adult patients and controls.	−.36***	4.06***	4.06***
Insight	Normal	.03	2.41*	1.31

*$p < .05$; **$p < .01$; ***$p < .001$

https://doi.org/10.1017/S1355617713001094 Published online by Cambridge University Press

(Luks et al., 2010), insight (Krueger et al., 2011), the cognitive correlates of eye movements (Hellmuth et al., 2012; Mirsky et al., 2011), social cognition (Shany-Ur & Rankin, 2011), and dissociable cognitive patterns in neurodegenerative disease (Possin et al., 2013). Applications of the optional IRT-generated Executive Composite and the Working Memory, Cognitive Control, and Fluency scales are illustrated by other manuscripts presented in this series.

As with all tests and batteries, NIH-EXAMINER will prove to be most suitable for specific applications. Individual tasks will have particular utility for researchers interested in cognitive neuroscience paradigms like flanker and set shifting, and who plan to analyze detailed trial-by-trial accuracy and reaction time data. The novel planning, insight, and social cognition tasks have broad potential applications in clinical research. Research that relies more on reliable and psychometrical robust measures of key executive constructs are more likely to use the IRT-generated scales. These may have particular utility for studies comparing different aspect of executive functioning where having psychometrically matched scales is important (see Schreiber et al. in this series). The IRT-generated scales are also well suited for longitudinal research and especially for clinical trials. Importantly, different subsets of the donor scales contributing to the Executive Composite can be selected depending on the specific research question and subject population, providing researchers with some degree of flexibility when designing studies.

Research to date, including the studies described in this series, support the utility and validity of NIH-EXAMINER. This work is still in a very early stage, however. Studies correlating EXAMINER with other batteries measuring executive functioning like DKEFS (Delis, Kaplan, & Kramer, 2001), CANTAB (Cambridge-Cognition, 1996), Behavioral Assessment of the Dysexecutve Syndrome (Wilson, Alderman, Burgess, Emslie, & Evans, 1996), and Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales (Flores, Ostrosky, & Lozano, 2008) will be important, and a more definitive sense of how the battery and its component parts are best used will require accumulation of experience in the years to come.

ACKNOWLEDGMENTS

This work was supported by NINDS contract HHSN271200623661C, P50 AG023501, 1R01AG022983 and 1R01AG032289. None of the authors have any conflicts to disclose.

Supplementary materials

To view supplementary material for this article, please visit http://dx.10.1017/S1355617713001094

REFERENCES

Amieva, H., Phillips, L., & Della Sala, S. (2003). Behavioral dysexecutive symptoms in normal aging. *Brain and Cognition*, *53*(2), 129–132.

Arnett, P.A., Rao, S.M., Grafman, J., Bernardin, L., Luchetta, T., Binder, J.R., & Lobeck, L. (1997). Executive functions in multiple sclerosis: An analysis of temporal ordering, semantic encoding, and planning abilities. *Neuropsychology*, *11*(4), 535–544.

Aron, A.R., Watkins, L., Sahakian, B.J., Monsell, S., Barker, R.A., & Robbins, T.W. (2003). Task-set switching deficits in early-stage Huntington's disease: Implications for basal ganglia function. *Journal of Cognitive Neuroscience*, *15*(5), 629–642.

Asimakopulos, J., Boychuck, Z., Sondergaard, D., Poulin, V., Menard, I., & Korner-Bitensky, N. (2012). Assessing executive function in relation to fitness to drive: a review of tools and their ability to predict safe driving. *Australian Occupational Therapy Journal*, *59*(6), 402–427. doi:10.1111/j.1440-1630.2011.00963

Barkley, R.A. (2010). Differential diagnosis of adults with ADHD: The role of executive function and self-regulation. *Journal of Clinical Psychiatry*, *71*(7), e17. doi:10.4088/JCP.9066tx1c

Boone, K.B., Miller, B.L., Lee, A., Berman, N., Sherman, D., & Stuss, D.T. (1999). Neuropsychological patterns in right versus left frontotemporal dementia. *Journal of International Neuropsychological Society*, *5*(7), 616–622.

Buckner, R.L. (2004). Memory and executive function in aging and AD: Multiple factors that cause decline and reserve factors that compensate. *Neuron*, *44*(1), 195–208.

Caeyenberghs, K., Leemans, A., Leunissen, I., Gooijers, J., Michiels, K., Sunaert, S., & Swinnen, S.P. (2012). Altered structural networks and executive deficits in traumatic brain injury patients. *Brain Structure and Function*. doi:10.1007/s00429-012-0494-2

Cahn-Weiner, D.A., Boyle, P.A., & Malloy, P.F. (2002). Tests of executive function predict instrumental activities of daily living in community-dwelling older individuals. *Applied Neuropsychology*, *9*(3), 187–191.

