NAB™

Neuropsychological Assessment Battery™

Psychometric and Technical Manual

Travis White, PhD

Robert A. Stern, PhD

PAR Psychological Assessment Resources, Inc. • 16204 N. Florida Avenue • Lutz, FL 33549 • 1.800.331.8378 • www.parinc.com

Development of the Neuropsychological Assessment Battery was made possible and funded in part by grants 1 R43 MH58501-01 and 2 R44 MH58501-02 from the National Institute of Mental Health.

9 8 7 6 5 4 3 2 1 Reorder #RO-5089 Printed in the U.S.A.

2

Development of the NAB

This chapter describes the development of the NAB, including the overall rationale, goals, standards, and processes used throughout the creation of the battery, as well as the specific steps and guidelines used for the development of each test.

NEUROPSYCHOLOGICAL FUNCTIONS MEASURED BY THE NAB

Decisions pertaining to the selection of specific neuropsychological functions to measure with the NAB were guided in large part by the results of the publisher's 1997 survey of neuropsychological assessment practices (Stern & White, 2000; see chapter 1 for additional information about the survey). The survey included 79 separate neuropsychological functions that the respondents were asked to rate with regard to how important they were for inclusion in a new, *briefer but comprehensive* neuropsychological test battery. Ratings were made according to a 4-point Likert-type scale: 1 = *Not at all important*, 2 = S*lightly important*, 3 = *Moderately important*, and 4 = *Very important*. The universe of functions to be included in the survey was based on a review of the major texts on neuropsychological assessment at the time (e.g., Lezak, 1995; Spreen & Strauss, 1991) and on the functions purportedly measured by existing neuropsychological batteries (e.g., Benton, Hamsher, & Sivan, 1994; Benton, Sivan, Hamsher, Varney, & Spreen, 1994; Golden et al., 1985; Goodglass & Kaplan, 1983; Reitan & Wolfson, 1993; Schmidt & Tombaugh, 1995; Wechsler, 1987; Williams, 1991).

Tables 2.1 through 2.6 summarize the percentage of respondents who rated each function as very important for inclusion in the new battery. Functions rated as very important by one third or more (≥33%) of the sample were included in the development of the NAB, with two exceptions. "Writing ability" (rated as very important by only 17% of respondents) and "Oral (speech) production" (not included in the survey) were included in the NAB at the recommendation of the development team's consulting

Table 2.1
Survey Results for the Attention Domain

Function	% of respondents who rated function as very important
Attentional capacity	70
Sustained attention/Vigilance	51
Information processing speed	43
Divided attention	37
Orientation	34
Mental tracking (i.e., working memory)	33
Psychomotor speed	24
Neglect/Hemi-inattention	21
Reaction time	14

Note. *N* = 888; functions rated as very important by 33% or more of the survey respondents were included in the NAB.

Table 2.2
Survey Results for the Language Domain

Function	% of respondents who rated function as very important
Auditory comprehension	51
Confrontation naming	40
Written comprehension/Reading ability	33
Calculation skills	31
Word and phrase repetition	19
Money skills	18
Writing ability	17
Prosodic comprehension	12
Prosodic expression	11
Humor comprehension	6

Note. $N = 888$; functions rated as very important by 33% or more of the survey respondents were included in the NAB.

Table 2.3
Survey Results for the Memory Domain

Function	% of respondents who rated function as very important
Verbal delayed recall	69
Verbal recognition memory	61
Word list learning and immediate recall	60
Visual/Nonmotor delayed recall	56
Prose/Paragraph immediate recall	55
Visual/Nonmotor learning and immediate recall	55
Visual/Nonmotor recognition memory	49
Verbal sensitivity to interference	31
Verbal paired-associate learning	25
Personal/Autobiographical remote recall	20
Incidental learning	17
Other remote recall (e.g., public events)	16
Priming	8

Note. $N = 888$; functions rated as very important by 33% or more of the survey respondents were included in the NAB.

aphasiologist/speech language pathologist. Other functions rated as very important by fewer than 33% of the survey sample were included only as part of NAB tests that are multifactorial in nature and are intended to tap into more than one functional domain, such as the NAB Daily Living tests.

More than one-third of survey respondents rated measures of malingering (e.g., symptom validity testing) and premorbid intelligence estimates as very important in a new neuropsychological battery. However, in order to keep the overall administration time under 3 hours (excluding Screening Module) and to provide a comprehensive assessment of the primary neuropsychological domains in a modular fashion, the development team addressed the issues of malingering and premorbid intelligence in an alternative manner. A simulated malingering study of the NAB was conducted as part of the development process (Ropacki, 2003; Turner, Ropacki, & Hinkin, 2003). For this study, the entire NAB was administered concurrently with existing

Table 2.4
Survey Results for the Spatial Domain

Function	% of respondents who rated function as very important
Visuoconstruction skills – Drawing	48
Visuoconstruction skills – Blocks or puzzles	46
Visual perception	39
Spatial analysis (i.e., visuospatial skill)	39
Visuoconstruction skills – Nonmanual tasks	30
Visual scanning	24
Right–Left orientation	16
Facial recognition	12
Geographic orientation	7

Note. *N* = 888; functions rated as very important by 33% or more of the survey respondents were included in the NAB.

Table 2.5
Survey Results for the Executive Functions Domain

Function	% of respondents who rated function as very important
Response set/Cognitive flexibility	59
Verbal abstraction/Conceptualization	55
Planning	51
Organization	49
Verbal fluency	47
Disinhibition/Impulse control	45
Visual abstraction/Conceptualization	42
Perseveration	41
Self-monitoring	24
Visual/Design fluency	22
Proverb interpretation	9
Cognitive estimation (size, shape, distance)	7

Note. *N* = 888; functions rated as very important by 33% or more of the survey respondents were included in the NAB.

symptom validity tests to a sample of 50 participants who were given specific "coaching" instructions to simulate malingering. The concurrent malingering tests included the Test of Memory Malingering (TOMM; Tombaugh, 1996), the Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 1997), and the Word Memory Test (WMT; Green, Allen, & Astner, 1995). The results of this study provide important information about NAB responding that may be difficult to feign in a sophisticated manner, as well as the relationship between selected NAB scores and existing malingering measures. The methodology and results of this study are presented in detail in chapter 6.

To address the issue of premorbid intelligence, a measure of intelligence was administered concurrently with the NAB during the standardization of the battery. All NAB standardization participants also completed the Reynolds Intellectual Screening Test (RIST; Reynolds & Kamphaus, 2003), a new measure of intelligence with strong psychometric properties. The correlations between the NAB and RIST scores are presented in chapter 6. Means and standard deviations of NAB scores are presented in Appendix B by four levels of estimated intelligence. In addition to relying on the information presented in this manual about the relationships between NAB scores and measures of malingering and

Table 2.6
Miscellaneous Survey Results

Function	% of respondents who rated function as very important
Malingering/Symptom validity testing/Effort testing	42
Premorbid Verbal IQ estimate	37
Premorbid Performance IQ estimate	33
Internal mood state	28
Self-awareness of cognitive skills and deficits (anosognosia)	23
Praxis	21
Reading achievement	20
Fine motor speed	20
Fine motor dexterity	19
Suicidal ideation/Risk	18
Syndromal depression	17
Anxiety	17
Socially inappropriate behavior	16
Arithmetic achievement	15
Personality assessment	14
Psychosis	12
Aggressive behavior	12
Apathy	11
Grip strength	10
Other auditory perception	10
Spelling achievement	10
Finger gnosis	9
Other tactile perception	8
Auditory rhythm discrimination	7
Fingertip writing perception	6
Odor identification	6

Note. *N* = 888; functions rated as very important by 33% or more of the survey respondents were included in the NAB or addressed by the NAB standardization and validation studies.

intelligence, the examiner can supplement the NAB by also administering in-depth measures of malingering and intelligence when the referral question so dictates.

GENERAL PRINCIPLES GUIDING THE DEVELOPMENT OF THE NAB

Once decisions were made about which neuropsychological functions to measure with the NAB, decisions about specific task creation were made. These decisions followed several general principles that are described in the following sections.

Tasks Must be Easy to Administer and Score

To increase the reliability and consistency of administration and scoring procedures among a variety of examiners (including non-doctoral technicians and experienced professional clinicians), NAB tasks were designed to have relatively simple administration and scoring procedures. Modular record forms, response booklets, and stimulus books are used, and all administration and scoring instructions are included in the record forms. In addition, the NAB includes only two manipulatives, the Screening and Spatial Modules Design Construction tans (i.e., flat, plastic geometric shapes) and the map used for the Spatial Module Map Reading test.

Stimuli Must be Attractive and Face Valid

To ensure that both examinees and examiners find the stimulus materials pleasing and enjoyable to work with, high quality graphics and artwork were carefully produced. For example, color photography was included in many tasks, rather than the more commonly used line drawings. Moreover, the Advisory Council rated the attractiveness and face validity of the NAB visual stimuli, and only those items and tasks with the highest ratings were included in the NAB.

Total Administration Time Must Be Three Hours or Less

The survey results clearly indicated the need for a comprehensive neuropsychological test battery that requires substantially less time to administer than a comparable battery of similar tests. The survey results clearly indicated the ideal time for the full NAB (excluding Screening Module) was 3 hours or less. Not exceeding 3 hours of administration time was a major factor in decisions about the nature and length of NAB tests throughout the item-writing, pilot testing, and standardization phases of development.

Large Pool of Items Represent a Wide Range of Difficulty Levels

Survey results underscored the need to avoid both floor and ceiling effects in a comprehensive neuropsychological battery. To this end, each task was initially created with items representing a wide range of difficulty levels. Advisory Council ratings of the difficulty level of all items and tests ensured appropriate variability, with specific emphasis on a range of difficulty represented in the Screening Module tests, wherever possible. The NAB was pilot tested in the spring of 2001 with a heterogeneous sample of healthy participants and patients with known neurological disorders. The pilot test results were used to refine the item pool, drop poorly performing items, establish the order of items in ascending difficulty, and revise administration, recording, and scoring procedures.

Relationship Between Screening Module and Main Module Tests and Items Must Be Meaningful

The primary goal for the Screening Module was the development of a brief yet sensitive measure of overall neuropsychological functioning that provides meaningful test data that facilitate decisions about the need for further, more in-depth neuropsychological testing. The Screening Module measures the same five functional domains that are measured by the main NAB modules: attention, language, memory, spatial skills, and executive functions. A related goal was the development of Screening Domain scores that psychometrically predict performance on the corresponding Module Index scores within the same functional domain. Screening Module tests are either (a) *similar* to main module tests but with different stimuli and task parameters (e.g., Shape Learning, Story Learning), (b) *shorter versions* of the same tests included in the main modules (e.g., Numbers & Letters, Mazes), or (c) *identical* to the main module tests (e.g., Orientation, Digits Forward). In the descriptions of the development of individual tests later in this chapter, the Screening Module tests are included under their associated domain module tests, in order to elucidate the relationship between the Screening Module and domain module versions.

