Test Validity and Performance Validity: Considerations in Providing a Framework for Development of an Ability-Focused Neuropsychological Test Battery

Glenn J. Larrabee*

Independent Practice, Sarasota, FL, USA

*Corresponding author at: 2650 Bahia Vista Street, # 308, Sarasota, FL 34242, USA. Tel.: +1-941-955-9596; fax: +1-941-957-3485.
E-mail address: glarrabee@aol.com (G. J. Larrabee).

Accepted 8 September 2014

Abstract

Literature on test validity and performance validity is reviewed to propose a framework for specification of an ability-focused battery (AFB). Factor analysis supports six domains of ability: first, *verbal symbolic*; secondly, *visuoperceptual and visuospatial judgment and problem solving*; thirdly, *sensorimotor skills*; fourthly, *attention/working memory*; fifthly, *processing speed*; finally, *learning and memory* (which can be divided into *verbal* and *visual* subdomains). The AFB should include at least three measures for each of the six domains, selected based on various criteria for validity including sensitivity to presence of disorder, sensitivity to severity of disorder, correlation with important activities of daily living, and containing embedded/derived measures of performance validity. Criterion groups should include moderate and severe traumatic brain injury, and Alzheimer's disease. Validation groups should also include patients with left and right hemisphere stroke, to determine measures sensitive to lateralized cognitive impairment and so that the moderating effects of auditory comprehension impairment and neglect can be analyzed on AFB measures.

Keywords: Assessment; Test construction; Meta-analysis; Head injury; Traumatic brain injury; Alzheimer's disease; Cerebrovascular disease/accident and stroke

Introduction

Bauer (2000) distinguished between two major approaches to neuropsychological assessment: first, the fixed battery approach, wherein everyone receives the same comprehensive battery of tests, regardless of the referral question or patient clinical presentation and second, a flexible battery, wherein there is a limited core set of procedures administered to all, to provide a basis for generating clinical hypotheses about the patient's neuropsychological status for purposes of additional evaluation. Bauer also discusses an intermediate approach, which he characterizes as multiple fixed battery, which can characterize population-specific batteries constructed for specific clinical disorders, for example, multiple sclerosis, traumatic brain injury (TBI), or domain-specific batteries, constructed for extensive evaluation of a particular process such as language or memory (e.g., Multilingual Aphasia Examination, Benton, Hamsher, & Sivan, 1994; Wechsler Memory Scale-IV; WMS-IV, Wechsler, 2009).

Surveys have established that the flexible battery approach has been the major practice orientation in clinical neuropsychology for several years (Rabin, Barr, & Burton, 2005; Sweet, Meyer, Nelson, & Moberg, 2011). In the most recent survey (Sweet et al., 2011), 78% of neuropsychologists endorsed using a flexible battery (core set of procedures with additional testing based on clinical history and core test findings), compared with 18% using a completely flexible approach (test selection governed entirely by referral question and patient clinical presentation), and 5% using a fixed standardized battery (e.g., Halstead-Reitan, Reitan & Wolfson, 1993; Luria-Nebraska, Golden, Purisch, & Hammeke, 1985; or Neuropsychological Assessment Battery, NAB, Stern & White, 2003; [note the NAB also has a feature allowing for more flexible application, with administration of a screening battery, which can be followed by more in depth examination of core areas of ability]). The assessment practice survey by Rabin et al. (2005) reflects high test use frequencies for select procedures such as 63.1% using the WAIS-R/III (Wechsler, 1981, 1997a),

42.7% using WMS-R/WMS-III (Wechsler, 1987, 1997b), 17.6% using the Trail Making Test (Reitan & Wolfson, 1993), and 17.3% using the CVLT/CVLT-II (Delis, Kramer, Kaplan, & Ober, 1987, 2000). This survey did not elicit data on frequency of combinations of tests, for example, the percentage of clinicians using the WAIS-R, WMS-R, CVLT, and Trail Making Test.

Despite the popularity of the flexible battery approach, there is no common set of tests comprising the core of a flexible battery, which is universally used. The purpose of the current paper is to lay out a framework for composing a standard neuropsychological test battery that can serve as the core for a flexible battery, supported by construct and criterion validity, and which also contains embedded/derived measures of performance validity (Performance Validity Tests, PVTs; Larrabee, 2012a).

The reader should be aware that Meyers (Meyers & Rohling, 2004) has developed a 2½ h core for a flexible battery, the Meyers Neuropsychological Battery (MNB), comprised of 22 tests, which also includes 11 embedded/derived PVTs (Meyers et al., 2014; Meyers & Volbrecht, 2003). While the MNB does follow some of the validity guidelines I will be presenting, it differs, largely in three ways: first, originally selecting tests on the basis of sensitivity to presence of brain dysfunction in a mixed sample of neurologic cases (Meyers, Miller, & Tuita, 2013); second, using only one test each to represent the domains of motor function, verbal and visual learning and memory; finally including tests not in common use (with the exception of those clinicians using the MNB): 1-Minute Estimation, and Dichotic Listening, as well as administering two tests that are usually only administered during evaluation of acquired aphasia (Sentence Repetition and the Token Test). The MNB does, however, present a model for how the core of a flexible battery can be composed of individually normed tests that in many ways functions as well as more extensive batteries of co-normed tests. This is likely because the MNB contains sensitive measures of memory and processing speed such as the Auditory Verbal Learning Test (AVLT), Complex Figure Test and Trail Making Test (cf. Rohling, Meyers, & Millis, 2003). As others have shown, tests of processing speed and memory are among the tests most sensitive to acquired brain impairment. This is particularly true when the brain damage is of a diffuse nature in conditions such as Alzheimer's disease (AD) or moderate and severe TBI (Backman, Jones, Berger, Laukka, & Small, 2005; Christensen, Griffiths, Mackinnon, & Jacomb, 1997; Dikmen, Machamer, Winn, & Temkin, 1995; Larrabee, Millis, & Meyers, 2008; Miller, Fichtenberg, & Millis, 2010; Powell, Cripe, & Dodrill, 1991).

In the following sections, I review a framework for developing an ability-focused neuropsychological battery (AFB) that is based upon multiple types of validity, consistent with the need to establish evidence-based standards for neuropsychological practice (Chelune, 2010). Key criterion groups are proposed to include moderate and severe TBI, and AD, with further analysis of subjects with left or right hemisphere stroke (cerebrovascular accident [CVA]). Three types of criterion validity will be considered: first, identification of the procedures sensitive to the presence of brain dysfunction; second, identification of the procedures sensitive to the severity of impairment; finally, identification of procedures that are the best predictors of competence in instrumental activities of daily living such as ability to drive a motor vehicle and independently manage one's finances. The moderating effects of aphasia in left hemisphere stroke, and neglect in right hemisphere stroke are considered in relation to subtests that are potential candidates for the core battery. Construct validity is addressed through review of factor analytic research, to identify core domains of ability, as well as the tests that are relatively pure measures of these core neuropsychological constructs. PVTs for evaluating whether the patient being examined is providing an accurate measure of actual level of ability (Larrabee, 2012a) are reviewed for tests that could serve as primary tests comprising the AFB.

A hypothetical battery will be offered based on this review. It is not the goal of this paper to present *the* common AFB, but rather to frame key issues related to battery specification that can provide guidance not only for clinicians determining their own evidence-based approach to assessment but also for potential inter-organizational efforts in this regard. Ultimately, adoption of a consensus AFB by the field would greatly enhance data analysis in both the individual case as well as in applied research by providing large datasets aggregated over a common set of procedures for a variety of neurological, psychiatric, and developmental conditions.

Criterion Validity

Effect Size and the Validity of Neuropsychological Tests

An effect size, generally defined, reflects the magnitude of a relationship between two variables, and can be represented by various statistics including the standardized mean difference, Pearson product-moment correlation, or odds ratios (see Borenstein, Hedges, Higgins, & Rothstein, 2009). In the present paper, the effect size, *d*, represents the standardized mean difference; in other words, the difference between the mean performance of two groups on a neuropsychological test in terms of the pooled *SD* (control plus clinical group; Cohen, 1988). Effect sizes of 0.20 are small, 0.50 are medium, and 0.80 or greater are large (Cohen, 1988). The larger the effect size, the greater the separation of the test performance of the two comparison groups. Cohen (Table 2.2.1, Cohen, 1988, p. 22) has provided percent non-overlap for various magnitudes of *d*, and Zakzanis, Kaplan and Leach (Table 2.1, Zakzanis, Kaplan, & Leach, 1999, p. 13) have reflected these values to demonstrate the percent of overlap as a function of the magnitude of *d*. For example, for a *d* of 1.0, the overlap is 44.6%, dropping to 18.9% for a *d* of 2.0;

thus the larger the value of d , the smaller the overlap percent, and the smaller the diagnostic error for both false positives and false negatives.

Effect size comparisons can provide information directly relevant to differential sensitivity of neuropsychological tests measuring the same construct. Loring et al. (2008) found a Cohen's d of 0.47 for Rey AVLT (Rey, 1964) scores of right vs. left temporal lobe epilepsy patients, which was substantially higher than the d of 0.29 for the same comparison employing the California Verbal Learning Test (CVLT; Delis et al., 1987). Comparing the performance of right versus left temporal lobe epilepsy groups on Boston Naming (Kaplan, Goodglass, & Weintraub, 1983) yielded a Cohen's d of 0.56, compared with a Cohen's d of 0.36 for the Benton Visual Naming Test (Benton, Hamsher, et al., 1994). This type of investigation yields information directly relevant to selection of neuropsychological tests for specific clinical groups (e.g., epilepsy) as well as for selection of measures of naming ability and verbal episodic memory for core measures in a common AFB.

Effect sizes are directly relevant to diagnostic statistics as the larger the effect size, the smaller the group overlap, and the smaller both false negative and false-positive error rates become. Simultaneously, there is an increase in both sensitivity (true positives, or correct identification of subjects with the condition of interest), and specificity (true negatives, or correct identification of subjects who do not have the condition of interest).