Cambridge-Cognition. (1996). *CANTAB®* . Cambridge: Cambridge Cognition Limited.

Case, R., Kurland, M.D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*, 386–404.

Chiaravalloti, N.D., & DeLuca, J. (2003). Assessing the behavioral consequences of multiple sclerosis: An application of the Frontal Systems Behavior Scale (FrSBe). *Cognitive and Behavioral Neurology*, *16*(1), 54–67.

Conway, A.R., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., & Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin Review*, *12*(5), 769–786.

Delis, D.C., Kaplan, E., & Kramer, J.H. (2001). *Delis-Kaplan Executive Function System*. San Antonio: The Psychological Corporation.

Flores, J., Ostrosky, F., & Lozano, A. (2008). *Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales* (Battery of Executive Functions and Frontal Lobe, Spanish) San Antonio: Pearson, Inc.

Foong, J., Rozewicz, L., Quaghebeur, G., Davie, C.A., Kartsounis, L.D., Thompson, A.J., … Ron, M.A. (1997). Executive function in multiple sclerosis. The role of frontal lobe pathology. *Brain*, *120*(Pt 1), 15–26.

Gerstenecker, A., Mast, B., Duff, K., Ferman, T.J., & Litvan, I. (2013). Executive dysfunction is the primary cognitive impairment in progressive supranuclear palsy. *Archives of Clinical Neuropsychology*, *28*(2), 104–113. doi:10.1093/arclin/acs098

Gioia, G.A., Isquith, P.K., Guy, S.C., & Kenworthy, L. (2000). Behavior rating inventory of executive function. *Child Neuropsychology*, *6*(3), 235–238. doi:10.1093/arclin/acs098

https://doi.org/10.1017/S1355617713001094 Published online by Cambridge University Press

Griffin, S.L., Mindt, M.R., Rankin, E.J., Ritchie, A.J., & Scott, J.G. (2002). Estimating premorbid intelligence: Comparison of traditional and contemporary methods across the intelligence continuum. *Archives of Clinical Neuropsychology*, *17*(5), 497–507. doi:S0887617701001366 [pii]

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hellmuth, J., Mirsky, J., Heuer, H.W., Matlin, A., Jafari, A., Garbutt, S., … Boxer, A.L. (2012). Multicenter validation of a bedside antisaccade task as a measure of executive function. *Neurology*, *78*(23), 1824–1831. doi:10.1212/WNL.0b013e318258f785

Hutchinson, A.D., & Mathias, J.L. (2007). Neuropsychological deficits in frontotemporal dementia and Alzheimer's disease: A meta-analytic review. *Journal of Neurology, Neurosurgery, and Psychiatry*, *78*(9), 917–928. doi:10.1136/jnnp.2006.100669

Kramer, J.H., Jurik, J., Sha, S.J., Rankin, K.P., Rosen, H.J., Johnson, J.K., & Miller, B.L. (2003). Distinctive neuropsychological patterns in frontotemporal dementia, semantic dementia, and Alzheimer disease. *Cognitive and Behavioral Neurology*, *16*(4), 211–218.

Kray, J., & Lindenberger, U. (2000). Adult age differences in task switching. *Psychology of Aging*, *15*(1), 126–147.

Krueger, C.E., Bird, A.C., Growdon, M.E., Jang, J.Y., Miller, B.L., & Kramer, J.H. (2009). Conflict monitoring in early frontotemporal dementia. *Neurology*, *73*(5), 349–355. doi:10.1212/WNL.0b013e3181b04b24

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. *Clinical Neuropsychologist*, *25*(5), 741–756. doi:10.1080/13854046.2011.569759

Levin, H.S., & Hanten, G. (2005). Executive functions after traumatic brain injury in children. *Pediatric Neurology*, *33*(2), 79–93.

Lezak, M.D. (2004). *Neuropsychological assessment* (4th ed.). New York, NY, USA.

Luks, T.L., Oliveira, M., Possin, K.L., Bird, A., Miller, B.L., Weiner, M.W., & Kramer, J.H. (2010). Atrophy in two attention networks is associated with performance on a Flanker task in neurodegenerative disease. *Neuropsychologia*, *48*(1), 165–170. doi:10.1016/j.neuropsychologia.2009.09.001

Malloy, P., & Grace, J. (2005). A review of rating scales for measuring behavior change due to frontal systems damage. *Cognitive and Behavioral Neurology*, *18*(1), 18–27.