Theoretical Foundation Must Combine Empiricism and Cognitivism

Historically, neuropsychological test construction has followed one of two underlying theoretical foundations: (a) empiricism or (b) cognitivism (Hebben & Milberg, 2002). Empiricism underlies the majority of existing tests and is based on the notion that *clinical prediction* is of primary interest, with content and neuropsychological meaning secondary. In contrast, cognitivism is based on the view that the underlying *neuropsychological constructs* are paramount, with clinical prediction secondary. The development of the NAB incorporated both of these traditions. That is, NAB tests were created to be sensitive and specific with regard to clinical prediction. Validity studies, presented in chapter 6, provide supportive evidence of this approach. However, underlying constructs rooted in cognitive psychology (e.g., Kellogg, 2002; Sternberg, 1999) and cognitive neuropsychology (e.g., Morris, 1997; Rapp, 2001) also guided the selection of task paradigms and item content. In all cases, tests were designed to measure one or more specific aspects of the five functional domains included in the NAB (i.e., Attention, Language, Memory, Spatial, and Executive Functions). The goal was the inclusion of items and tests that provide a broad and representative sampling of the domains being measured.

Test Names Should Describe the Content and/or Procedures Involved

In most cases, test names were selected to describe the *content* of the test materials and/or *procedures* involved in the task, rather than any purported or hypothesized underlying neuropsychological function/construct that the test may possibly measure (e.g., “working memory,” “planning”). This decision was made for two primary reasons. First, an individual test may measure more than one underlying neuropsychological construct. Second, the actual constructs underlying a specific test will be determined through a dynamic process over years of future validation and experimental research. This does not mean, however, that theoretical constructs were not considered in the design of each test. Such constructs underlying the NAB tests are described later in the discussion of the development of each individual NAB test. In addition, results of convergent validity analyses, presented in chapter 6, generally confirm the existence of these constructs, insofar as existing neuropsychological tests measure the constructs in question. It is possible, however, that future research with the NAB will suggest modifications to the constructs attributed to each test.

Advisory Council Ratings Must Inform Development Activities

Throughout the development of the NAB, the NAB Advisory Council members and the NAB language/aphasia consultant provided important guidance in the creation, refinement, and final selection of test items and administration procedures. The NAB Advisory Council members are listed in Appendix A. In the earlier stages of test design, these advisors provided guidance about the spectrum of neuropsychological functions to be assessed with the NAB. This process is an important component of test development (Anastasi & Urbina, 1997; Haynes, Richard, & Kubany, 1995) and was used to facilitate high levels of content validity for the NAB tests. Advisory Council members also provided open-ended feedback about task design, stimuli, and individual items. This feedback was then used to completely or partially revise tasks and stimuli. Another major focus of the Advisory Council was rating the initial pool of test items on a variety of dimensions. The test/item characteristics rated are listed in Table 2.7. Each characteristic was rated on a 5-point Likert-type scale (1 = *low*, 5 = *high*). The Advisory Council reviewed every item, question, and stimulus, and then rated the test characteristics relevant to that specific test. The ratings were used to narrow the pool of items and to divide items into two initial forms.

Advisory Council ratings were averaged across the raters. Tasks and items with unacceptably high levels of potential sex, education, ethnic/racial/cultural, or U.S. geographic/regional biases were eliminated first. Test satisfaction ratings, both overall satisfaction with a test and satisfaction with a specific feature, such as visual design or photograph selection, were used to further eliminate items/tasks. Only those tasks with adequate or better satisfaction ratings were

Table 2.7
Test/Item Characteristics Rated by the Advisory Council

Test/Item characteristics
Ability to be verbally encoded (Shape Learning, Visual Discrimination)
Clinical utility (Judgment)
Design satisfaction (Shape Learning, Design Construction, Visual Discrimination)
Difficulty
Ecological validity (Daily Living tests)
Education bias
Ethnic/Racial/Cultural bias
Sex bias
Linguistic demands (Bill Payment)
Overall satisfaction with task
Photograph satisfaction (Naming, Reading Comprehension, and Categories)
Quality of artwork (Driving Scenes)
Reading difficulty (Reading Comprehension)
Satisfaction for phonemic cue (Naming)
Satisfaction for semantic cue (Naming)
Stimulus satisfaction (Dots, Map Reading)
Task appropriateness
U.S. regional bias

retained. Additional test-specific ratings were also used to further eliminate items and to assign items to Form 1 or Form 2. Finally, difficulty ratings were used to initially order the items in ascending difficulty and to equate the two forms for difficulty. The specific procedures used to create the NAB tests are described in the subsequent sections of this chapter.

ATTENTION MODULE

Orientation

Background

Impaired orientation is one of the most common symptoms of a variety of brain disorders. Disorientation to place and to time are the most common of these difficulties and are associated with disorders in which the patient has significantly impaired attention and/or retention (Lezak, 1995). Because of this, questions about orientation to time and to place are included in most mental status tests (e.g., Folstein, Folstein, & Fanjiang, 2001), dementia examinations (e.g., Jurica et al., 2001), and memory test batteries (e.g., Wechsler, 1997b). The NAB Orientation test includes 16 questions that measure orientation to self, time, place, and situation.

Task Creation

Because of the unique nature of orientation questions, this is the one NAB test for which the items are identical across the two forms and, within form, across the Screening Module and Attention Module. Seven questions about orientation to self (i.e., name, age, date of birth, street address, city, state, phone number), five questions about orientation to time (i.e., year, month, date, day of week, time), three questions pertaining to place (i.e., name of current location, city, state), and one question pertaining to situation (i.e., "Why are you here?") are included in the Orientation test.

Advisory Council Ratings and Equivalent Forms

All 16 items were rated by the Advisory Council for difficulty level, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, and overall task satisfaction. All items received excellent bias ratings as well as overall satisfaction ratings. There was minimal variability in difficulty ratings. On the basis of these ratings, all original items were retained in the test.

Screening Module

The identical Orientation test is used for both the Attention Module and the Screening Module.

Digits Forward

Background

The repetition of orally presented digits, frequently referred to as digit span or digit repetition, is another task included in most mental status examinations (e.g., Folstein et al., 2001), neuropsychological screening tests (e.g., Randolph, 1998), memory batteries (e.g., Wechsler, 1997b; Williams, 1991), and dementia examinations (e.g., Mattis, 2002), as well as intelligence tests (e.g., Wechsler, 1997a). This paradigm is the most common method of assessing auditory attentional capacity (also referred to as the span of immediate recall). The NAB Digits Forward test is based on the standard approach to digit repetition utilized in most existing tests.

Task Creation

The series of digits for this task were randomly generated with the random-number-generation function in Microsoft Excel™. These numbers were compiled into a master list to which exclusion criteria were then applied to eliminate specific types of number sequences. The exclusion criteria included (a) repeating numbers within a sequence, (b) zeros, and (c) more than two forward or reversed consecutive sequential numbers. The resulting series were then checked for sequences that occurred more than once, and these items were also eliminated. Additionally, any three-digit sequence (within any span length) that was the same as a telephone area code of a major U.S. city was eliminated. On the basis of these criteria, a total of six sequences of digits for each span length from 3 to 9 were created (i.e., a total of 42 sequences). Three sequences for each span length were included in each of the two forms.

Advisory Council Ratings and Equivalent Forms

The Advisory Council rated difficulty level and overall task satisfaction for each of the 42 span sequences. On the basis of these ratings, one sequence was eliminated at each span length for each form. That is, each form started with three trials per span length, with one of these three trials eliminated following the Advisory Council ratings. Sequences with the highest task satisfaction ratings were retained. Difficulty ratings were then used to assign items to Form 1 and Form 2. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Screening Module

The identical task with identical items is used in both the Attention Module and the Screening Module.

Digits Backward

Background

Tests requiring the examinee to reverse orally presented digits are almost always linked to digit-repetition tasks and are also included in most mental status examinations (Folstein et al., 2001), memory batteries (Wechsler, 1997b; Williams, 1991), dementia evaluations (Mattis, 2002), and intelligence tests (e.g., Wechsler, 1997a). In most existing tests, both digit-repetition and digit-reversal tasks are combined into one score. However, the two paradigms (digit repetition and digit reversal) most likely measure distinct functions (Kaplan, Fein, Morris, & Delis, 1991; Lezak, 1995), with digit reversal measuring both attentional capacity and working memory. The Digits Backward test in the NAB is a distinct test from Digits Forward, with completely distinct scores.

Task Creation

The same item-generation procedures were used for Digits Backward as were used for Digits Forward. That is, the series of digits for this task were randomly generated with the random-number-generation function in Microsoft Excel™. These numbers were compiled into a master list to which exclusion criteria were applied to eliminate specific types of number sequences. These exclusion criteria included (a) repeating numbers within a sequence, (b) zeros, and (c) more than two forward or reversed consecutive sequential numbers. The resulting series were then checked for sequences that occurred more than once, and these were also eliminated. Additionally, any three-digit sequence (within any span length) that was the same as a telephone area code of a major U.S. city was eliminated. On the basis of these criteria, a total of six sequences of digits for each span length from 3 to 9 were created (i.e., a total of 42 sequences). Three sequences for each span length were included in each of the two forms.

Advisory Council Ratings and Equivalent Forms

The Advisory Council rated the difficulty level and overall task satisfaction for each of the 42 span sequences. On the basis of these ratings, one sequence was eliminated at each span length for each form. Sequences with the highest task satisfaction ratings were retained. Difficulty ratings were then used to assign items to Form 1 and Form 2. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Screening Module

The identical task with identical items is used in both the Attention Module and the Screening Module.

Dots

Background

Delayed-recognition span tests have been used in both animal and human investigations of working memory (Chodosh, Reuben, Albert, & Seeman, 2002; Lacreuse, Herndon, Killiany, Rosene, & Moss, 1999; Martin, Pitrak, Pursell, Mullane, & Novak, 1995; Moss, Albert, Butters, & Payne, 1986; Moss, Killiany, Lai, Rosene, & Herndon, 1997). The spatial delayed-recognition span paradigm typically involves an array of dots that is exposed for a brief period, followed by a blank interference page, followed by a new array with one additional dot that the examinee is asked to recognize and point to. This measure of visual working memory and visual scanning has been found to be sensitive to a variety of human disorders, including dementia (Moss et al.), HIV infection (Martin et al.), and basal ganglia disorders (Partiot et al., 1996), among others. The NAB Dots test is based on the spatial delayed-recognition span test most commonly used in both experimental and clinical settings.