The effect size is also related to the area under the receiver operating characteristic (ROC) curve. The ROC is derived by plotting the false-positive error rate ($1 - \text{specificity}$) on the x -axis and the true-positive rate (sensitivity) on the y -axis for each potential cutting score comparing two groups on a diagnostic test (Hsaio, Bartko, & Potter, 1989; Swets, 1973). When the distributions for the false-positive errors and sensitivity each are normally distributed, there is a one to one correspondence between ROC area under curve (AUC) and the effect size, d (Rice & Harris, 2005). Consequently, literature review yielding effect sizes for discrimination of neurologically impaired versus control subjects can yield information similar to that provided by ROC analysis, even though an actual ROC curve has not been plotted. Given the fact that ROC curves plot the false-positive rate and sensitivity at each possible test score, in studies that report both d and ROC AUC, the AUC is the preferred statistic. Additionally, ROC AUC is the preferred statistic if the data are skewed (Fawcett, 2003). ROC AUC can vary between 0.00 and 1.0, and AUC of 0.50 represents chance discrimination of the two groups. Interpretive guidelines suggest that a minimum ROC AUC of 0.70 is required for acceptability; AUC between 0.80 and 0.90 is considered excellent, and AUC in excess of 0.90 is considered outstanding (Hosmer & Lemeshow, 2000; Miller et al., 2010).

Aggregation of effect sizes, across multiple studies, is the basis of meta-analysis (Borenstein et al., 2009). Meta-analysis has been used to evaluate neuropsychological outcome in mild TBI, showing essentially full recovery by 3 months post-trauma (Belanger, Curtiss, Demery, Lebowitz, & Vanderploeg, 2005; Binder, Rohling, & Larrabee, 1997; Frencham, Fox, & Maybery, 2005; Rohling et al., 2011; Schretlen & Shapiro, 2003). Meta-analysis has been applied to characterize severity of neuropsychological impairment as a function of severity of TBI, showing a linear increase in impairment as a function of time-to-follow commands (TFCs), up to and including 30 days or more of coma (Rohling et al., 2003).

In a very interesting application of meta-analysis, Zakzanis et al. (1999) use meta-analytically derived effect sizes, per neuropsychological test domain (verbal skills, performance skills, memory acquisition, memory delay, attention/concentration, cognitive flexibility and abstraction, and manual dexterity), to define different profiles of performance in various patient groups, aggregating across multiple investigations of patients with disorders such as AD and Parkinson's disease. Meta-analysis has also been used to demonstrate the comparative average sensitivity (effect sizes) of individual tests as well as specific domains of performance in comparing the neuropsychological performance of depressed patients versus normal controls, and depressed patients versus those with AD (Christensen et al., 1997).

Effect sizes can also be used to analyze cross-sectional age-related changes in level of performance on tests of various neuropsychological functions. For example, data from the manual for the Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981), comparing 70- to 74-year-old persons with persons ages 20–24, show effect sizes of 0.60 for Verbal IQ, versus 1.80 for Performance IQ. Using data from Heaton, Miller, Taylor, and Grant (2004), age effect sizes (70–74 versus 20–24) are 1.40 for the Category Test and 1.80 for the Tactual Performance Test, for these measures of visuospatial/visuoperceptual problem solving. Attention/Working Memory age effect sizes are 0.50 for Seashore Rhythm Test, 0.33 for Digit Span, and 0.50 for WAIS-R Arithmetic. Processing Speed age effects are 1.83 for WAIS-R Digit Symbol, and 1.60 for Trail Making B. Motor Function age effects are 0.90 for Finger Tapping, 1.20 for Grip Strength, and 2.00 for the Grooved Pegboard. Verbal memory age effect sizes are 2.50 for learning on the Selective Reminding Test, and 3.31 for delayed recall (Larrabee, Trahan, Curtiss, & Levin, 1988). Visual memory age effects are 1.87 for Continuous Visual Memory Test (CVMT) total correct over the learning trials, and 2.01 for CVMT delayed recognition (Trahan & Larrabee, 1988).

These data show that age effects are smaller for verbal intellectual functions, and attention/working memory, compared with larger age effects for visuoperceptual/visuospatial problem solving, processing speed, fine motor skill, and verbal and visual learning and memory. As noted earlier, measures of processing speed, and verbal and visual learning and memory also tend to be the processes most sensitive to diffuse brain dysfunction caused by factors such as TBI and dementia (Backman et al., 2005;

Christensen et al., 1997; Dikmen et al., 1995; Powell et al., 1991). This underscores the critical importance of having normative data corrected for age and, where necessary, educational attainment and sex, across the age range, to reduce the likelihood of false-positive findings, particularly in the elderly.

Sensitivity to Presence and Severity of Brain Dysfunction

In the past, “brain damage” or “brain dysfunction” was considered as a unitary (present vs. absent) or unidimensional (more or less of) construct and it was common to mix various etiologies of disorders in one group, such as TBI, brain tumors and stroke. Over the years, it has become obvious that “brain damage” is not a unitary or unidimensional construct, and in modern neuropsychology, the criterion is typically presence or absence of a particular *type* of brain dysfunction, and its differential impact on key neuropsychological abilities (cf. Zakzanis et al., 1999), with commonly seen disorders including those resulting from TBI, stroke, and dementia (Lezak, Howieson, Bigler, & Tranel 2012).

Certain modifiers of criterion validity are also important to consider, such as disease/injury severity, presence/absence of language comprehension impairment in left hemisphere stroke, and presence/absence of neglect in right hemisphere stroke. As will be shown, tests that are sensitive to presence/absence of a disease/injury may not be the same tests that are sensitive to severity of disease/injury or to the everyday functional consequences of a particular disorder such as Alzheimer’s disease. Failure of a task such as WAIS-IV Block Design may represent a visuospatial problem-solving deficit in a person with a right hemisphere stroke, whereas failure of the same task in a patient with left hemisphere stroke may represent the consequence of comprehension impairment secondary to Wernicke aphasia, rather than representative of a pure visuospatial deficit.

In TBI, persisting impairments at 1 year post-injury are not typically found on most neuropsychological assessment tools, until the initial TFCs is between 1 and 24 h (1–24 h of coma). The only measures in a comprehensive neuropsychological battery that were sensitive to persistent deficit in this injury severity group were Verbal Selective Reminding (Buschke, 1973; Larrabee et al., 1988), a sensitive measure of verbal supraspan learning, and Trail Making Part B (Reitan & Wolfson, 1993), a measure of psychomotor speed and set shifting. Moreover, the effect size for Verbal Selective Reminding, 0.46, was three times the effect size for Trail Making B, 0.15, reflecting greater sensitivity of verbal memory than processing speed. In a mixed neurological group, comprised primarily of TBI and seizure disorder patients, performance on the AVLT Trial V (Lezak et al., 2012; Rey, 1964) was more sensitive to discriminating the neurologic group from a normal control group than any other measure of performance, including tasks of verbal and visual concept formation and problem solving, processing speed, attention/working memory, and visual memory function (Powell et al., 1991). These data demonstrate that measures of verbal supraspan learning are among the most sensitive of neuropsychological tests.

Comparison of effect sizes across performance domains for Alzheimer’s disease, Parkinson’s disease with dementia, and major depressive disorder (Zakzanis et al., 1999) showed greatest effect sizes for measures of delayed recall for Alzheimer’s and depression, relative to other abilities within each group. Within the Alzheimer’s group, the delayed recall effect size of 3.23 was nearly four times the effect of manual dexterity, $d = 0.85$. In contrast, in Parkinson’s with dementia, the delayed recall effect size of $d = 1.82$ was *smaller* than the manual dexterity effect size $d = 2.42$, consistent with the major effects of this disease on motor functions. These data show how different disorders can differentially impact performance on the major neurobehavioral domains of ability.

Neuropsychological effects are clearly related to severity of TBI, as defined by TFCs (Dikmen et al., 1995; Rohling et al., 2003). Using an overall test battery mean (OTBM), represented as an average z score, effect size for neuropsychological performance at 1 year post-trauma was $d = -0.02$ for TFCs of <1 h, increasing linearly to $d = -0.22$ for 1–23 h TFC, $d = -0.45$ for 1–6 days TFC, $d = -0.68$ for 7–13 days TFC, $d = -1.33$ for 14–28 days TFC, and $d = -2.31$ for >28 days TFC (Rohling et al., 2003). Thus, the most severely injured group ($d = -2.31$) performed over 2 SD worse than the least severely injured group, whose performance was essentially identical to that of orthopedic trauma controls, at $d = -0.02$. Donders, Tulsky, and Zhu (2001) found that WAIS-III Digit Symbol, Symbol Search, and Letter–Number Sequencing subtests discriminated between patients with moderate–severe TBI and normal subjects or patients with mild TBI, who did not differ from one another. They also showed that of the four WAIS-III factors, processing speed was the most sensitive to the effects of TBI. As noted earlier, processing speed (Trail Making B) and verbal learning and memory (Verbal Selective Reminding) were the most sensitive measures for identifying residual impairment at 1 year post-trauma, in the group requiring 1–23 h TFC, in the Dikmen and colleagues (1995) investigation.

In AD, tests most sensitive to discriminating patients with AD from normal elderly are those measuring learning and memory, particularly delayed recall (Larrabee, Largent, & Levin, 1985; Welsh et al., 1994; Zakzanis et al., 1999). Despite the sensitivity of memory tests to detection of cognitive impairment associated with AD, memory tests may not be sensitive to severity of the disorder. Larrabee, Largent, et al. (1985) found that Verbal Selective Reminding, the most sensitive measure discriminating AD from normal elderly, did not correlate at all with severity of AD; rather, WAIS Information and Digit Symbol reflected significant

correlation with disease severity, as measured by the Clinical Dementia Rating Scale (CDR; Hughes, Berg, Danziger, Coben, & Martin, 1982), or functional adaptive impairment, as measured by the Blessed, Tomlinson, and Roth (1968) dementia rating scale. Consistent with the findings of Larrabee, Largen, et al. (1985), Griffith and colleagues (2006) found that subjects with mild cognitive impairment (MCI), many of whom are likely in the beginning stages of AD, are discriminated from normal controls by the Hopkins Verbal Learning Test (HVLT; Brandt, 1991; $d = 1.50$), but the HVLT did not discriminate MCI from AD ($d = 0.06$). In contrast, semantic fluency discriminated AD and MCI ($d = 0.71$). Again, a procedure sensitive to detecting early stages of AD from age-peer controls (HVLT) did not discriminate severity of AD (i.e., MCI vs. AD), whereas a non-memory cognitive task (semantic fluency) was sensitive to severity of dementia.