Mirsky, J.B., Heuer, H.W., Jafari, A., Kramer, J.H., Schenk, A.K., Viskontas, I.V., … Boxer, A.L. (2011). Anti-saccade performance predicts executive function and brain structure in normal elders. *Cognitive and Behavioral Neurology*, *24*(2), 50–58. doi:10.1097/WNN.0b013e318223f6c6

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. *Cognitive Psychology*, *41*(1), 49–100.

Monsell, S. (2003). Task switching. *Trends in Cognitive Science*, *7*(3), 134–140. doi:S1364661303000287 [pii]

Moorhouse, P., Song, X., Rockwood, K., Black, S., Kertesz, A., Gauthier, S., & Feldman, H. (2010). Executive dysfunction in vascular cognitive impairment in the consortium to investigate vascular impairment of cognition study. *Journal of Neurological Science*, *288*(1-2), 142–146. doi:10.1016/j.jns.2009.09.017

Mungas, D., Reed, B.R., & Kramer, J.H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology*, *17*(3), 380–392.

Munoz, D.P., & Everling, S. (2004). Look away: The anti-saccade task and the voluntary control of eye movement. *Nature Reviews: Neuroscience*, *5*(3), 218–228.

Owen, A.M., McMillan, K.M., Laird, A.R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59. doi:10.1002/hbm.20131

Paulsen, J.S. (2011). Cognitive impairment in Huntington disease: Diagnosis and treatment. *Current Neurology and Neuroscience Reports*, *11*(5), 474–483. doi:10.1007/s11910-011-0215-x

Peters, M., Giesbrecht, T., Jelicic, M., & Merckelbach, H. (2007). The random number generation task: Psychometric properties and normative data of an executive function task in a mixed sample. *Journal of International Neuropsychological Society*, *13*(4), 626–634. doi:S1355617707070786

Possin, K.L., Feigenbaum, D., Rankin, K.P., Smith, G.E., Boxer, A.L., Wood, K., … Kramer, J.H. (2013). Dissociable executive functions in behavioral variant frontotemporal and Alzheimer dementias. *Neurology*, *80*(24), 2180–2185.

Psychology Software Tools, I. (2009). E-Prime 2. Sharpsburg, PA.

Ravizza, S.M., & Ciranni, M.A. (2002). Contributions of the prefrontal cortex and basal ganglia to set shifting. *Journal of Cognitive Neuroscience*, *14*(3), 472–483.

Reed, B.R., Eberling, J.L., Mungas, D., Weiner, M., Kramer, J.H., & Jagust, W.J. (2004). Effects of white matter lesions and lacunes on cortical function. *Archives of Neurology*, *61*(10), 1545–1550.

Reitan, R.M., & Wolfson, D. (1985). *The Halstead–Reitan Neuropsychological Test Battery*. Tucson: Neuropsychology Press.

Schenk, A., Berhel, A., Verde, S., Widmeyer, M., & Kramer, J.H. (2008). Assessing the mental rigidity of FTD patients using a random number generation task [Abstract]. *Journal of the International Neuropsychological Society*, *14*(S1), 266.

Shallice, T., & Burgess, P.W. (1991). Deficits in strategy application following frontal lobe damage in men. *Brain*, *114*, 727–741.

Shany-Ur, T., & Rankin, K.P. (2011). Personality and social cognition in neurodegenerative disease. *Current Opinion in Neurology*, *24*(6), 550–555.

Slachevsky, A., Villalpando, J.M., Sarazin, M., Hahn-Barma, V., Pillon, B., & Dubois, B. (2004). Frontal assessment battery and differential diagnosis of frontotemporal dementia and Alzheimer disease. *Archives of Neurology*, *61*(7), 1104–1107.

Stuss, D.T. (2011). Traumatic brain injury: Relation to executive dysfunction and the frontal lobes. *Current Opinion in Neurology*, *24*(6), 584–589. doi:10.1097/WCO.0b013e32834c7eb9

Stuss, D.T., Floden, D., Alexander, M.P., Levine, B., & Katz, D. (2001). Stroop performance in focal lesion patients: Dissociation of processes and frontal lobe lesion location. *Neuropsychologia*, *39*(8), 771–786.

Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Bauer, P.J., … Gershon, R.C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S54–S64. doi:10.1212/WNL.0b013e3182872ded

Wilson, B., Alderman, N., Burgess, P.W., Emslie, H., & Evans, J.J. (1996). *Behavioural Assessment of the Dysexecutive Syndrome*. Bury St. Edmunds: Thames Valley Test Company.