Task Creation

Three forms, each consisting of 15 items, were initially created. Each individual item consisted of three 8½ in. x 11 in. pages presented in landscape orientation: (a) an initial presentation page (page "A"), consisting of a spatial array of colored dots (ranging from 3 to 17 dots, each dot 7⁄16 in. in diameter); (b) a mask page, consisting of a 7 in. x 9½ in. solid rectangle printed in the same color as the dots on the previous page; and (c) a recognition page (page "B"), consisting of a spatial array of colored dots identical to the corresponding page "A" but with one additional dot (i.e., the target "new dot"). The placement of the dots on the pages was initially made by a computer program designed to create a pseudorandom array of dots (given input of a specific number of initial dots) with the four quadrants of the page equally represented. Once the 15 original items were created, two additional sets were created by modifying the original set. The first additional set was derived by rotating each of the original items 180 degrees along the horizontal axis; the second additional set was derived by rotating (mirroring) the original items 180 degrees along the vertical axis.

Advisory Council Ratings and Equivalent Forms

The 45 three-page items (15 items in each of the three sets) were rated by the Advisory Council members for difficulty level. In addition, each of the three sets was rated for overall task satisfaction and satisfaction with the stimulus. The two sets with the best satisfaction ratings also had the most similar mean difficulty ratings; these two sets were, therefore, retained. The two resulting 15-item (sample and 14-items) sets were then subjected to extensive pilot testing. On the basis of the results of pilot testing, the two largest array items (with 16 and 17 dots on the "A" page, respectively) for each set were deleted because of excessive difficulty. The 3-dot item in each set was then used as the sample item. The 4- to 7-dot items were included as Items 1 to 4, respectively, and are presented with a 5-second delay (i.e., 5-second duration of mask presentation). The 8- to 15-dot items were included as Items 5 to 12, respectively, and are presented with a 10-second delay. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Numbers & Letters

Background

Letter- and/or symbol-cancellation tasks (e.g., Diller et al., 1974; Mesulam, 2000) are commonly included in neuropsychological evaluations as measures of sustained attention, visual scanning, neglect, and psychomotor speed. In addition, cancellation tasks that use a controlled search paradigm (e.g., selecting a specific target among similar distractors) are frequently employed as measures of selective or focused attention (Ruff & Allen, 1996). Tests based on the Trailmaking Test paradigm (Reitan & Wolfson, 1993) are also commonly used in neuropsychological evaluations to measure psychomotor speed and divided attention. Although not typically assessed with traditional paper-and-pencil cancellation tasks, information processing speed (or mental processing speed) is also considered a component of the broad domain of attention (Williamson, Scott, & Adams, 1996). Because many of these important facets of attention are interrelated, four separate subtests were created for the NAB to measure sustained attention, psychomotor speed, selective attention, divided attention, information processing speed, impulsivity, and disinhibition. The subtests have a similar format, but each has different task demands and requirements. The goal was to provide the examinee with similar stimuli in each subtest but to increase the complexity of task demands in a manner that measured these seven components of attention.

Task Creation

The Numbers & Letters test includes four separate subtests, each created with a similar underlying structure and format. The first subtest, Part A, is a letter-cancellation task that requires the examinee to mark a slash through target X's embedded in 24 rows of numbers and letters. Part A was designed to measure sustained attention, selective attention, and psychomotor speed. Each row has a total of 40 numbers and letters, with 9 or 10 X's embedded as the target cancellation letter. In addition to the targets, each row also contains approximately 10 numbers (including only 1, 2, and 3) and approximately 30 letters (excluding A, I, O, and Z). The target X's were placed in each row in a pseudorandom fashion, with 4 or 5 X's appearing in the left half of the line and 4 or 5 X's appearing in the right half of the line. Each row had no more than two instances in which two X's appeared sequentially, and instances in which three or more X's appeared next to each other were not permitted.

Part B was designed to measure selective attention and information processing speed. Part B follows a similar format to Part A, although Part B requires the examinee to count the number of X's in each row (without marking a slash through the X's) and to write the total in a space provided at the end of each row. Part B has 8 rows of 40 numbers and letters, and each row has 9 to 12 target X's and 9 to 10 numbers.

Part C is similar to Part B, although it has greater information processing demands. Part C involves the examinee *adding* the numbers in each row (again, without marking a slash through the X's) and writing the sum in a space provided at the end of each row. In order to limit the dependency on calculation skills for successful performance, only the numbers 1, 2, and 3 were used. Part C has 8 rows, and each row has 9 to 10 X's and 10 numbers.

The final subtest, Part D, adds complexity to the task demands in that it requires the examinee to mark a slash through the X's *and* to simultaneously add the numbers and write the sum at the end of each row. Part D was designed to measure selective attention, divided attention, and psychomotor speed. Part D has 4 rows, and each row has 10 to 12 X's and 10 numbers.

Advisory Council Ratings and Equivalent Forms

Three parallel forms of each of the four Numbers & Letters subtests were created initially. Identical design rules were used for each form. The Advisory Council rated each subtest of each of three forms for difficulty level and overall task satisfaction. These ratings were used to eliminate one form of each subtest, so that two forms of each subtest with highly similar difficulty level and overall satisfaction ratings were retained. The results of pilot testing were used to empirically determine the difficulty level of each task and to equate tasks across the two equivalent forms.

Screening Module

The Screening Module includes similar, abbreviated versions of Parts A and D of the Attention Module Numbers & Letters Test. Screening N&L Part A is similar to Attention N&L Part A, although it includes only 4 rows of numbers and letters and the target cancellation letter is A instead of X. Each row includes 10 target A's, 10 numbers, and 20 additional letters. Screening N&L Part B is similar to Attention N&L Part D, although it includes only 2 rows of numbers and letters and the target cancellation letter is also A instead of X. The same procedures for Advisory Council ratings and equivalent forms creation were employed.

Driving Scenes

Background

For the Daily Living test of the Attention Module, the goal was to create a multifactorial measure that taps several key aspects of attention and that is both face valid and likely to be related to everyday living. Existing measures of attention in everyday life typically include several different and lengthy tasks to meet these goals (e.g., Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994). One specific approach to examining attention as it relates to everyday life involves the "useful field of view" (UFOV) paradigm (Ball, Beard, Roenker, Miller, & Griggs, 1988; Ball & Roenker, 1998), a computerized method of assessing visual attention that has been found to predict motor vehicle crashes in the elderly (Ball, Owsley, Sloane, Roenker, & Bruni, 1993). Aspects of the UFOV paradigm were incorporated into the NAB Driving Scenes test, although the task was designed to be administered without a computer and to measure several different aspects of visual attention, including working memory, visual scanning, attention to detail, and selective attention.

Task Creation

Artwork for this task was initially created by pen-and-ink hand drawings. These drawings were then scanned into a computer and altered (smoothed, colorized, shaded) by a graphic artist. One original form was first created, with a base stimulus depicting a driving scene on a two-lane road in a small town business area, as viewed from behind the steering wheel of a car, along with five additional scenes built on the base scene, but with specific modifications (additions, changes, and subtractions of details from scene to scene). This initial series of scenes was then pilot tested, and changes were made. Once the original form was finalized, two additional forms (each with a base scene and five subsequent sequential scenes) were created. For each form, specific criteria were followed for the design of each base scene (Scene 1) and for changes in subsequent scenes, including (a) approximately equal numbers of stimuli in both sides of each scene; (b) similar numbers of "dangerous" items (e.g., vehicles approaching, people crossing), dashboard items (e.g., changes in speedometer, fuel gauge), and minor/detail items (e.g., birds flying, kite in sky) across the three forms; and (c) approximately equal total number of new and missing items in each scene across the three forms.

Advisory Council Ratings and Equivalent Forms

Each of the 18 scenes (6 scenes for each of three forms) were rated by the Advisory Council members for difficulty level, sex bias, ethnic/racial/cultural bias, quality of artwork, and overall task satisfaction. In addition, each of the forms was rated on an overall basis on these same characteristics. All scenes were rated as having minimal sex or ethnic/racial/cultural biases. All scenes across the three forms received similar difficulty ratings. However, one form consistently received lower satisfaction ratings than the other two and was, therefore, eliminated.

The results of pilot testing were used to empirically determine the difficulty level of each set of scenes and to equate panels across the two equivalent forms.

LANGUAGE MODULE

Oral Production

Background

Oral production (i.e., speech output, or verbal fluency) is a key element in the assessment of aphasia with regard to both differential diagnosis and recommendations for therapeutic interventions. Most tests of verbal fluency (e.g., Benton, Hamsher et al., 1994), however, involve the examinee's generating words that begin with a specific letter or from a specific semantic category (e.g., animals). Although performance on these tasks is diminished in patients with nonfluent aphasia, it is also frequently impaired in patients with executive dysfunction without aphasia (Boone, 1999; Mitsrushina et al., 1998). In fact, the term "fluency" is possibly misleading for these word-generation tasks (Marshall, 1986). Assessment of *propositional speech output* is a more appropriate method of examining fluency in the aphasic patient. Unfortunately, this important aspect of assessment is frequently based on subjective or qualitative observation (e.g., having the examinee orally describe what is happening in the Cookie Theft picture of the Boston Diagnostic Aphasia Examination [BDAE], Goodglass et al., 2000), rather than on a more objective, quantified approach. Yorkston and Beukelman (1980) developed a system for measuring the content units and content units per minute for oral descriptions of the Cookie Theft picture. Nicholas and Brookshire (1995) described another approach to measuring the main concepts in speech during storytelling. This content-unit approach to measuring speech output was incorporated into the NAB Oral Production test.

Task Creation

Artwork for this test was initially created by pen-and-ink hand drawings. These drawings were then scanned into a computer and altered (smoothed, colorized, shaded) by a graphic artist. Three similar family scene drawings were created. Each of these three forms was designed to be equated for the number of possible content units, as well as the type of information included (e.g., an element of danger; parents being unaware of the potential danger; different types of food present; balance of background details with foreground details; several potential nouns, adjectives, and verbs appropriate to describe the scene).