Data from Appendix B of Dikmen and colleagues (1995, p. 90) also demonstrate that the tests most sensitive to detection of impairment are not always the tests most sensitive to the severity of impairment. As already noted, comparison of the test performance of TBI subjects taking > 1 h but < 24 h to follow commands with that of orthopedic trauma controls yielded an effect size of 0.46 for Verbal Selective Reminding. Comparing these same two groups on Finger Tapping, a simple motor speed test, yielded an effect size of 0.27 which was non-significant. In contrast, comparing the performances of those TBI subjects who took between > 24 h but < 7 days to follow commands to those who followed commands > 1 h but < 24 h, the effect size was 0.50 for Finger Tapping but only 0.07 for Verbal Selective Reminding. Again, this demonstrates that tests most sensitive to detection of impairment may not be the same tests sensitive to severity of impairment.

Prediction of Activities of Daily Living

Validity is also evaluated by correlation of neuropsychological performance with important instrumental activities of daily living such as driving a car and making financial decisions, as well as with prediction of vocational abilities. This area of criterion validity has also been referred to as *ecologic validity*.

Dikmen and colleagues (1994) predicted return to work following TBI. Return to work was associated with severity of injury, with 82% who followed commands within 1 h back at work in 1 year, contrasted with only 6% back to work who had taken over 28 days to follow commands. At 1 year, 77% of patients who could simply undergo neuropsychological testing < 1 month post-trauma had returned to work, compared with only 6% who were untestable 1 month post-trauma. Halstead Impairment Index (HII) scores of 0.2 or less were associated with a 96% return to work, compared with 66% back to work with HII scores of 0.5–0.7, and 35% back to work who had HII of 0.8 or greater (*note*: HII scores represent the proportion of scores out of 7 total, that fall in the range of impaired performance).

Williams, Rapport, Hanks, Millis, and Greene (2013) found that neuropsychological tests predicted outcome on the Disability Rating Scale (DRS), and return to work, above and beyond the predictions made by injury severity (e.g., admission Glasgow Coma Scale) and CT scan abnormalities. Particularly, significant predictors were Trail Making A and B, Grooved Pegboard, the Symbol Digit Modalities test, and measures of visuospatial ability. It is noteworthy that list learning (AVLT or CVLT), which per the above review tends to be among the most sensitive measure for detecting presence of impairment, was not a sensitive predictor of important activities of daily living.

Driving ability has been correlated with performance on Trail Making B in patients who have suffered severe TBI (Novack et al., 2006). Similarly, driving competence in patients with questionable dementia was related to performance on Trail Making B (Whelihan, DiCarlo, & Paul, 2005). Brown and colleagues (2005) found that the NAB (Stern & White, 2003) Driving Scenes subtest correlated 0.55 with a 108 point open road driving score.

Financial capacity in Alzheimer's disease, assessed via an 8-part Financial Capacity Instrument (Earnst et al., 2001) was related to a variety of neuropsychological test performances. Digits Forward related to understanding a bank statement, whereas Digits Reversed related to all four aspects of basic monetary skills. WAIS-III Letter–Number Sequencing related to several domains of monetary capacity, and the Arithmetic subtest related to basic monetary skills, checkbook, and bank statement management in Alzheimer's disease (Earnst et al., 2001). In a subsequent investigation, Sherod and colleagues (2009) found that written arithmetic skill (WRAT-3, Wilkinson, 1993) predicted financial capacity for control subjects, those with mild AD, and for those with amnesic Mild Cognitive Impairment (MCI).

Capacity to make medical decisions was related to word fluency (Controlled Oral Word Association), but not to memory performance or overall severity of cognitive impairment, in patients with AD (Marson, Ingram, Cody, & Harrell, 1995). This was despite significant differences in global cognitive function, and memory function, between patients with AD and normal controls. Again, this is consistent with the findings of Larrabee, Largen, et al. (1985), Griffith and colleagues (2006), and Earnst and colleagues (2001), in demonstrating that despite memory tests being the most sensitive discriminators of AD and normal elderly, measures of non-memory cognitive skills, specifically, phonemic fluency/word retrieval skills, may be more sensitive to severity of dementia, and accompanying impairments in activities of daily living.

Moderating Effects of Aphasia and of Neglect in Subjects with Unilateral Cerebrovascular Damage

In brain dysfunction criterion groups comprised of patients with unilateral brain damage such as can occur with stroke, particularly important moderator variables are comprehension impairment in left hemisphere stroke, and neglect in right hemisphere stroke. Benton, Sivan, Hamsher, Varney, and Spreen (1994) have analyzed performance on a variety of visuo-perceptual and visuo-spatial tasks in relation to language comprehension impairment, and visual field defect. For example, performance on Facial Recognition, a task requiring the subject to match a black and white photograph of an unfamiliar person to photographs of the same person presented in different shading contrasts, is performed more poorly by patients with posterior right hemisphere lesions (53% failure rate) than anterior right hemisphere lesions (26% failure rate). In contrast, Facial Recognition is passed by 100% of left hemisphere stroke patients without aphasia (anterior and posterior), and 100% of left hemisphere stroke patients with aphasia (anterior and posterior), but who have normal auditory comprehension. In contrast, 29% of anterior left hemisphere stroke patients, and 44% of left posterior stroke patients who have auditory comprehension impairment fail the Facial Recognition Test. Although Larrabee (1986) did not analyze auditory comprehension impairment specifically, overall severity of language dysfunction was significantly correlated with WAIS Verbal IQ (-0.77) and Performance IQ (-0.74), as well as with a variety of “non-verbal” subtests, including Block Design (-0.44) and Object Assembly (-0.72), in a group of patients with left hemisphere damage due to a variety of etiologies. Of course, aphasics with greater degree of language impairment typically also manifest significant impairment in language comprehension.

The above data clearly demonstrate that language comprehension is a moderating variable for performance on visual cognitive tasks, and must be considered in interpretation of “non-verbal” performance in aphasic patients. Benton, Sivan, et al. (1994) do provide data showing that performance on Judgment of Line Orientation does not seem to be affected by the presence/absence of auditory comprehension impairment, making this task important for the differential diagnosis of cognitive impairment secondary to one versus multiple infarctions.

Hemispatial neglect is to neuropsychological effects of right hemisphere disease as auditory comprehension impairment is to the cognitive effects of left hemisphere disease. Hemispatial neglect is a cognitive rather than sensory phenomena (Heilman, Watson, & Valenstein, 2012), and represents a failure of directed attention. Patients with a visual field cut without neglect will move the to-be-perceived object so that it will fall in the preserved visual field. Patients with neglect do not compensate for the field cut. On the Facial Recognition Test, patients with posterior right hemisphere stroke and field cut had a 58% failure rate, whereas those without field cut had a 40% failure rate (*note*: Benton et al. did not differentiate the field cut group as to which subjects had or did not have neglect, but neglect is frequently associated with presence of visual field cut).

The attentional impairment associated with neglect may reflect a more generalized attentional impairment in right hemisphere stroke. Trahan, Larrabee, Quintana, Goethe, and Willingham (1989) found a 56% rate of impairment for acquisition, and 48% rate of impairment for delayed recall on the Expanded Paired Associate Test (EPAT; Trahan et al., 1989) for left hemisphere stroke patients, which was a substantially higher failure rate than that of patients who had right hemisphere stroke (25% for acquisition and 23% for delayed recall). Performance on WAIS-R Digit Span, a measure of attention and working memory, was related to EPAT performance for the right but not the left hemisphere stroke patients, suggesting an attentional basis to poor EPAT test performance in the right hemisphere stroke group. Unfortunately, data were unavailable to determine whether there was a higher rate of neglect in those right CVA with attentional impairment who performed poorly on the EPAT.

Factor Analysis and the Construct Validity of Neuropsychological Tests

Factor analysis, when used appropriately, can be a powerful tool for determination of the construct validity of neuropsychological test procedures (Delis, Jacobson, Bondi, Hamilton, & Salmon, 2003; Larrabee, 2003d). Construct validity refers to the degree to which a test is a valid measure of a hypothetical underlying construct. The goals of factor analysis are to summarize patterns of correlations among observed variables, reduce a larger number of observed variables into a smaller number of factors, provide an operational definition for an underlying process (e.g., memory) by using observed variables (i.e., memory test scores), and to test a theory about the nature of underlying processes (Tabachnick & Fidell, 2005). Factor analysis addresses this through statistical analysis of the pattern of intercorrelations among a set of variables. Variables that are intercorrelated with one another but relatively independent of other subsets of variables are combined into factors. The basic assumption is that tests loading on a particular factor (i.e., correlated with that factor) are explained by the underlying factor. For example, if a test is truly a measure of the construct of verbal memory, then it should load on a factor defined by other tests known to be measures of verbal memory, characterizing a factor that is distinct from other underlying factors such as verbal intelligence or attention, otherwise the verbal memory test is nothing more than another way of measuring verbal intelligence or attention.

Since factor analysis derives from analyses of correlations or covariances, the results of a factor analysis can be distorted by effects of method variance, which occur when multiple scores derived from the same test are included in the same factor analysis,

thereby weighting that test multiple times. Although at least two scores representative of an underlying factor are needed to define that factor, these scores should be derived from independent tests; otherwise, a spurious factor can occur. A common error here is including both immediate and delayed recall scores for tests such as Logical Memory and Visual Reproduction in the same factor analysis. Factor solutions can also be distorted by insufficient representation of tests that are expected to identify underlying factors.

A good example of these issues is the factor analysis conducted by Brown, Roth, Saykin, and Beverly-Gibson (2007) in an attempt to demonstrate the construct validity of the Brown Location Test (BLT), a newly designed measure of visual memory. Brown et al. included eight scores from the BLT, six scores from the CVLT, plus WASI (or WAIS-III) T scores for Vocabulary and Matrix Reasoning, Full Scale IQ, and a score from a visual cancellation test, in the same factor analysis. Not surprisingly, they obtained a visual memory factor defined by the eight BLT scores, a verbal memory factor, defined by the six CVLT scores, and an IQ factor defined by Vocabulary, Matrix Reasoning, and IQ (which is comprised of both Vocabulary and Matrix Reasoning), whereas the visual scanning test did not load on any factor. This is clearly a spurious factor analytic result. Rather, at least two verbal intelligence subtests (e.g., Vocabulary and Similarities), two visual intelligence subtests (e.g., Matrix Reasoning and Block Design), two working memory subtests (Digit Span and Arithmetic), two processing speed measures (the visual scanning test they used, plus Digit Symbol) should have been included, as well as additional measures of verbal memory (Logical Memory), and visual memory (Visual Reproduction), with learning and retention scores included in separate, independent factor analyses (Larrabee, 2003d; Larrabee & Curtiss, 1995).

In the following section, I review the results of factor analyses that have included sufficient tests to define multiple domains of abilities, while minimizing the effects of method variance. The names I have chosen for the factors are based as much as possible on descriptions common to neuropsychologists and characteristic of past factor descriptors. I have relied as well on cognitive neuropsychological descriptions of constructs, in particular, by avoiding use of the term “cognitive,” since cognition is a general term that applies to multiple mental processes, such as verbal symbolic processes, perception, attention, and memory (Purves et al. 2008).

Factor analyses of neuropsychological test batteries (Holdnack, Zhou, Larrabee, Millis, & Salthouse, 2011; Larrabee, 2000; Larrabee & Curtiss, 1992, 1995; Leonberger, Nicks, Larrabee, & Goldfader, 1992; Tulskey & Price, 2003) generally define six domains of function:

- (1) *Verbal symbolic abilities* (word definition such as Wechsler Adult Intelligence Scale-IV/WAIS-IV Vocabulary, Wechsler, 2008; word fluency such as Controlled Oral Word Association, Benton, Hamsher, et al., 1994; verbal concept formation such as WAIS-IV Similarities).
- (2) *Visuoperceptual and visuospatial judgment and problem solving* including tests such as Facial Recognition and Line Orientation (Benton, Sivan, et al., 1994), WAIS-IV Visual Puzzles, Block Design, Matrix Reasoning (Wechsler, 2008).
- (3) *Sensorimotor function* (Finger Tapping, Reitan, & Wolfson, 1993; Grooved Pegboard, Heaton et al., 2004; Finger Localization and Tactile Form Perception, Benton, Sivan, et al., 1994; these tests have had limited investigation in factor analysis, with loadings of Grooved Pegboard and Purdue Pegboard on a *visual* factor, along with Benton Tactile Form Perception and WAIS-R *visuoperceptual visuospatial* tests (Block Design, Object Assembly), with a separate *motor* factor on which Finger Tapping and Grip Strength load (Larrabee & Curtiss, 1992, see Larrabee, 2000). Carroll (1993) considers tasks such as strength of grip, tapping speed, and fine manual dexterity as being psychomotor abilities (also see Frazier, Youngstrom, Chelune, Naugle, & Lineweaver, 2004, who found that Finger Tapping and the Grooved Pegboard loaded on a *processing speed* factor),
- (4) *Attention/working memory* (Wechsler Adult Intelligence Scale-IV/WAIS-IV Digit Span, Arithmetic, Letter–Number Sequencing, Wechsler, 2008; Wechsler Memory Scale-IV/WMS-IV Symbol Span, Wechsler, 2009),
- (5) *Processing speed* (Trail Making Test, Reitan & Wolfson, 1993; WAIS-IV Symbol Search and Coding [Digit Symbol], Wechsler, 2008),
- (6) *Learning and memory-verbal* (WMS-IV Logical Memory, Wechsler, 2009; CVLT-II, Delis et al., 2000) and *learning and memory-visual* (WMS-IV Visual Reproduction; CVMT; Trahan & Larrabee, 1988). Note that combined rather than separate *verbal* and *visual learning and memory* factors have been found, with other evidence for separate rather than combined factors (Holdnack et al., 2011).

Some of these tests do show cross-loadings (i.e., on more than one factor), for example, Controlled Oral Word Association loading with *processing speed* (Larrabee, 2000), and WAIS-IV Arithmetic loading with *verbal symbolic abilities* (Holdnack et al., 2011). Tests described as measures of *executive function* (Lezak et al., 2012) typically show loadings on factors of *processing speed* (Trail Making B; Controlled Oral Word Association), *working memory* (Letter–Number Sequencing), or *visuoperceptual visuospatial problem solving* (Category Test; Wisconsin Card Sorting Test, WCST; see Larrabee, 2000; Leonberger et al., 1992), rather than on

a unique *executive function* factor (also see Barbey, Colom, & Grafman, 2013; Keifer & Tranel, 2013; Salthouse, 2005; for the relationship of tests of executive function to tests of problem solving, general intelligence, and processing speed).

Achievement testing is common in neuropsychological assessment, for primary assessment of learning disability. The WRAT/WRAT-R/WRAT-3 (Jastak & Wilkinson, 1984; Wilkinson, 1993) was the 18th most commonly used test in the Rabin and colleagues (2005) test use survey. The WRAT versions do not represent comprehensive achievement test batteries but can serve as a useful screen, in combination with clinical history, for the presence of learning disability. Additionally, the Reading subtest provides a quick assessment of reading ability for administration of the MMPI-2-RF. As noted previously, the WRAT-3 Arithmetic subtest is a significant predictor of financial capacity in the elderly (Sherrod et al., 2009). In the Larrabee and Curtiss (1992) factor analysis (see Larrabee, 2000), the WRAT-R subtests did not form a separate *achievement* factor; rather they showed primary loadings on the *verbal symbolic* factor, with secondary loadings on an *attention/concentration* factor.

The factor structure of collections of neuropsychological tests appears to be invariant of age over the adult years (Crook & Larrabee, 1988; Larrabee & Curtiss, 1995), with the exception of failure to differentiate between the Perceptual Organization and Processing Speed indices of the WAIS-III in the very old (Wechsler, 1997c), although this was not found with the WAIS-IV (Wechsler, 2008). Salthouse and Saklofske (2010) have also found that the WAIS-IV subtests measure the same aspects of cognitive functioning in adults under and over age 65. Overall, it appears that the same constructs are identified over the adult age range, particularly in normal adults. Of course, factor analysis based on a moderately demented sample of individuals would not be expected to generate the same factor structure as a factor analysis based on the performance of healthy, age, and education-matched peers, due to floor effects in the dementia sample.

The six factors/domains of performance identified in this review show significant similarities to the broad abilities identified by Carroll (1993) and the Cattell-Horn broad abilities (McGrew, 2009). Although the Cattell-Horn-Carroll (CHC) model of intelligence grew out of educational psychology research, the 10 broad abilities identified by McGrew (2009) as representative of the CHC model, map fairly closely with the six neuropsychological domains reviewed previously, with some modification and collapsing of CHC broad abilities into the six neuropsychological domains. For example, CHC *fluid reasoning* is related to the *visuo-perceptual and visuospatial judgment and problem-solving* domain, CHC *comprehension knowledge* to the *verbal symbolic* domain, CHC *short-term memory* is related to the *attention/working memory* domain, CHC *visual processing* to the *visuo-perceptual and visuospatial judgment and problem-solving* domain, CHC *auditory processing* (speech sound discrimination, musical discrimination, and judgment) appears related to the *attention/working memory* domain, CHC *long-term storage and retrieval* is related to the *learning and memory* domain, and CHC *processing speed* (as well as *reaction and decision speed*) to the *neuropsychological processing speed* domain. As noted in the review of the neuropsychological factor analyses, CHC *reading and writing* and *quantitative knowledge* would be expected to show a primary association with the *verbal symbolic* domain, with a secondary association with *attention/working memory*.

Performance Validity and Symptom Validity

In the context of external incentives, the most valid test instruments can yield totally invalid data secondary to invalid performance by an examinee that does not provide an accurate measure of their actual level of ability. Invalid performance as detected by PVTs can obscure expected relationships between severity of neurological insult and test performance, for example, Green (2007) found no difference in CVLT performance comparing TBI patients with or without CT abnormalities, until those failing a PVT were excluded; Green, Rohling, Iverson, and Gervais (2003) did not find the expected dose–response relationship between admission Glasgow Coma Scale and olfactory identification, until those TBI subjects failing a PVT were excluded. Invalid performance can also result in spurious associations between symptom complaint and test performance, as demonstrated by Gervais, Ben-Porath, Wygant, and Green (2008), who only found a correlation between memory complaints and performance on the CVLT in persons failing a PVT; the correlation disappeared in subjects passing the PVT. Rohling and colleagues (2011) provide other examples of the effect of invalid test performance on attenuation of expected predictor criterion relationships in neuropsychological research.

Malingering is defined as the intentional fabrication and/or exaggeration of symptoms and deficits, in the context of external incentive such as financial gain in civil litigation, or avoidance of prosecution in criminal settings (DSM-V; American Psychiatric Association, 2013), and occurs commonly, with estimated frequencies up to 40% for litigating mild TBI (Larrabee, 2003a; Mittenberg, Patton, Canyock, & Condit, 2002), 54.3% for criminal defendants (Ardolf, Denney, & Houston, 2007), and 45.8% in Social Security Disability applicants (Chafetz, 2008). Due to these substantial frequencies of invalid performance, it is essential that assessment of performance validity is built into any neuropsychological test battery.