Advisory Council Ratings and Equivalent Forms

The three forms (i.e., three family scene drawings) were rated by the Advisory Council members for difficulty level, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, and overall task satisfaction. One of the three forms received uniformly high U.S. regional bias, educational bias, and ethnic/racial/cultural bias ratings, as well as poorer overall task satisfaction; therefore, that form was eliminated. The remaining two forms had identical difficulty ratings. The results of pilot testing were used to empirically determine the difficulty level of each scene and to equate each scene across the two equivalent forms.

Auditory Comprehension

Background

Auditory comprehension is another important component in the assessment of language impairment and aphasia. There are several methods of examining auditory comprehension, with the most common approaches involving the examiner's giving oral commands of increasing complexity to the examinee, who then responds by manipulating objects placed in front of him/her (e.g., Boller & Vignolo, 1966; DeRenzi & Vignolo, 1962; Benton, Hamsher, et al., 1994). Other methods of assessing auditory comprehension include pointing commands (e.g., to body parts) and yes/no questions (e.g., Goodglass et al., 2000). The NAB Auditory Comprehension test was designed to be a comprehensive assessment of auditory comprehension, incorporating most of the previous existing methods of assessment, including asking the examinee to perform various one- to four-step commands; questions pertaining to the concepts of before/after, above/below, and right/left; body-part identification; and yes/no questions.

Task Creation

Six separate subtests were created to measure Auditory Comprehension. They require the examinee to listen to orally presented commands and to respond by pointing to stimuli such as (a) colored rectangles (Colors, 7 items), (b) geometric shapes (Shapes, 12 items), and (c) colored geometric shapes with numbers printed on them (Colors/Shapes/Numbers, 13 items); or (d) by pointing to body parts or places in the room (Pointing; 6 items); (e) by answering orally presented pairs of yes/no questions (Yes/No Questions, 5 pairs of items); and (f) by folding paper according to one- to four-step commands (Paper Folding, 6 items). For all six subtests, three forms were initially created, and the Advisory Council ratings were used to eliminate one entire form.

Colors. The stimuli for this pointing task are four colored (red, black, yellow, and blue) rectangles measuring 1½ in. x 3½ in., arranged vertically on a stimulus book page. The three original forms had identical stimuli except that the order of the colors on the page was varied. Of the seven items, the first four consist of one-step commands (e.g., "point to blue"), with one item per color. The last three consist of two-step commands (e.g., "point to yellow and then to blue"). The commands for the different forms were similar, except for the order of the colors.

Shapes. The stimuli for this second pointing task are three solid geometric shapes (circle, square, triangle), all printed in filled black ink, measuring 2 in. across, arranged vertically on a stimulus book page. The three original forms had identical stimuli, except that the order of the shapes on the page was varied. Of the 12 items, the first 3 consist of one-step commands (e.g., "point to the square"), with one item per shape. Two items require additional comprehension of numbers (e.g.,"…three sides"). Two items require relational understanding (e.g., "…below the square"). One item requires comprehension of the concept of between. The final 5 items require comprehension of ordering/sequencing (e.g., "and then," "after," "before"). The commands for the different forms were similar except for the order of the shapes.

Colors/Shapes/Numbers. The stimuli for this third related pointing task are six geometric shapes (two triangles, two squares, and two circles), with one of each pair of shapes printed in red and the other in blue. Each shape also has a number printed on it, with three two-digit numbers and three single-digit numbers. The three forms had similar stimuli except that the order of the shapes on the page was varied and the numbers on the shapes were different. Of the 13 items, the first 3 consist of one-step commands involving numbers (e.g., "point to the number 8"). Three items require comprehension of both color and shape together (e.g.,"…blue triangle"). Three items require relational understanding as well as comprehension of shape names (e.g., "…triangle that is below the square"). The final 4 items require a variety of additional comprehension skills (e.g., "not," number facts). The commands for the three original forms were similar, although with different orders of shapes, colors, and numbers and with different number concept items.

Pointing. This subtest requires the examinee to point to three parts of the room and three body parts. The three parts of the room were identical across the three original forms. However, the three body parts were different between forms, although they represented similar locations on the body.

Yes/No Questions. Five pairs of yes/no questions were created for each of the three original forms. The questions were paired in order to control for chance responding, given the 50% chance of responding correctly to individual questions. The correct combinations and ordering of responses (i.e., "yes" vs. "no") for each pair were constant across the three forms. Each form included a pair of questions about clothing, a pair about time, a pair about familial relationship and age, a pair about eating utensils, and a pair about compass directions.

Paper Folding. The final Auditory Comprehension subtest consists of six commands of increasing complexity and involving the folding and other manipulation of a piece of paper. The three original forms were very similar, with the exception that the folding paper had different marks on its vertical and horizontal bisectors on either side of the page. The first two items involve one-step commands. The third and fourth items involve two-step commands and the ordering of the responses. The fifth item involves a three-step command and the ordering of the responses. The sixth item involves a four-step command and ordering of the responses.

Advisory Council Ratings and Equivalent Forms

As just described, three forms were created for all six subtests. Each of the items on the first three subtests (Colors, Shapes, Colors/Shapes/Numbers) was rated by the Advisory Council members for difficulty level, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, and overall task satisfaction. In addition, each of the three stimulus sets for each of the three forms was rated for satisfaction with the stimuli. The individual items for each of the three forms of the Pointing subtest were rated for difficulty level, task appropriateness, and overall task satisfaction. All of the items for the three forms of the Yes/No Questions subtests were rated for difficulty level, linguistic demands, sex bias, U.S. regional bias, task appropriateness, and overall task satisfaction. The Paper Folding items were rated for difficulty level and overall task satisfaction, as well as for overall satisfaction with the stimuli for the folding sheets.

The Advisory Council ratings were used to eliminate one form for each of the six subtests. All bias ratings were adequate for each of the three forms. Within each subtest, the one form with the lowest satisfaction ratings was eliminated. The remaining two forms of each subtest have very similar difficulty ratings. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Screening Module

The identical Colors, Shapes, and Colors/Shapes/Numbers subtests, with identical items, are used in both the Language Module and the Screening Module.

Naming

Background

Almost all patients with aphasia, regardless of specific syndrome, have some difficulty with word-finding, or naming (Benson & Ardilla, 1996; Goodglass & Wingfield, 1997). The most common method of measuring word-finding is through visual confrontation naming, in which the examinee is asked to state the name of an object depicted in a drawing and then is provided with semantic and phonemic cues, if necessary. Existing confrontation naming instruments, such as the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983), however, are greatly influenced by the examinee's educational level (e.g., Hawkins et al., 1993; Welch, Doineau, Johnson, & King, 1996). Therefore, for the NAB Naming test, care was taken to combine sensitivity to naming deficits with a lack of influence from educational attainment.

Task Creation

Stimuli included in the Naming test were culled from thousands of digital stock photography images of common objects. Photographs were eliminated if they depicted an object that could be referred to by more than one word or required a compound word. A photograph was also eliminated if it (a) was not solely of the target object (i.e., nothing else could appear in the photograph to distract from the target object to be named), (b) was not a prototypical representation of the target object, or (c) appeared "dated" in any fashion. This selection process resulted in a total of 84 potential photographs. For each of the 84 items, a semantic cue and a phonemic cue were created. Because confrontation naming is more likely to be associated with the frequency of usage in *spoken* language, rather than with written language, traditional word frequency ratings could not be used (e.g., Zeno, Ivens, Millard, & Duvvuri, 1995). Therefore, Advisory Council ratings of spoken language usage were included as an estimate of word frequency.

Advisory Council Ratings and Equivalent Forms

Each of the 84 items was rated by the Advisory Council for frequency of usage in spoken English, sex bias, educational bias, U.S. regional bias, ethnic/racial/cultural bias, satisfaction with the semantic cue, satisfaction with the phonemic cue, and overall task satisfaction. Any item with unacceptable biases was eliminated first. Next, items with low overall satisfaction were eliminated. Attempts were then made to create pairs of items for the two forms, equated for word usage and type of object (e.g., fruit, animal). On the basis of this process, 62 items were retained, with 31 items for each of the two Naming forms. Finally, Advisory Council ratings of semantic and phonemic cues were examined, and any cues with unacceptable ratings were revised according to specific recommendations of the Advisory Council members. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Screening Module

Of the original 84 items just described, 20 were retained for the Screening Module, with 10 items for each of the two forms. Ten items with relatively high usage frequency ratings and 10 with relatively low frequency ratings were included. The identical process described for the Language Module was used in the Screening Module for item retention and selection of items for specific forms.

Reading Comprehension

Background

Assessment of reading in patients with aphasia typically involves the patient's matching written words with objects and comprehending sentences (Benson & Ardilla, 1996). Additional, more in-depth reading assessment may require the examinee to read letters, syllables, logotomes, and paragraphs. However, as part of the neuropsychological evaluation, the NAB test would focus on single-word and sentence reading by requiring the examinee to select a single target word from a group of foil words to match a photograph of an object and to select a single target sentence from a group of foil sentences to match a photograph scene that shows people interacting.

Task Creation

Stimuli to be included in the Reading Comprehension Words subtest were culled from hundreds of digital stock photography images of common objects. Photographs were eliminated if they depicted an object that (a) could be referred to by more than one word, (b) required a compound word, or (c) represented an atypical spelling. This selection process resulted in a total of 30 individual photographs. Under each photograph, the target word was printed, along with five foils; the order of the words was pseudorandomly assigned. For each item, the same criteria for creating the five foils were followed: (a) a word within the same semantic category as the target (e.g., "wolf" for the target "bear"); (b) a word that has the same last three letters as the target (e.g., "pear" for the target "bear"); (c) a word that is in the same semantic category as a foil (e.g., "mango" for the foil "pear"); (d) a word that begins with the same first letter as the target (e.g., "ball" for the target "bear"); and (e) a word that is unrelated to the target or other foils (e.g., "window" for the target "bear"). Word length was kept similar across the target and foils for each item.

A similar process was used in the creation of the Reading Comprehension Sentences subtest. Photographs were culled from hundreds of available digital stock photography images that depicted scenes of at least one person interacting with another person, with other people, or with an animal, and engaged in a clearly defined activity. This selection process resulted in a total of 30 photographs. Target sentences were created according to specific criteria, including (a) written in the present tense; (b) beginning with the principal subject, followed by a verb, followed by the secondary subject; and (c) written at the 8th-grade reading level or lower, according to the Flesch-Kincaid reading formula (Flesch, 1994). Three foil sentences were also created for each item, again following a consistent set of criteria: (a) reversal of subjects; (b) one part, but not all parts, of the sentence are accurate; and (c) not descriptive of the stimulus, but not a nonsense sentence. The order of the target and foil sentences in each item was pseudorandomly assigned.