I will not be reviewing the diagnostic criteria for malingering proposed by Slick, Sherman, and Iverson (1999), other than to note that these criteria were important for being the first proposed criteria to objectively define the diagnosis of malingering of neurocognitive dysfunction. This has led to criterion groups research designs, which along with simulation studies resulted in

the development of both stand-alone and embedded/derived PVTs. This research is reviewed in Boone (2007, 2013), Larrabee (2007), and Morgan and Sweet (2009). I also will not be reviewing the area of symptom validity tests (SVTs; Larrabee, 2012a), which allow assessment of whether an examinee is giving an accurate report of actual symptom experience on pain scales (Larrabee, 2003b) or on self-report omnibus personality tests such as the MMPI-2 (Larrabee, 2003c) or MMPI-2-RF (Tellegen & Ben-Porath, 2008/2011). The reader should note, however, that the MMPI-2-RF includes several SVTs including F-r, Fp-r for evaluation of exaggeration of severe psychopathology, and Fs, FBS-r and RBS for assessment of exaggeration of injury, illness, and cognitive complaints (Ben-Porath, 2012; also see Wygant et al., 2007). The following discussion is focused on embedded/derived PVTs (Larrabee, 2012a).

Objective performance cutoffs on PVTs are determined to discriminate either non-injured persons dissimulating impairment or persons diagnosed as definite or probable malingerers (based on Slick et al., 1999 criteria) from non-litigating patients with moderate/severe TBI, depression, and other psychiatric, neurologic, or developmental conditions (Boone, 2007; Larrabee, 2007, 2012b). Typically, these cutoffs are set such that 90% or more of the non-litigating, bona fide clinical groups are classified as non-malingering (i.e., the false-positive rate is 10% or less). Moreover, in normally motivated clinical patients without any obvious external incentives, scores on free-standing PVTs are uncorrelated or weakly correlated due to performance at ceiling. Consequently, the chance of multiple scores exceeding cutoff representing a “false-positive” diagnosis of malingering is actually small.

Relying on multiple PVT and SVT failures improves the diagnosis of malingering and/or determination of invalid performance by improving sensitivity without substantially altering specificity (Larrabee, 2003a, 2014; Victor, Boone, Serpa, Buehler, & Ziegler, 2009). Larrabee (2008) demonstrated that with failure of two independent PVTs, each with a sensitivity of 0.50 and specificity of 0.90, the posterior probability of malingering using chained likelihood ratios and a base rate of .40 was .94; adding failure of a third independent PVT with the same sensitivity and specificity yielded a posterior probability of .99.

PVT procedures have been derived from performance patterns on standard neuropsychological measures of perception, motor function, attention, processing speed, memory, and problem-solving. These procedures are extensively reviewed in Boone (2007, 2013), Larrabee (2007), and Morgan and Sweet (2009). Performance on PVTs derived from these standard neuropsychological tests is atypical for bona fide clinical disorder. This can manifest as inconsistent patterns of performance, such as better performance on fine motor as opposed to gross motor tasks (Greiffenstein, Baker, & Gola, 1996), better performance on memory in comparison to attention (Mittenberg, Azrin, Millsaps, & Heilbronner, 1993), better performance on WAIS-R Vocabulary than Digit Span (Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995), better performance on recall relative to recognition on verbal memory tasks (Millis, Putnam, Adams, & Ricker, 1995), and production of errors that are not typical for neurologically impaired patients, such as excessive failure-to-maintain set errors on the WCST (Suhr & Boyer, 1999). Atypical performance can also manifest as scores that are excessively impaired such that they are rarely found in patients with moderate/severe TBI, including abnormally poor performance on the Benton, Sivan, et al. (1994) Visual Form Discrimination (Larrabee, 2003a), Finger Tapping (Arnold et al. 2005; Larrabee, 2003a), Digit Span (Babikian, Boone, Lu, & Arnold, 2006), or Reliable Digit Span (RDS; Greiffenstein, Baker, & Gola, 1994).

The above atypical patterns of performance can be captured by single scores, or by use of empirically derived statistical formulas via discriminant function analysis (Mittenberg et al., 1993; 1995) or logistic regression and Bayesian Model Averaging (Millis & Volinsky, 2001). At present, the literature is sufficiently developed to define PVTs for many common measures of core neuropsychological abilities, which will be reviewed in the next section on construction of a core battery. Failure of multiple PVTs and SVTs does not automatically equate to malingering, as there must be an external incentive present, with no other viable explanation for failure such as severe neurologic, psychiatric, or developmental disorders that often require a supervised living setting (Boone, 2007, 2013; Larrabee, 2007). Regardless of whether there is an external incentive, multiple PVT and SVT failure does call into question the validity of findings on the entire battery, such that poor performances are more likely the result of intentional underperformance, while normal range scores themselves may reflect underestimates of actual ability.

Constructing a Core Neuropsychological Battery for Adults

Proof of Concept

Investigations supporting the development of a core for an AFB include the research of Larrabee et al. (2008), Miller et al. (2010), and Rohling et al. (2003). Larrabee et al. (2008) have demonstrated that an AFB comprised measures of language (H-Words; timed generation of words beginning with the letter H), fine motor skill (Grooved Pegboard), working memory (WAIS-R Arithmetic), processing speed (WAIS-R Digit Symbol), verbal and visual memory (Wechsler Memory Scale delayed Logical Memory and delayed Visual Reproduction), verbal intelligence (WAIS-R Similarities), and visual intelligence (WAIS-R Block Design) generated an ROC AUC of 0.86, compared with an AUC of 0.83 for the seven primary scores of the HRB

(Category Test, TPT Total Time, Memory and Location, Finger Tapping, Seashore Rhythm, and Speech Sounds Perception) for discrimination of neurologically normal patients from patients with a variety of neurological disorders. Logistic regression with Bayesian Model Averaging of the Ability-Focused subtests, primary HRB scores, and Trail Making B selected four tests as consistent discriminators: H-Words, Trail Making B, the Grooved Pegboard, and Finger Tapping. The Grooved Pegboard had the largest Cohen's d , 1.08, compared with *any* other neuropsychological test.

Subsequent to this investigation, Miller and colleagues (2010) evaluated the diagnostic discrimination of a group of brain injured subjects (primarily suffering TBI, stroke, or dementia) from subjects who had cognitive complaints but no evidence for acquired neurological dysfunction (a “pseudoneurologic” control group), using an AFB covering five domains: language/verbal reasoning, visual-spatial reasoning, attention, processing speed and memory, using WAIS-III domain scores and select measures of neuropsychological function such as the CVLT-II and Trail Making Test. ROC AUC was 0.89 based on the five domains, and 0.88 based on an average of the five domain scores. Based on processing speed and memory alone, the ROC AUC was 0.90.

Importantly, Rohling and colleagues (2003) have demonstrated that an AFB (the MNB, Meyers & Rohling, 2004; Vollbrecht, Meyers, & Kaster-Bundgaard, 2000) based on individually normed tests (computed using published norms for individual tests that were statistically adjusted using regression analyses based on data from independent clinical patients and normal subjects to smooth the norms for effects of age, education, handedness, and gender) yielded essentially identical T scores of impairment in association with severity of TBI as did a co-normed HRB augmented with measures of learning and memory. Using five groups of TBI severity ranging from <1 h TFCs up to 14–28 days TFC, the within group correlation of the OTBM with TBI severity was 0.99 for the MNB and 0.96 for the co-normed HRB, with essentially identical slopes, -2.6 (MNB) and -3.1 (HRB), and intercepts, 47.0 (MNB) and 48.1 (HRB). The OTBM, collapsed over the five levels of TBI severity was $T = 39.2$ for the MNB and $T = 38.9$ for the HRB, with a correlation of 0.97 between the MNB and HRB OTBMs associated with each of the five severity levels of TBI.

The study of Rohling and colleagues (2003) is important for it not only shows the equivalent sensitivity of a non-HRB battery to an augmented HRB battery but also shows equivalency of a battery of individually normed tests to a co-normed battery. Regarding this latter point, this equivalency depends upon adequately normed individual tests, with appropriate corrections for important demographic factors such as age, education, and sex, when such corrections are necessary. Other investigations support this point by demonstrating essentially equivalent results when a common dataset is scored using either the Heaton and colleagues (2004) norms or the meta-analytically derived norms published by Mitrushina, Boone, Razani, and D'Elia (2005) (Hill, Boettcher, et al., 2013; Rohling, Axelrod, & Wall, 2008). There is also comparability for the Mitrushina and colleagues (2005) norms, the norms comprising the Meyers MNB, and the Heaton and colleagues (2004) norms when all three normative sets are used to score common test procedures (M. L. Rohling, personal communication, February 5, 2014).

Of course, widespread adoption of a core battery would allow for co-norming, on a large-scale basis, providing a data source preferable to individually normed tests. However, the striking similarity of results based on individually normed compared with co-normed tests reported by Rohling and colleagues (2003), does support relying upon an aggregated set of norms, pending development of co-normed test procedures. Moreover, subjecting data from individually normed tests to the statistical analyses proposed by Rohling, Miller, and Langhinrichsen-Rohling (2004) for aggregation of test scores into domains of ability, computation of an OTBM, and comparing these to one another and to estimated premorbid level of ability, can further enhance interpretation of ability-focused neuropsychological assessment based on individually normed tests.