Advisory Council Ratings and Equivalent Forms

Each of the 30 Reading Comprehension Words items was rated by the Advisory Council for reading difficulty, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, satisfaction with the photograph, and overall task satisfaction. Items with unacceptable biases were eliminated first. Additional items with low photograph and/or overall satisfaction ratings were then eliminated. Finally, items were retained to provide a spectrum of difficulty levels.

This process resulted in 12 items, 6 for each form, equated for reading difficulty across forms.

A similar item-reduction and form-equivalence process was used for Reading Comprehension Sentences. Each of the 30 target sentences was rated by the Advisory Council for reading difficulty, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, task appropriateness, and overall task satisfaction. Similarly, the group of three foils for each item was rated on these same variables. Finally, each photograph was rated for sex bias, U.S. regional bias, ethnic/racial/cultural bias, task appropriateness, and overall task satisfaction. All items with unacceptable biases were eliminated first. Additional items with low task appropriateness and/or overall satisfaction ratings were then eliminated. Finally, items were retained to provide a spectrum of difficulty levels. This process resulted in 14 items, 7 for each form, equated for difficulty across forms. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Writing

Background

Writing is an important aspect of the assessment of language, because most individuals with aphasia have some difficulty with writing, or agraphia (Benson & Ardilla, 1996). Writing is frequently measured in mental status exams or brief/screening neuropsychological tests by having the examinee write a single sentence. However, this approach does not fully sample the variety of writing difficulties that may be tapped by a longer narrative writing task. One of the most common narrative writing tasks employed in neuropsychological and speech/language evaluations is the Cookie Theft picture of the BDAE (Goodglass et al., 2000). However, the assessment of the examinee's writing sample is typically based on a qualitative observation of the sample, rather than on a quantitative measurement of specific aspects of the written production. Therefore, the NAB Writing test was created to allow for quantification of several major features of narrative writing, including legibility, syntax, spelling, and conveyance of the major themes depicted in the stimulus picture.

Task Creation

The identical stimulus is used for the Writing and Oral Production tests. As described in the Oral Production section earlier in this chapter, the artwork for this test was initially created by pen-and-ink hand drawings. These drawings were then scanned into a computer and altered (smoothed, colorized, shaded) by a graphic artist. Three similar family scene drawings were created. Each of these three drawings was designed to be equated for the number of possible content units, as well as for the type of information depicted (e.g., an element of danger; parents being unaware of the potential danger; different types of food present; balance of background details with foreground details; several potential nouns, adjectives, and verbs appropriate to describe the scene).

The scoring system for the narrative writing samples was designed to be relatively simple and reliable to complete, yet to provide sensitive markers of the major aspects of writing: legibility, syntax, spelling, and conveyance of the major themes depicted in the stimulus. Because of the similarity between the different forms (i.e., stimuli), the scoring criteria were identical for each form. A 0- to 2-point scale is used for legibility, whereas a 0- to 3-point scale is used for the other three scores. Each score has specific anchor points to facilitate scoring and to improve interrater reliability.

Advisory Council Ratings and Equivalent Forms

The three family scene drawings were rated by the Advisory Council members for difficulty, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, and overall task satisfaction. One of the three forms received uniformly higher U.S. regional bias, educational bias, and ethnic/racial/cultural bias ratings, as well as poorer overall task satisfaction ratings; therefore, that form was eliminated. The remaining two forms had identical difficulty ratings. The results of pilot testing were used to empirically determine the difficulty level of each scene and to equate the scenes across the two equivalent forms.

Bill Payment

Background

The adequate use of language and communication in everyday living is an important aspect of functional independence for many individuals. Some patients with aphasia may exhibit significant deficits in circumscribed areas of language functioning in formal office-based or bedside testing but may still be able to communicate effectively in the "real world." Other patients may exhibit only mild language difficulties upon formal testing, although they are not able to perform more complicated tasks in everyday life. Several tests and test batteries have been developed to address this issue (e.g., Frattali, Thompson, Holland, Wohl, & Ferketic, 1995; Holland, 1980; Holland, Frattali, & Fromm, 1999). The NAB Language Module Daily Living test was created to provide a real world situation (i.e., household utility bill payment) that requires intact functioning in many areas of language functioning, including auditory comprehension, reading comprehension, writing, simple calculations, and speech output.

Task Creation

The Bill Payment test was designed to be relevant to the language demands of everyday functioning. It includes four stimuli that are facsimiles of a household utility bill statement, a blank check, a check ledger, and an envelope. Eight items involve five questions and three commands that involve these stimuli. Three forms of the task were created initially, each using a different household utility (i.e., telephone, cable television, and electricity). Identical procedures were used to create the content of the stimuli across forms (e.g., number of words in the company name, number of digits in the account number), and the stimuli for each form were presented identically. The five questions were created to require different types of oral responses (including yes/no, numerical only, word only, and combination of word and numerical), along with reading comprehension and number comparisons (including date, dollar and cents, and words). The three commands were each designed to require multiple-step, written responses that involve words, numbers, and simple calculations. The questions and commands for each of the three forms were identical (with the exception of the utility type), although the responses were different, based on the unique content information provided in each form.

Advisory Council Ratings and Equivalent Forms

The three forms of the items were rated by the Advisory Council for difficulty level, sex bias, U.S. regional bias, educational bias, ethnic/racial/cultural bias, linguistic demands, task appropriateness, and overall task satisfaction. In addition, the three forms of stimuli were rated for overall satisfaction with the stimulus. The ratings across all three forms were nearly identical for all questions, commands, and stimuli. However, the form with the cable television bill received slightly higher overall educational bias ratings. Therefore, the telephone bill and electric bill forms were retained. The results of pilot testing were used to empirically determine the difficulty level of each item, and to equate items across the two equivalent forms.

MEMORY MODULE

Decisions about tests to include in the Memory Module were based on several key factors. First, the various types of learning and memory tests and procedures available at the time of initial NAB development were included in the publisher's survey (Stern & White, 2000). As shown in Table 2.3, the survey results indicated the respondents' desire to include measures of word-list learning and immediate recall, prose/paragraph immediate recall, verbal delayed recall, verbal recognition memory, visual/nonmotor learning and immediate recall, visual/nonmotor delayed recall, and visual/nonmotor recognition memory. On the basis of these findings and a review of the literature on learning and memory assessment (e.g., Cermak, 1994; Lezak, 1995; Squire & Butters, 1992; Tulving & Craik, 2000), it was decided that the NAB Memory Module would include (a) a list-learning test (List Learning), with immediate free recall, delayed free recall, and delayed forced-choice recognition trials; (b) a story-learning test (Story Learning), with two learning trials, separate measures of verbatim and gist recall, and immediate free recall and delayed free recall trials; (c) a visual/nonmotor learning test (Shape Learning), with immediate multiple-choice recognition, delayed multiple-choice recognition, and delayed forced-choice recognition trials; and (d) a Daily Living Memory test with information likely to be encountered in everyday life, such as medication instructions and a person's name, address, and phone number. The Daily Living Memory test involves immediate free recall, delayed free recall, and delayed multiple-choice recognition trials.

The decision to include yes/no forced-choice recognition trials in the List Learning and Shape Learning tests was based on research findings that the response biases and discriminability measures resulting from this paradigm can provide important information about various patient groups with memory impairments (e.g., Snodgrass & Corwin, 1988).

The decision to include both list-learning and story-(prose) learning measures in the NAB was further supported by research findings that suggest differential performance on these two paradigms by different patient groups. For example, patients with impaired executive functioning have been found to perform worse on list-learning tasks than on logically organized story-learning tests (e.g., Tremont, Halpert, Javorsky, & Stern, 2000).

Decisions about the length of delay intervals were based on (a) existing research literature that indicates little, if any, difference in recall performance between delays ranging from 10 to 60 minutes (Berry & Carpenter, 1992; Chapman, White, & Storandt, 1997; Somerville & Stern, 2001); (b) research findings that suggest that relatively brief delay intervals (2 to 10 minutes) are best at differentiating patients with Alzheimer's dementia from other patient groups and from control respondents (see Albert & Moss, 1992); and (c) the "flow," order, and maximum administration time of the Memory Module and Screening Module. Therefore, the delay intervals for the List Learning, Story Learning, and Shape Learning tests are 15 minutes, and the delay intervals for the two subtests of the Daily Living Memory test and the two Screening Module memory tests are 5 to 10 minutes.

List Learning

Background

Verbal list-learning tests are an important component of the assessment of memory. These tasks allow for measures of learning curve (i.e., recall improvement with repetition trials), sensitivity to interference, the use of semantic encoding strategies, intrusion, perseveration, and differences between free recall and forced-choice recognition. Existing word-list-learning tests (e.g., Brandt & Benedict, 2001; Delis, Kramer, Kaplan, & Ober, 2000; Williams, 1991) are commonly used in clinical practice as well as in research settings because of the rich quantity and quality of data they provide. In addition to providing important information about differential diagnosis (e.g., Brandt, Corwin, & Krafft, 1992; Crossen, Sartor, Jenny, Nabors, & Moberg, 1993; Curtiss, Vanderploeg, Spencer, & Salazar, 2001; Massman, Delis, Butters, Dupont, & Gillin, 1992), list-learning tests have also been found to be significantly correlated with measures of central nervous system change in various diseases (e.g., Killiany et al., 2002).

In the design of the NAB List Learning test, three primary goals were followed: (a) to create a three-trial learning test to avoid the potential difficulties that five-trial tasks represent for impaired individuals; (b) to include three semantic categories to allow for examination of the use of semantic clustering as a learning strategy; (c) to avoid sex, education, and other potential biases; and (d) to include both free recall and forced-choice recognition paradigms.

Task Creation

The List Learning test involves three learning trials of a 12-word list, followed by an interference list, and then by short-delay free recall, long-delay free recall, and long-delay forced-choice recognition tasks. The word list includes three embedded semantic categories with four words in each category. This approach is common to existing list-learning paradigms (e.g., Brandt, 1991; Delis, Kramer, Kaplan, & Ober, 1987).

The following criteria were used to develop the items for the word lists. All words had a middle-range frequency of occurrence in the English language as defined by a Standard Frequency Index (SFI) range of 42.0 to 53.0 (Zeno et al., 1995). No compound words (e.g., babysitter) were used, and only single words with less than four syllables were included. The words were concrete nouns that clearly fell within a distinct category; within each category, words with distinct phonemes were used. The words were intended to be unbiased with respect to sex, ethnic, religious, occupational, and regional characteristics. Finally, words were not included if they were likely to have a personal (e.g., medical) meaning for the examinees.

The final two forms (after word elimination) have the following format: List A consists of 12 words with three embedded semantic categories (A, B, C) of 4 words each. List B consists of 12 words with three embedded semantic categories (A, D, E). One of the categories (A) is the same as one of the categories from List A; the remaining two categories (D, E) are new. The 36-word recognition list includes the following: all 12 words from List A; all 12 words from List B; 2 distractors from Category B of List A; 2 distractors from Category D of List B; and 2 distractors from each of four additional, new categories (F, G, H, I).

There were four forms initially created, each with 150% of the total number of words that were needed. For example, although only 8 words are included in the final form for Category A, 12 words were initially included. Therefore, the Advisory Council ratings facilitated the selection of the best words from each category for each of the four forms, as well as the selection of the two overall best forms.

Advisory Council Ratings and Equivalent Forms

Each word from each of the four initial forms was rated by the Advisory Council members for difficulty level, ethnic/racial/cultural bias, sex bias, and overall task satisfaction. Within each form, those words with the highest biases and lowest overall task satisfaction were eliminated first. Then, those words with high difficulty ratings were eliminated. This process resulted in the final pool of words for each form. Once the four forms were finalized, the two forms with the highest mean overall satisfaction ratings were retained. The results of pilot testing were used to empirically determine the difficulty level of each word and list, and to equate words and lists across the two equivalent forms.

Shape Learning

Background

Visual memory has historically been tested with a variety of paradigms, the most common of which involves the drawing and recall of geometric shapes and figures (e.g., Meyers & Meyers, 1995; Sivan, 1992; Wechsler, 1997b). These tasks, by definition, depend on intact graphomotor functioning and visuoconstruction skills. Several visual memory tests use a motor-free recognition response format in order to overcome these potential confounds, although some of these tests involve stimuli that are pictures of common objects (e.g., Hannay, Levin, & Grossman, 1979) and are readily encoded through verbal mediation. Several visual recognition tasks involve geometric shapes (e.g., Williams, 1991) or more abstract or nonsense stimuli (Benedict, 1997; Trahan & Larrabee, 1988), but these tests are also not free of potential verbal mediation (Lezak, 1995). In the design of the NAB

Shape Learning test, four primary goals were followed: (a) to create a motor-free, visual recognition learning task; (b) to utilize a three-trial learning paradigm to mirror the List Learning test; (c) to create stimuli that are very difficult to encode verbally; and (d) to make the stimuli visually pleasing and "acceptable" to examinees in spite of the absence of verbal mediation.

Task Creation

The Shape Learning test involves three learning trials of nine target stimuli. Each learning trial is composed of an initial presentation of the nine targets (one at a time), followed by nine 4-stimulus, multiple-choice recognition items (each including a target and three related foils). After a 15-minute delay, there is another multiple-choice recognition trial, followed by an 18-item forced-choice, yes/no recognition trial composed of the 9 original stimuli and 9 foils. The stimuli consist of computer-generated "swirls," "swatches," and "blobs." Foils for each target were created by manipulating specific parameters of each graphic (e.g., for swirls, increasing or decreasing the degree of rotation or changing the direction of the rotation). Although the stimuli are printed in color, color is not part of the to-be-remembered information. Three of the nine items involve only one stimulus (one swatch or one swirl or one blob). Three items contain two stimuli each, and three items contain three stimuli each (one swatch and one swirl and one blob). The location of the stimuli on the card is a component of the to-be-remembered information for the three-stimulus items. Three complete forms were initially created, each including 9 targets, 9 sets of three foils, and 18 forced-choice recognition foils.

Advisory Council Ratings and Equivalent Forms

The Advisory Council rated each item on each form for difficulty level, verbal encodability, satisfaction with the design, and overall task satisfaction. None of the items received low task satisfaction ratings or high verbal encodability ratings, and all three forms had very similar mean overall difficulty levels. The one form with the lowest overall task satisfaction ratings was eliminated. This process resulted in the final two forms. The results of pilot testing were used to empirically determine the difficulty level of each item and set of items, as well as to equate the items and sets of items across the two equivalent forms.

Screening Module

The Screening Shape Learning test was designed to be similar to the Shape Learning test in the main Memory Module but differs along the following dimensions; (a) it uses a different type of stimuli, (b) it has only five items, and (c) it involves only one learning trial, followed by an immediate recognition trial and a delayed recognition trial. The examinee is presented with a series of five geometric designs, presented individually. After stimulus presentation, the examinee is shown a series of five cards, each of which contains one target stimulus along with three foils; the examinee is asked to recognize which design was previously seen. The principles that guided the development of these stimuli were similar to those described previously for Shape Learning in the main Memory Module. The first goal was to develop computer-generated stimuli that could be manipulated in a systematic manner to produce foils and alternate forms. Second, the stimuli were created to be as resistant as possible to verbal encoding yet still be engaging and enjoyable for the examinee.

Following the work of Benedict (1997), D'Elia, Satz, Uchiyama, & White (1996), and Vanderplas and Garvin (1957), the initial target stimuli were based on randomly generated polygons with eight points. On the basis of these principles, 80 initial stimuli were created during the early stages of task development. A group of five clinicians and technicians rated these figures for ease of verbal encoding; those items with easier verbal encoding were deleted. Three forms were then initially developed. Each form consisted of five target stimuli and five recognition pages, each with a target stimulus and three foils. The foils were all created by altering the target in varying degrees, such that one foil was very similar to the target, one was moderately similar, and one was only slightly similar. The targets and foils across forms were rotations or mirror images of the stimuli on the initial form. Each of the targets and associated recognition pages (each including a target and three foils) for each of the three original forms was then rated by the Advisory Council for difficulty level, verbal encodability, satisfaction with the target, satisfaction with the foils, and overall task satisfaction. On the basis of these ratings, the two forms with the best overall satisfaction, most similar difficulty levels, and most similar verbal encodability ratings were retained for the final two forms.

Story Learning

Background

Story learning and recall tasks are included in most memory assessment batteries (e.g., Denman, 1987; Randt & Brown, 1986; Tombaugh & Schmidt, 1992; Wechsler, 1997b; Williams, 1991) and flexible neuropsychological evaluations (Lezak, 1995). These tasks have been found to be very sensitive to early memory impairment and to discriminate significantly among patient groups (e.g., Locascio, Growdon, & Corkin, 1995; Morris et al., 2001; Wefel, Hoyt, & Massman, 1999). The NAB Memory Module Story Learning test was designed to incorporate several important features of existing memory tests, including two learning trials, separate measures of verbatim recall (i.e., phrase unit recall), gist recall (i.e., thematic unit recall), and both immediate and delayed recall trials.

Task Creation

Thirty unique stories were initially written according to the following criteria. The stories had 20 memory elements comprising five sentences of four phrase units each. No story had more than 65 words. The stories were written at or below the 6th-grade Flesch-Kincaid Reading Level and at or above a Flesch Reading Ease Index of 65.0 (Flesch, 1994). All stories were written in past tense and used active voice. There was no repetition of key phrase unit elements within a story. The content of the stories involved various actions of people, and the stories were written to have a slightly emotional valence. The 30 initial stories were reviewed by the development team who selected the "best" eight.

Advisory Council Ratings and Equivalent Forms

Each of the eight original stories was rated by the Advisory Council for difficulty level, sex bias, ethnic/racial/cultural bias, "other" bias, face validity, and overall task satisfaction. The four stories with the best overall task satisfaction were initially retained. From these four, the two with the most equivalent difficulty ratings and lowest bias ratings were selected for the final two forms.

Screening Module

The Screening Story Learning task was designed to be a single-trial prose-learning task that had an immediate free recall trial and a delayed free recall trial. Development was nearly identical to that of the Story Learning test in the main Memory Module. Twenty-five unique stories were initially written according to the following criteria. The stories each contained two sentences with a total of 12 phrase units each (thematic units are not scored for Screening Story Learning). There was no repetition of phrase units within a story. All stories were written at or below the 6th-grade Flesch-Kincaid Reading Level and at or above a Flesch Reading Ease Index of 65.0 (Flesch, 1994). The stories involved various actions of people and had a slightly emotional valence. The development team reviewed the 25 stories and selected the "best" eight. Following the same Advisory Council rating procedures used for the Memory Module Story Learning test, two of the eight stories were retained for the final two forms.

Daily Living Memory

Background

Traditional memory tests involving learning and recall of word lists, stories, or figures provide important information about the specific strengths and weakness of various learning and memory functions. However, there can frequently be a dissociation between an individual's performance on formal memory measures and real-world memory functioning. The need for greater ecological validity in memory testing led to the development of the Rivermead Behavioural Memory Test (RBMT; Wilson, Cockburn, & Baddeley, 1985), a battery of learning and recall tasks related to everyday functioning that has been found to be a better predictor of everyday memory than traditional memory tests (e.g., Makatura, Lam, Leahy, Castillo, & Kalpakjian, 1999). The NAB Daily Living Memory test focuses on the ecological validity of the to-be-remembered information. That is, in contrast to the more novel or unfamiliar tasks of List Learning, Shape Learning, and Story Learning, this test was designed to employ information that people are often required to remember in their everyday lives: a name, address, and phone number or medication dosing instructions. The latter was selected because of the direct implications this information can have for the safety of prescribed medical treatment (Haynes, McDonald, & Garge, 2002).

Task Creation

Eight forms of both components of the Daily Living Memory test (i.e., Medication Instructions and Name, Address, and Phone Number) were initially created. All Medication Instructions forms were developed according to explicit criteria for the two target sentences and the three delayed recognition foils for each of the two sentences (e.g., modifications to the number, the color, and the type [pill vs. capsule] of medication). Similarly, all eight forms of Name, Address, and Phone Number were developed according to a set of very explicit criteria for the target stimuli and the delayed recognition foils (e.g., substitutions for the first or last number of the area code, substitutions for ending of town name ["…ville" vs. "…burg"]).

Advisory Council Ratings and Equivalent Forms

Each of the eight forms of the Daily Living Memory materials was rated by the Advisory Council members for difficulty level, ethnic/racial/cultural bias, sex bias, other biases, face validity, ecological validity, and overall task satisfaction. Forms with unacceptable levels of bias were then eliminated. Forms were sorted from highest to lowest overall satisfaction rating, which was used as the primary basis for choosing the final two forms. The final two forms were selected on the basis of the highest satisfaction rating as well as approximately equal difficulty level and face validity. The results of pilot testing were used to empirically determine the difficulty level of each item and to equate items across the two equivalent forms.