General Issues in Constructing the Battery

Construction of a core battery requires that measures be selected for each of the core neuropsychological domains, supported by factor analysis, including first, *verbal symbolic abilities*; secondly, *visuoperceptual and visuospatial judgment and problem solving*; thirdly, *sensorimotor skills*; fourthly, *attention/working memory*, fifthly, *processing speed*; finally, *learning and memory*. Certain domains may also require specification of verbally mediated and visually mediated abilities, including *attention/working memory*, *processing speed*, and *learning and memory*. Selection of procedures for each domain should include, at minimum, tests clearly representative of the domain, with additional evidence of other indicia of validity such as sensitivity to presence of brain dysfunction, and/or sensitivity to severity of dysfunction, and/or prediction of important activities of daily living. It is not assumed that each test assigned to a domain include evidence for all three areas of validity nor that every test included contain an embedded/derived measure of performance validity. For example, the AVLT might be included because, in addition to being a good representation of verbal learning and memory, it is sensitive to presence of brain injury or dysfunction, and contains embedded/derived PVTs. Sufficient data currently exist to conduct meta-analyses of various candidate tests, particularly in relation to evidence for sensitivity to the neuropsychological effects of TBI and AD.

If one were to start this project *de novo*, each domain of function (*verbal symbolic* abilities, *sensorimotor*, etc.) would be over-sampled, with procedures administered to persons with AD or TBI. These two subject groups are important for validation, because both are widely seen by neuropsychologists, and both have validated means for assessing severity of dysfunction/impairment (Glasgow Coma Scale, TFCs for TBI; CDR ratings for Alzheimer's Disease), that are independent of neuropsychological test performance. Moreover, the TBI group could yield sub-samples who also have unilateral mass lesions allowing investigation of lateralized neuropsychological effects (see [Levin, Benton, & Grossman, 1982](#), Fig. 5-5, p. 112). Within the TBI group, each measure could be contrasted for ability to discriminate moderate and severe TBI from normal subjects, as well as for correlation of each measure with severity of trauma, defined by GCS, TFCs, and duration of Post-Traumatic Amnesia. Correlation of test performance with rating scales of adaptive function (DRS, [Rappaport, Hall, Hopkins, Belleza, & Cope, 1982](#); Mayo-Portland Adaptability Inventory, [Malec et al., 2003](#)) should be analyzed for TBI, as well as for AD (CDR, [Hughes et al., 1982](#); also see review of various measures of basic and instrumental activities of daily living by [Loewenstein & Mogosky, 1999](#)).

Once the tests are identified for the core battery, these could be compared against sub-batteries developed for patients with left and right CVA. These sub-batteries of specialized measures of language dysfunction (e.g., Multilingual Aphasia Examination, [Benton, Hamsher, et al., 1994](#)), and spatial ability would be constructed for persons suffering left and right cerebrovascular accidents, with particular attention paid to the moderating effects of auditory comprehension impairment and neglect. These sub-batteries would then be compared with see which subtests discriminate patients with right versus left CVA on the basis of lateralized neuropsychological deficits, a task which could also employ various measures of gross and fine motor skill and tactual/perceptual skills to determine which of these measures are best for group discrimination.

The measures developed for the sub-batteries would then be administered along with the core battery procedures to explore interrelationships and contingencies of performance. Hence, it could be determined that a left CVA patient with normal WAIS-IV Vocabulary, Controlled Oral Word Association and Animal Naming would not need to be administered the Boston Naming Test or be evaluated further for language impairment, and that same patient who had normal Block Design, would not have to be administered the Line Orientation test.

Each of the core areas of function should be represented by a minimum of three measures. This is notwithstanding the research of [Donders and Axelrod \(2002\)](#), who found that the WAIS-III Verbal Comprehension, and Working Memory indices could be adequately measured by two rather than three measures (Processing Speed is already measured by two indicators). Although two tests per domain are a minimal requirement, a stronger argument can be made for at least three measures, both for yielding extractions of an underlying factor ([Carroll, 1993](#)) as well as optimally defining a reliable measure of the domain ([Rohling et al., 2004](#)). Moreover, requiring at least three tests per domain also allows for potential increase in variability of performance, allowing for analysis of intra-individual variability (IIV). Increases in IIV have been related both to acquired neuropsychological impairment, as well as to the presence of invalid test performance ([Hill, Rohling, Boettcher, & Meyers, 2013](#)).

Candidates for the Core Battery

The following section of this paper considers candidates for each of the core domains of ability. As noted, selection of procedures should be guided by factor analytic data, a minimum of three tests per domain, including measures that include evidence of at least one of the following features: first, sensitive to presence of disorder; secondly, sensitive to severity of disorder; thirdly, showing a predictive relationship to important basic and complex activities of daily living; finally, possessing embedded or derived measures of performance validity. Obviously, if two tests show relatively identical sensitivity to both presence and severity of disorder, but one also shows correlation with instrumental activities of daily living and contains an embedded/derived PVT, the test addressing more validity indicators would be given preference in the core battery.

Verbal Symbolic Ability

Controlled Oral Word Association ([Benton, Hamsher, et al., 1994](#); a phonemic fluency task requiring rapid production of words beginning with specific letters of the alphabet) and Semantic Category Fluency (rapid generation of words from a semantic category, such as animals) are good candidates, due to the ubiquitous nature of word-finding impairment in aphasic conditions, and the ability to use dissociations in performance between phonemic and semantic abilities in differential diagnosis of AD ([Salmon, Heindel, & Lange, 1999](#)). Controlled Oral Word Association is also associated with medical decision-making capacity in AD ([Marson et al., 1995](#)), and semantic category fluency discriminated patients with MCI from normal elderly, as well as from mild AD ([Griffith et al., 2006](#)). Additional *Verbal Symbolic* ability candidates include WAIS-IV Vocabulary, Information, and Similarities subtests. Comparisons of Vocabulary with Digit Span can yield information about symptom validity ([Mittenberg et al., 1995](#)). The Similarities subtest is one of the more sensitive WAIS verbal subtests to brain dysfunction ([Loring & Larrabee, 2006](#)), although the WAIS-IV technical and interpretive manual reports a larger Cohen's *d* for the discrimination of

TBI, AD, and MCI from normative subjects for the Information subtest contrasted with both Similarities and Vocabulary (Wechsler, 2008). Consideration could also be given to including a visual confrontation naming test, such as the Boston Naming Test (Kaplan et al., 1983), given the loading of the Benton Visual Naming Test (Benton, Hamsher, et al., 1994) on a *verbal symbolic* factor (Larrabee & Curtiss, 1992; Larrabee, 2000), with a similar finding reported for Boston Naming by Frazier and colleagues (2004). Alternatively, empirical research might show that administration of visual confrontation naming is not necessary given the presence of measures of semantic and phonemic fluency, in addition to WAIS-IV subtests such as Similarities and Information (i.e., the construct of word retrieval is already sufficiently covered for purposes of a core battery). It is also important to consider including measures of academic achievement such as the Wide Range Achievement Test-IV (Wilkinson & Robertson, 2006) for two reasons: first, the aforementioned sensitivity of written computations to financial capacity (Sherod et al., 2009); second, ability to ascertain evidence suggestive of a premorbid learning disability. In the Larrabee and Curtiss (1992) factor analysis, all three Wide Range Achievement Test-Revised (WRAT-R; Jastak & Wilkinson, 1984) subtests (Spelling Reading and Arithmetic) showed primary loadings on a *verbal symbolic* factor, with secondary loadings on *attention/concentration* (see Larrabee, 2000). Moreover, oral reading tasks such as the WRAT-R Reading have been used to estimate premorbid level of function (though note that single word reading tasks employing irregularly spelled words that cannot be decoded phonetically, such as the Wechsler Test of Premorbid Function from the Advanced Clinical Solutions (ACS), Pearson, 2009, appear to be superior for this purpose compared with the WRAT-R; Lezak et al., 2012). Thus, the Reading and Arithmetic sections of the Wide Range Achievement Test-IV (Wilkinson & Robertson, 2006) should be considered for the *verbal symbolic* domain, as well.

Visuoperceptual and Visuospatial Judgment and Problem Solving

Candidates for this domain include WAIS-IV Visual Puzzles, Matrix Reasoning, and Block Design (one of the WAIS measures most sensitive to brain dysfunction; Loring & Larrabee, 2006; Russell & Starkey, 1993). The WAIS-IV technical and interpretive manual shows larger Cohen's *d* for Visual Puzzles compared with either Block Design or Matrix Reasoning for discriminating TBI from normative subjects, as well as for discrimination of AD from normative subjects. The older WAIS Performance IQ (containing Block Design as well as a processing speed test, Digit Symbol) was correlated with stability of employment following TBI (Machamer, Temkin, Fraser, Doctor, & Dikmen, 2005). As noted in the earlier review of factor analytic investigations, the WCST (Heaton, Chelune, Talley, Kay, & Curtiss, 1993) and the Category Test (Reitan & Wolfson, 1993) both load on a *visuoperceptual and visuospatial judgment and problem-solving* factor rather than defining a separate executive function factor (Leonberger et al., 1992; Larrabee, 2000), hence both would be considered candidates for this domain. Both have derived measures of performance validity (Greve, Bianchini, Mathias, Houston, & Crouch, 2002; Larrabee, 2003a; Suhr & Boyer, 1999; Tenhula & Sweet, 1996). The WCST categories achieved was significantly correlated with wages earned and hours worked in predicting work outcome in a sample of schizophrenics (McGurk, Mueser, Harvey, LaPuglia, & Marder, 2003). The Benton, Sivan, et al. (1994) Visual Form Discrimination is sensitive to invalid performance (Larrabee, 2003a). Line Orientation is sensitive to spatial impairment, and is not affected by auditory comprehension impairment in aphasics (Benton, Sivan, et al., 1994). Meyers, Galinsky, and Volbrecht (1999) have reported a cutting score for invalid performance for the Line Orientation Test, although the utility of this has been questioned (Iverson, 2001).

Sensorimotor Skills

Four motor procedures are candidates: Grip Strength, Finger Tapping, the Purdue Pegboard, and the Grooved Pegboard, although Lezak et al. (2012) observe that the Grooved Pegboard has gradually replaced the Purdue Pegboard in popularity of use over time. Larrabee et al. (2008) found that the Grooved Pegboard had the largest Cohen's *d*, 1.08, of any neuropsychological test in discriminating pseudoneurologic controls from brain dysfunction patients (the next largest *d* was 0.89 for Trail Making B). Greiffenstein et al. (1996) demonstrated how probable malingerers show the reverse pattern of declining gross to fine motor skill typical of neurologically based motor function impairment; that is, in the malingering group, the best performance was on the Grooved Pegboard, followed by Finger Tapping and Grip Strength, whereas the group who had bona fide upper motor neuron dysfunction showed the reverse pattern. Finger Tapping can also be analyzed for validity of performance (Arnold et al., 2005; Larrabee, 2003a). Separate assessment of tactile skills may not be necessary as part of a core battery (note that Benton Tactile Form Perception loaded on a visuospatial problem-solving factor, as did Grooved Pegboard and the Purdue Pegboard; Larrabee & Curtiss, 1992; see Larrabee, 2000), but could be a supplemental consideration in unilateral stroke cases.

Attention/Working Memory

Candidates for this domain include WAIS-IV Digit Span, Arithmetic, and Letter–Number Sequencing. Digit Span provides important information relative to performance validity (Jasinski, Berry, Shandera, & Clark, 2011) and is correlated with financial

capacity in AD (Earnst et al., 2001). Letter–Number Sequencing is sensitive to effects of TBI (Donders et al., 2001) as well as correlated with financial capacity in AD (Earnst et al., 2001). WMS-IV visual working memory tasks can also be considered. Although there is a theoretical rationale for including separate verbal and visual working memory tests (phonological loop vs. visuospatial sketchpad; Baddeley, 2007), the clinical utility of separate modality-specific working memory tasks remains to be demonstrated. The Spatial Addition subtest of the WMS-IV, one of two measures comprising the WMS-IV Visual Working Memory Index, is not normed for persons older than 69. The same is true for Letter–Number Sequencing on the WAIS-IV; however, normative data are available to age 89 for the slightly different Letter–Number Sequencing task on the WMS-III.

Processing Speed

Candidates for this domain include WAIS-IV Symbol Search and Coding (Digit Symbol), and the Trail Making Test. Digit Symbol is sensitive to both presence and severity of the effects of TBI (Dikmen et al., 1995) and sensitive to presence and severity of AD (Larrabee, Lengen, et al., 1985). Trail Making B is sensitive to residual effects of TBI at 1 year post-trauma (Dikmen et al., 1995) and was one of the most sensitive discriminators of neurologic from non-neurologic patients (Larrabee et al., 2008). In the WAIS-IV technical and interpretation manual, Symbol Search yields larger *d* values for both TBI, 1.13, and AD, 1.64 than the values of 0.73 for the TBI contrast, and 1.41 for the AD contrast using Coding (Digit Symbol). Trail Making B is correlated with driving ability in AD (Whelihan et al., 2005) and TBI (Novack et al., 2006). Both Trail Making B and Digit Symbol were correlated with amount of time worked since TBI (Machamer et al., 2005). The Stroop (Golden, 1978; Trenerry, Crosson, DeBoe, & Leber, 1989) and the Symbol Digit Modalities Test (Smith, 1983) are additional candidates for the processing speed domain, and might define a separate verbal modality of processing speed, distinct from the visuomotor speed demands of Trail Making and the WAIS-IV processing speed tasks.

Learning and Memory

Although I have listed a single domain for *learning and memory*, separation of this domain into sub-domains of *verbal* and *visual learning and memory* is supported, both clinically and psychometrically. For *verbal learning and memory*, both the CVLT-II (Delis et al., 2000) and Rey AVLT (Rey, 1964; Schmidt, 1996) are candidates for a measure of verbal supraspan learning. Both are sensitive to effects of TBI and AD on memory function (Delis et al., 2000; Jacobs & Donders, 2007; Schmidt, 1996) and both have embedded and/or derived measures of performance validity (Barrash, Suhr, & Manzel, 2004; Boone, Lu, & Wen, 2005; Davis, Millis, & Axelrod, 2012; Meyers & Volbrecht, 2003; Wolfe et al., 2010). As discussed earlier in this paper, Loring and colleagues (2008) found a much larger effect size for discrimination of right versus left temporal lobe epilepsy for the AVLT in comparison to the CVLT (the original CVLT). The original CVLT was correlated with hours worked in schizophrenics able to return to work (McGurk et al., 2003; Evans et al., 2004). The Hopkins Verbal Learning Test-Revised (HVLT-R; Brandt & Benedict, 2001) is another candidate, although the use of 12 items in three categories may make it too easy for younger subjects, in contrast to the CVLT or AVLT. Additionally, there is no measure of performance validity that has been derived for the HVLT-R. Other paradigms for evaluation of verbal memory include text recall and paired associate learning, as contained within the various editions of the Wechsler Memory Scale. Both Logical Memory and Paired Associate Learning have derived measures of performance validity on the WMS-III (Killgore & DellaPietra, 2000; Langeluddecke & Lucas, 2003). On the WMS-IV, Logical Memory Delayed Recognition and Verbal Paired Associate Delayed Recognition yield measures of performance validity (ACS, Pearson, 2009), but can only be administered to persons up to age 69. The test stimuli and administration procedure for Logical Memory and Verbal Paired Associates changes at age 70.

Candidates for *visual learning and memory* include the Rey Complex Figure Test (Meyers & Meyers, 1995; Osterrieth, 1944; Rey, 1941) and CVMT (Trahan & Larrabee, 1988), both of which have demonstrated sensitivity to effects of TBI and AD (Lezak et al., 2012; Strauss, Spreen, & Sherman, 2006; Trahan & Larrabee, 1988), and stroke (Lezak et al., 2012; Strauss et al., 2006; Trahan, Larrabee, & Quintana, 1990) and both of which have derived measures of performance validity (Larrabee, 2009; Lu, Boone, Cozolino, & Mitchell, 2003). Another candidate is the Brief Visuospatial Memory Test-Revised (BVM-T-R; Benedict, 1997) which has the advantage of multiple parallel forms, but does not include a derived performance validity measure. The WMS-IV Visual Reproduction I and II also includes a delayed recognition trial that is useful in evaluating performance validity, particularly when combined with the ACS Word Choice Test, RDS, Logical Memory Recognition, and Verbal Paired Associates Recognition (ACS; Pearson, 2009). The Designs subtest of the WMS-IV is a test that is new to the WMS-IV without much independent supporting research at the present time. Additionally, the Designs subtest cannot be administered to subjects older than age 69. Important considerations in the *visual learning and memory* domain include the confounding effects of constructional and spatial skills in performing design reproduction from memory tasks, resulting in higher factor loadings on *visuoperceptual and visuospatial judgment and problem solving* than occur on either a general *learning and memory* or *visual learning and memory*

factor for immediate reproduction scores, a pattern which reverses in the delayed recall format (Larrabee, Kane, Schuck, & Francis, 1985; Larrabee & Curtiss, 1995). Loadings suggesting a visuospatial confound do not appear to occur for measures of visual recognition memory such as the CVMT (Larrabee & Curtiss, 1995), which suggests the advisability of limiting selection to only one design reproduction from memory test for the *visual learning and memory* domain. A final consideration is that most visual memory tests such as the Rey Complex Figure and the CVMT employ abstract visual geometric patterns. Consequently, test selection should also consider procedures using meaningful stimuli, such as the recurring familiar figures comprising the Continuous Recognition Memory Test (Hannay, Levin, & Grossman, 1979), which also contains an embedded/derived PVT (Larrabee, 2009). Brown and colleagues (2007) have published a test of visual location learning and memory, the BLT, requiring memory for location of colored tokens placed on a grid, including five learning trials, short- and long-delay free recall and a delayed recognition trial. Normative data are available for ages 17–88, and performance is significantly poorer for right compared with left temporal lobectomy (Brown et al., 2010). This novel test is of interest, given the non-verbalizable stimuli, not requiring a drawing response, presented in a format similar to the CVLT and AVLT, including a recognition trial that may lead, with subsequent research, to a derived PVT which presently does not exist. The BLT is also available in two alternate forms.

Hypothetical AFB

Per the above, a hypothetical AFB could include: first, *verbal symbolic ability*: Controlled Oral Word Association, Animal Naming, WAIS-IV Information and Similarities, WRAT-IV Reading and Arithmetic; secondly, *visuo perceptual and visuospatial judgment and problem solving*: Benton Visual Form Discrimination, WAIS-IV Block Design, Visual Puzzles, WCST; thirdly, *sensorimotor skills*: Grip Strength, Finger Tapping, Grooved Pegboard; fourthly, *attention/working memory*: WAIS-IV Digit Span, Arithmetic, Letter–Number Sequencing, WMS-IV Symbol Span; fifthly, *processing speed*: Trail Making Test, WAIS-IV Symbol Search, Coding, and the Stroop; finally, *learning and memory verbal*: the AVLT, WMS-IV Logical Memory and Verbal Paired Associates; *learning and memory visual*: WMS-IV Visual Reproduction, CVMT, Hannay and colleagues (1979) Continuous Recognition Memory.

This hypothetical AFB would contain 27 measures (11 of which each require 5 min or less to administer, with a total estimated time of 4.5 h), and include 10 embedded/derived PVTs. This compares with 34 measures if one were to administer all of the tests comprising the Heaton and colleagues (2004) normative data (23), plus all of the WAIS-R subtests (11) in this data base, and 36 tests if the entire NAB is administered. The Meyers MNB (Meyers & Rohling, 2004) contains 22 measures, with 11 PVTs, but uses single tests to represent motor and tactile ability, verbal and visual memory, and test selection was not guided by the validity criteria proposed in the current review.

Once a core set of procedures is finalized for the AFB, additional research could establish a core screening battery, created either by selecting the most sensitive test per domain or by employing procedures such as logistic regression, which may define a screening battery based primarily upon measures of processing speed and memory, per the research of Larrabee and colleagues (2008) and Miller and colleagues (2010). If a patient screened negative for evidence of acquired impairment, there would be no need for additional assessment with procedures sensitive to severity of impairment or prediction of activities of daily living. Finally, anyone using the AFB should also administer free-standing PVTs, and conduct personality assessment per published practice guidelines (American Academy of Clinical Neuropsychology, 2007; Bush et al., 2005; Heilbronner et al., 2009).