SPATIAL MODULE

Visual Discrimination

Background

Intact visuospatial functioning requires basic visual perceptual accuracy. That is, without adequate visual perception, performance on drawing, assembly, visual organization, and other more multifactorial spatial tasks would likely be impaired. The visual match-to-target paradigm is commonly used to measure visual perception (e.g., Visual Form Discrimination Test; Benton, Sivan et al., 1994). The NAB Visual Discrimination test is based on this paradigm, but unlike most similar tasks, it relies on stimuli that are not easily verbally encoded.

Task Creation

Visual Discrimination stimuli were created with computer graphics software. Several hundred stimuli were initially created with the following three basic styles: (a) single-color, green, solid irregular geometric shapes with perimeters consisting of several curves and/or straight lines; (b) two-tone blue geometric shapes, with one shape superimposed on another; and (c) single, thin purple lines with multiple curves and/or angles. These initial stimuli were pared to 72 stimuli, with approximately equal representation for each of the three styles. Computer graphics software was then used to create three foils for each of the 72 targets. Each foil was made by modifying one characteristic: orientation (e.g., 180-degree rotation), minor exaggeration or reduction in concave or convex details, or moderate exaggeration or reduction in concave or convex details. The specific placements of the target and three foil types on the page were pseudorandomly distributed.

Advisory Council Ratings and Equivalent Forms

The original 72 items, each consisting of a target on top and four choices (target and three foils) depicted below, were rated by the Advisory Council for difficulty level, design satisfaction, and overall task satisfaction. These ratings were used to eliminate items and to create the two equivalent forms for the Spatial Module. Items with poor design satisfaction and/or overall satisfaction ratings were initially eliminated. Eighteen items were eventually retained for each of the two forms, with each form containing six items for each of the three basic styles. Pairs of items (i.e., one for each form within the same style) were selected on the basis of similarity of difficulty ratings. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Screening Module

From the original 72 items, 12 items (i.e., 6 for each form) were retained for the Screening Module. Three pairs of items with low difficulty ratings and three pairs of items with high difficulty ratings were used.

Design Construction

Background

Visuoconstruction is a multifactorial function involving a combination of visual perception, motor output and integration, and spatial analysis. Visuoconstruction tasks can be separated into two major classes: assembly and drawing. Because performance on these two types of tasks does not consistently covary in neurologically impaired patients, it is important to measure them separately (Lezak, 1995). The assembly tasks included in the NAB Design Construction tests were adapted from the ancient Chinese puzzle game that uses a set of seven geometrically shaped puzzle-like pieces, or tans, to copy two-dimensional designs, or tangrams.

Task Creation

Traditional tangram puzzles are always based on the same seven proportional shapes (tans): five triangles, one square, and one rhomboid. The Screening Module Design Construction and Spatial Module Design Construction tests use only five of these original tans (two large triangles, one small triangle, one square, and one rhomboid). Ten designs of increasing difficulty were initially created for the Spatial Module version of the test. The first three items do not have the number of tans needed for correct reproduction printed at the top of the design (i.e., the examinee is provided with all five tans and is not told how many are to be used to complete the design). These three items require 2, 3, and 4 tans to complete, respectively, and each was created so that none of the individual shapes shared any contiguous sides. The remaining seven items have the number of tans needed for correct reproduction printed at the top of the design; they required from 2 to 4 tans to complete, and each design has at least one shared contiguous side. The 10 designs were pilot tested to ensure both (a) that the designs increased in difficulty and (b) that the most difficult item could be completed by most pilot test participants in 240 seconds or less. Modifications to the designs were made on the basis of the pilot test results. Once the initial 10 designs were completed, two additional forms were created by specific rotations of entire designs and/or specific tans within a design. This process resulted in a total of 30 designs.

Advisory Council Ratings and Equivalent Forms

The 30 designs were rated by the Advisory Council members for difficulty level and overall task satisfaction. Items with low satisfaction ratings were eliminated first. Difficulty ratings were then used to equate and select pairs of designs (i.e., one design for each of two forms) and to order the designs according to increasing difficulty. Eight items were retained for each of the two forms. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate the two equivalent forms.

Screening Module

The Screening Module Design Construction test is identical to the Spatial Module Design Construction test, except the Screening Design Construction test has only three items. The first item requires two tans and does not provide the number of tans required to correctly reproduce the target. The second item requires three tans and does provide the number of tans required. The third item requires five tans and does provide the number of tans required. Items 2 and 3 have tans that share two or more contiguous sides. A total of nine designs were rated by the Advisory Council members for difficulty and overall task satisfaction. The pairs of items with the best combination of highest satisfaction and similar difficulty levels were retained, one item for Form 1 and one item for Form 2.

Figure Drawing

Background

Drawing tasks are a common component of virtually every mental status examination (e.g., Folstein et al., 2001), dementia evaluation (e.g., Jurica et al., 2001; Morris et al., 1989), neuropsychological screening test (e.g., Randolph, 1998), and bedside neuropsychological battery (e.g., Kessler, 1998), as well as flexible neuropsychological evaluations (e.g, Walsh & Darby, 1999), due to their sensitivity to a variety of neurologic disorders and conditions (Lezak, 1995). Drawing tasks vary widely, from simple copying of crosses or pentagons, to clock drawing, to copying of complex geometric figures. The NAB Figure Drawing task was designed to be less complex than the commonly used Rey-Osterrieth Complex Figure (Rey, 1941) in order to avoid floor effects with significantly impaired individuals but to still be sensitive to the executive aspects (e.g., organization) of figure drawing (Freeman et al., 2000; Somerville, Tremont, & Stern, 2000). A free recall condition immediately following the copy condition was included in order to provide a measure of the examinee's encoding or processing of spatial information (Westervelt, Somerville, Tremont, & Stern, 2000), not as a measure of memory, per se.

Task Creation

One figure was created with a main rectangle, horizontal and vertical bisectors, one external element on each side of the figure (one triangle, one rectangle), and distinct elements within each of the four quadrants, such that an equal number of total elements appeared in each half (right vs. left) of the figure. Two internal elements (a large "X" or "cross" in one quadrant and an oval divided between two quadrants) were designed to "pull" for fragmentation. In order to be sensitive to planning ability, the figure was designed so that some elements extend to halfway points on the outer rectangle, other elements extend to quarter-way points, and some extend to points not easily discernable. The figure was centered on an 8½ in. x 11 in. page in a landscape (horizontal) orientation. The initial figure was field tested to ensure that (a) it was neither too easy to draw by healthy control examinees, nor too difficult to draw by neurologically impaired patients with dementia, (b) it was reported to be "enjoyable" to draw, and (c) it took much less time to draw than the Rey-Osterrieth Complex Figure. Two additional figures were then created: One figure was a mirror image of the original, although with the internal elements in the right side of the original figure rotated 180 degrees. The other was also a mirror image of the first, although with the internal elements in the left side of the original figure rotated 180 degrees. These three figures were then submitted for Advisory Council ratings in order to determine their overall satisfaction with the figures and to eliminate one figure.

The scoring system for the figure was designed to be completed quickly (less than 5 minutes) and reliably and to provide an overall summary score (based on the presence, accuracy, and placement of the individual elements). The scoring system also provides quantifiable measures of important qualitative features of the drawing: fragmentation, planning, and organization. Some aspects of the Boston Qualitative Scoring System for the Rey-Osterrieth Complex Figure (BQSS; Stern et al., 1999) were included in the Figure Drawing scoring system. So that the order of pen strokes used in completing the drawing (for the scoring of fragmentation and planning) can be recorded, a pen-switching procedure is used during the administration of Figure Drawing. Some have suggested that switching pens is overly distracting to the examinee and may result in more fragmentation during figure drawing (Meyers & Meyers, 1995). However, a prospective study using the BQSS (Ruffolo, Javorsky, Tremont, Westervelt, & Stern, 2001) found that pen-switching resulted in no more fragmentation or other difficulties than single-pen use (with the examiner keeping a flow-chart). In addition, pen-switching made recording easier for the examiner, and the resulting figures required significantly less time to score.

Advisory Council Ratings and Equivalent Forms

The three initial figures were rated by the Advisory Council for difficulty level, verbal encodability, and overall task satisfaction. All three figures received very high overall task satisfaction ratings. The two with the most similar difficulty and verbal encodability ratings were retained as Form 1 and Form 2 stimuli for Figure Drawing. The results of pilot testing were used to empirically determine the difficulty of each figure and to equate the two equivalent forms.

Map Reading

Background

One goal for the Spatial Module Daily Living test was to create a task that does not require a motor response and that does not depend on speed for successful performance. However, in keeping with the overall criteria for the NAB Daily Living tests, the Spatial Module Daily Living test must have (a) multifactorial task demands, (b) face validity to the examinee, and (c) similarity to a task of everyday living. The Map Reading test was, therefore, designed to be a measure of visuospatial skill, spatial/directional orientation, right–left orientation, and visual scanning.

Task Creation

In the design of the task, one city map was created, with avenues and boulevards traversing north and south and streets and roads traversing east and west. Two highways were included with an intersecting exit ramp. Once the map was finalized, another map was created by rotating the original map 180 degrees. Different street names (including changing boulevards to lanes) and route numbers were used. A number of questions (*n* = 22) for each of two forms were initially created, with an equal balance of questions requiring the mileage legend (e.g., how many miles between point A and point B) and the compass rose (i.e., east, west, north, south directions) and those requiring right–left orientation. Two additional sample questions were created for each form. Care was taken to balance the quadrants in which the questions had starting and ending points. The questions for Form 1 were created initially, and then parallel questions were written for Form 2.

Advisory Council Ratings and Equivalent Forms

The Advisory Council rated each of the two maps (Map Form 1 and Map Form 2) for overall stimulus satisfaction. Both maps received high satisfaction ratings. In addition, each of the 24 questions (2 sample items and 22 test questions) for each of the two forms was rated for difficulty level, sex bias, U.S. regional bias, ethnic/racial/cultural bias, task appropriateness, and overall task satisfaction. On the basis of these ratings, the number of items per form was reduced to the final 14 (2 sample items and 12 test items).

As just described, the two maps were nearly identical spatially, although they differ by 180 degrees rotation. In the selection of the final 14 questions for each form, care was taken to retain pairs of questions. That is, only items in which both the parallel Form 1 and Form 2 versions of the items received high overall Advisory Council ratings were retained. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

EXECUTIVE FUNCTIONS MODULE

Mazes

Background

Planning and foresight are important aspects of executive functioning (Stern & Prohaska, 1996) and are frequently impaired in patients with frontal lobe dysfunction (Damasio & Anderson, 2003). There are, however, few formal measures of planning (Lezak, 1995). Tower tests (e.g., Shallice, 1982) are commonly used as measures of planning; however, they are not easily amenable to the development of alternative forms without significant practice effects, and they require somewhat cumbersome manipulatives. Among the few other paradigms used to measure planning, maze-tracing tasks have historically been found to be especially sensitive to frontal lobe lesions (Milner, 1968; Porteus, 1959). The NAB Mazes test is based on this maze-tracing paradigm and was designed to avoid both floor and ceiling effects found in existing maze tests.

Task Creation

Seven mazes were initially created, with increasing complexity from very simple to very difficult. All mazes were 9 in. wide and, with the exception of the first and easiest maze (which was 3 in. high), were 6 in. high. The alley width for the first two mazes was kept constant at 1 in. The third maze had ¾ in. alleys. The fourth maze had ½ in. alleys. The last three mazes had ¼ in. alleys. Care was taken so that the paths traversed all four quadrants of the maze. The "start" and "end" points for the mazes were divided between right and left and between center and perimeter. Initial designs were pilot tested and subsequently modified to ensure increasing completion times across the seven mazes. Once the seven mazes were finalized, two additional alternate forms of each maze were created in the following manner. The first alternate form was created by rotating the original maze 180 degrees along its vertical axis (keeping the "start" and "end" points in the original, unrotated position). The second alternate form was created by rotating the original maze 180 degrees along its horizontal axis (again, keeping the "start" and "end" points in the original, unrotated position). In this manner, a total of 21 mazes was created.

Advisory Council Ratings and Equivalent Forms

The 21 mazes (three alternate forms of 7 mazes each) were rated by the Advisory Council members for difficulty level and overall task satisfaction. The two mazes from each three-maze group with the best combination of highest satisfaction and similar difficulty levels were retained. That is, seven mazes were eliminated, and two forms with seven mazes each were retained. The results of pilot testing were used to empirically determine the difficulty level of each item, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Screening Module

The Screening Module Mazes test was created according to identical methods used to create the Executive Functions Module Mazes test, except there were only three mazes per form. All three mazes were 9 in. wide. The first maze was 3 in. high, and the second and third were 6 in. high. The alley width for the first two mazes was kept constant at 1 in. The third maze had ½ in. alleys. The same rotation procedures were used to create the second and third forms from the original set of three mazes. The three forms (i.e., 9 mazes) were rated by the Advisory Council for difficulty level and overall task satisfaction. The two forms (3 mazes each) with the best combination of highest satisfaction and similar difficulty levels were retained.

Judgment

Background

Caregivers and coworkers of patients who have damage to the prefrontal cortex and associated executive dysfunction often complain of the patients' poor judgment in daily living (Parker, 1990). Additionally, "impaired judgment" is considered an important diagnostic feature of dementia (Knopman et al., 2001). Therefore, assessment of judgment is a central aspect of mental status testing, in general, and in the examination of decisional capacity (i.e., competency), in particular. Informal and unstandardized assessments of judgment are included in most mental status examinations, and some neuropsychological and intelligence tests include a small number of judgment-related questions (e.g., Kiernan et al., 1987; Wechsler, 1997a). However, there have been surprisingly few formal measures of judgment as it pertains to critical aspects of independence in daily living. The Daily Living test for the Executive Functions Module was designed to measure this important area of functioning. It includes a series of questions about home safety, health, and medical issues likely to be encountered in everyday life.

Task Creation

A minimum of 10 questions was generated for each of six major categories: (a) home safety, (b) personal hygiene, (c) medication safety, (d) motor vehicle driving safety, (e) medical decision making, and (f) general judgment. All questions were written at approximately an 8th-grade reading level as determined by the Flesch-Kincaid reading formula (Flesch, 1994). A total of 77 questions was created.

Advisory Council Ratings and Equivalent Forms

The 77 original items were rated by the Advisory Council members for difficulty level, sex bias, U.S. regional bias, ethnic/racial/cultural bias, clinical utility, task appropriateness, and overall task satisfaction. Items with poor ratings on any of the bias categories were eliminated initially. Items with poor clinical utility, low task appropriateness, or low overall task satisfaction ratings were also eliminated. All items about driving safety and general judgment were eliminated on the basis of these item-reduction rules. Thirteen items were retained for each form. Selection was based on an iterative process in which the following two goals/factors were maximized: (a) there should be pairs of similar items across the two forms (e.g., "What should you do if you take too much of a prescription medication?" and "What is the best thing to do if you accidentally take someone else's medication?"), and (b) the overall Advisory Council difficulty level ratings for the group of 13 items should be similar across the two forms. On the basis of the results of the national standardization, three additional item pairs (i.e., three from each form) were eliminated due to either poor interrater reliability or inconsistent comprehension of the questions by examinees. This process resulted in the final 10 items per form. The results of pilot testing were used to empirically determine the difficulty level of each of the original 13 items, to order the items in ascending difficulty, and to equate items across the two equivalent forms.

Categories

Background

Concept formation, cognitive flexibility, generativity, and novel problem solving are all major functions subsumed under the overall domain of executive functioning. These skills and functions are the focus of some of the most commonly used and time-honored neuropsychological instruments purported to be sensitive to frontal systems dysfunction (e.g., Berg, 1948; Delis et al., 2001; Heaton, Chelune, Talley, Kay, & Curtiss, 1993; Reitan & Wolfson, 1993). In the design of the NAB Categories test, the strengths of existing sorting and classification tasks were incorporated, along with additional features that allow for an alternate form with little practice effects, no additional manipulatives or cards, and face-valid and visually attractive stimuli.

Task Creation

Three complete forms of the Categories test were created initially. Each form consisted of two separate panels; each panel contained photographs of six adults along with associated identifying information. Hundreds of photographs were initially culled from catalogs of digital stock photography. The selection of the final 36 photographs used in the initial three forms was made through an iterative process involving extensive pilot testing and regrouping and reselecting photographs in order to yield forms with similar possible categorization solutions based on the visual information included in the photographs (e.g., eye glasses vs. no eye glasses, round shirt collar vs. button-down shirt collar). Once photographs were selected, the background of each photograph was digitally modified to be either blue-gray or yellow (according to categorization decision rules). For each of the six panels (i.e., two panels per form), additional categorization rules were used to create distinct outlines (e.g., three punch holes vs. spiral-binding holes) and backgrounds (thin-lined border vs. thick-lined border) for each photograph, in addition to the identification information printed below each photograph. For each form, the first panel included six identifying information categories (e.g., name, occupation, place of birth), and the second panel included five identifying information categories. Each category (except marital status) was designed to have several possible solutions for 1- and 2-point scores; the possible correct responses for the first panel were not possible correct responses for the second panel. Once again, pilot testing results guided modifications to the category information such that each panel had a similar number of possible solutions across the three forms.

Advisory Council Ratings and Equivalent Forms

The three forms (6 panels) were rated by the Advisory Council members for difficulty level, ethnic/racial/cultural bias, educational bias, U.S. regional bias, sex bias, and overall task satisfaction. The two forms (two panels each) with the lowest bias ratings, highest task satisfaction ratings, and most similar difficulty ratings were retained. The results of pilot testing were used to empirically determine the difficulty level of each panel and to equate panels across the two equivalent forms.

Word Generation

Background

Word-generation tasks (e.g., Spreen & Benton, 1977; Benton, Hamsher et al., 1994) are sensitive indicators of dementia (e.g., Small, Herlitz, Fratiglioni, Almkvist, & Bäckman, 1997), brain damage, in general (e.g., Mutchnick, Ross, & Long, 1991), and frontal systems dysfunction, in particular (e.g., Miceli, Caltagirone, Gainotti, Masullo, & Silveri, 1981). These tests are often the most highly correlated with all other executive functioning tests (e.g., Somerville et al., 2000), and they are strongly associated with caregiver reports of instrumental activities of daily living in elderly patients (Cahn-Weiner, Boyle, & Malloy, 2002). However, existing word-generation tasks, in which the examinee is typically asked to say as many words as he/she can think of that begin with a specific letter within a 60-second time limit, are highly associated with educational level (e.g., Ruff, Light, & Parker, 1996) and are affected by language-related impairments in which there is reduced lexical retrieval. The NAB Word Generation test was designed to be a measure of generativity, similar to existing word-generation tasks. However, the NAB Word Generation task was created to be more specific to executive impairment than to language impairment and to be less influenced by educational level than current measures. Therefore, in the NAB test, all examinees are provided with the identical set of letters from which to generate as many three-letter words as they can within a specific time limit.

Task Creation

Three forms were initially developed. Each form included eight letters: two vowels and six consonants. The same two vowels were used in each form ("a" and "o"). All possible three-letter combinations (using each letter only once per word) for each form were created with an internet-based anagram server (http://www.wordsmith.org/anagram/index.html). The total number of acceptable words (as defined by published dictionaries including *Merriam-Webster's Official SCRABBLE Players Dictionary, Third Edition*, 1999) was then established for each form. Each of the resulting words was then checked for the frequency of use in the English language as defined by the Standard Frequency Index (SFI; Zeno et al.,1995). The letters used in each form were altered in an iterative process, such that the number of possible (nonproper noun) words (30–31) and the mean SFI (49.3–50.6) of the possible words were similar across the three forms.

Advisory Council Ratings and Equivalent Forms

The three original forms were rated by the Advisory Council for difficulty level and overall task satisfaction. The two forms with the best combination of highest satisfaction and similar difficulty levels were retained. The results of pilot testing were used to empirically determine the difficulty level of each vowel/consonant set and to equate the vowel/consonant sets across the two equivalent forms.

Screening Module

The Screening Word Generation test is identical to the Executive Functions Module Word Generation test except that the total number of letters per form is six: two vowels and four consonants. The same two vowels ("e" and "u") are used in all three forms. The procedures used to create the three Screening Word Generation forms were the same as those used for the main Executive Functions Module version. The letters used in each form were altered in the same iterative process, such that the number of possible (non-proper noun) words (11–12) and the mean SFI (46.5–51.7) of the possible words were similar across the three forms. The three original forms were rated by the Advisory Council for difficulty level and overall task satisfaction. The two forms with the best combination of highest satisfaction and similar difficulty levels were retained.