The above core measures can be augmented by additional procedures in specialized populations (see Bauer, 2000, for further discussion of this approach). For example, the Paced Auditory Serial Addition Test (Gronwall, 1977) and Auditory Consonant Trigrams (Stuss, Stethem, Hugenholtz, & Richard, 1989) have shown utility for assessment of deficits following TBI, but are inappropriate due to difficulty level, for older persons being evaluated for suspected AD. Moreover, impairments on aspects of the core battery should trigger additional evaluation with specialized measures that yield further information on the impaired construct; for example, detailed language assessment should be performed for someone showing impaired phonemic and semantic category fluency; assessment of tactile perceptual skills should be conducted in a stroke patient showing unilateral impairment in motor skills.

Finally, the core battery that has been discussed is open to modification, should new and improved test procedures be developed for one of the core domains. As an example, consider the possibility that a new measure of verbal supraspan learning is developed to compete with the version already selected for the core battery. The two procedures could be compared by administering both with samples of moderate TBI, severe TBI, MCI, and AD, with a research design that allows for a small sample of control subjects obtained by examining relatives of the patients, so that the sensitivity of each test to presence and severity of disorder could be determined. A small dissimulation study could be conducted to determine if performance can differentiate non-injured simulators from the TBI, MCI, and AD patients. Additionally, the construct validity of the newly proposed test can be evaluated through use of multiple regression. Considering that the *verbal learning and memory* domain should consist of three subtests: first, supraspan list learning; secondly, text recall; finally, paired associate learning, the R^2 obtained by predicting performance on the original

supraspan learning by text recall and paired associate learning can be contrasted with the R^2 obtained by predicting the new supraspan learning task by the existing measures of text recall and paired associate learning. This analysis would address convergent validity. Discriminant validity could be established by comparing the R^2 obtained by predicting the original supraspan learning test by the test variables most representative of the remaining test domains (i.e., the highest loading variables for the five remaining factors). If the newly designed test shows a larger R^2 when predicted by text recall and paired associate learning than the original test, shows a lower R^2 when predicted by the highest loading subtest for each of the remaining test domains, plus shows greater sensitivity to the presence and severity of the effects of TBI and AD, and shows superior discrimination of feigned versus bona fide neuropsychological deficits, then serious consideration could be given to replacing the existing test with the new procedure.

Summary

A framework for developing a core neuropsychological battery is proposed, based on both test validity as well as incorporating embedded/derived measures of performance validity for six separate domains of performance: first, *verbal symbolic ability*; secondly, *visuoperceptual and visuospatial judgment and problem solving*; thirdly, *sensorimotor skills*; fourthly, *attention/working memory*; fifthly, *processing speed*; finally, *learning and memory (verbal and visual)*. It is recommended that each domain comprises at least three tests that are chosen on the basis of sensitivity to detection of the presence of disorder, sensitivity to severity of disorder, correlation with external criteria relevant to important basic and instrumental activities of daily living, including safely living independently, financial competence and driving a motor vehicle.

Select tests comprising the AFB may serve more than one purpose, for example, Trail Making B is very sensitive to detection of presence of disorder, severity of disorder, and correlates with external criteria such as driving a motor vehicle. The AVLT is sensitive to presence of disorder, and also yields performance validity information. Key clinical groups for derivation of the core battery include moderate-to-severe TBI (including subsets with unilateral mass lesions), and AD. Secondary groups also used in developing specialized sub-batteries, include subjects suffering left or right hemisphere stroke, to further elucidate the effects of lateralized cerebral dysfunction, aphasia with and without auditory comprehension deficit, and neglect, on core battery subtests and domains.

The proposed framework for battery development serves two purposes. First, the framework presents a psychometrically sound, evidenced-based rationale for battery composition for the current individual practitioner. Second, this approach presents guidelines to consider for development of a core battery for common use as might result from coordinated inter-organizational efforts of groups such as the National Academy of Neuropsychology, and Society for Clinical Neuropsychology of the American Psychological Association.

Determination and adoption of a core adult neuropsychological test battery has a primary advantage of allowing aggregation of data from multiple clinical sites, which can advance the interpretation of individual cases by yielding modal profiles for various clinical disorders, expanding on the work of Zakzanis et al. (1999). Accumulation of large data sets can lead to multivariable research investigations such as logistic regression analysis to contrast the neuropsychological performance of mild AD with the effects of Parkinson's disease. Aggregation across clinical sites can also lead to additional research on the test validity criteria considered in the current review, including measures sensitive to presence and severity of disorder, and predictive of activities of daily living. Such large data sets can also lead to additional factor analytic investigations of large-scale data sets specific to disease categories, including structural equation modeling (cf. Tabachnick & Fidell, 2005).

Finally, specification of a common core neuropsychological test battery containing embedded or derived PVTs can advance detection of invalid neuropsychological test performance by development of logistic regression equations that discriminate between probable malingerers and non-litigating patients with moderate–severe TBI, major depression, anxiety disorder, and other conditions relevant to differential diagnosis. At present, such determinations are based on aggregation of individually developed PVTs (Larrabee, 2008), referred to as a naïve Bayesian approach (Holdnack, Millis, Larrabee, & Iverson, 2013). Utilization of a common set of tests that include embedded/derived measures of performance validity allows for development of PVTs using logistic regression which has two advantages over the individually aggregated approach: first, logistic regression allows for variable intercorrelation, assumed to be negligible in valid-performance groups in the individually aggregated approach and secondly, logistic regression allows for differential weighting of salient variables, which are unit-weighted in the individually aggregated approach (Holdnack et al., 2013).

Conflict of Interest

GJL is a co-author of the Continuous Visual Memory Test (Trahan & Larrabee, 1988) and receive royalties from Psychological Assessment Resources for sales of this test. GJL is the editor of *Assessment of Malingered Neuropsychological Deficits* (2007) and

Forensic Neuropsychology: A Scientific Approach (2nd Ed.) (2012), and receives royalties from Oxford University Press for sales of these books.

Acknowledgements

This paper is based on lectures presented at the Vivian Smith International Neuropsychological Society Summer Institute, June, 2007, Xylocastro, Greece; the Houston Neuropsychological Society, October, 2009, Houston, TX; The International Academy of Applied Neuropsychology and Akademie bei König and Mueller, September, 2010, London, UK; Brooks Army Medical Center, January, 2011, San Antonio, TX; and Womack Army Medical Center, June, 2012, Ft. Bragg, NC.

References

- American Academy of Clinical Neuropsychology. (2007). American Academy of Clinical Neuropsychology (AACN) practice guidelines for neuropsychological assessment and consultation. *The Clinical Neuropsychologist*, 21, 209–231.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders-text revision* (5th ed.). Washington, DC: Author.
- Ardolf, B. R., Denney, R. L., & Houston, C. M. (2007). Base rates of negative response bias and malingered neurocognitive dysfunction among criminal defendants referred for neuropsychological evaluation. *The Clinical Neuropsychologist*, 20, 145–159.
- Arnold, G., Boone, K. B., Lu, P., Dean, A., Wen, J., Nitch, S., et al. (2005). Sensitivity and specificity of finger tapping test scores for the detection of suspect effort. *The Clinical Neuropsychologist*, 19 (1), 105–120.
- Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various digit span scores in the detection of suspect effort. *The Clinical Neuropsychologist*, 20, 145–159.
- Backman, L., Jones, S., Berger, A.-K., Laukka, E. J., & Small, B. J. (2005). Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology*, 19, 520–531.
- Baddeley, A. (2007). *Working memory, thought and action*. New York: Oxford.
- Barbey, A. K., Colom, R., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human intelligence. *Neuropsychologia*, 51, 1361–1369.
- Barrash, J., Suhr, J., & Manzel, K. (2004). Detecting poor effort and malingering with an expanded version of the Auditory Verbal Learning Test (AVLTX): Validation with clinical samples. *Journal of the International Neuropsychological Society*, 26, 125–140.
- Bauer, R. M. (2000). The flexible battery approach to neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 419–448). Mahwah, NJ: Lawrence Erlbaum Associates.
- Belanger, H. G., Curtiss, G., Demery, J. A., Lebowitz, B. K., & Vanderploeg, R. D. (2005). Factors moderating neuropsychological outcomes following mild traumatic brain injury: A meta-analysis. *Journal of the International Neuropsychological Society*, 11, 215–227.
- Benedict, R. H. B. (1997). *Brief visuospatial memory test-revised*. Odessa, FL: Psychological Assessment Resources.
- Ben-Porath, Y. S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis: University of Minnesota Press.
- Benton, A. L., Hamsher, K. de S., & Sivan, A. B. (1994). *Multilingual aphasia examination* (3rd ed.). Iowa City: AJA.
- Benton, A. L., Sivan, A. B., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1994). *Contributions to neuropsychological assessment. A clinical manual* (2nd ed.). New York: Oxford University Press.
- Binder, L. M., Rohling, M. L., & Larrabee, G. J. (1997). A review of mild head trauma. Part I: Meta-analytic review of neuropsychological studies. *Journal of Clinical and Experimental Neuropsychology*, 19, 421–431.
- Blessed, G., Tomlinson, B. F., & Roth, M. (1968). The association between quantitative measures of dementia and of senile change in the cerebral gray matter of elderly subjects. *British Journal of Psychiatry*, 114, 797–811.
- Boone, K. B. (Ed.). (2007). *Assessment of feigned cognitive impairment. A neuropsychological perspective*. New York: Guilford.
- Boone, K. B. (2013). *Clinical practice of forensic neuropsychology. An evidence-based approach*. New York: Guilford.
- Boone, K. B., Lu, P., & Wen, J. (2005). Comparisons of various RAVLT scores in the detection of noncredible memory performance. *Archives of Clinical Neuropsychology*, 20, 301–319.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley.
- Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new verbal memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5, 124–142.
- Brandt, J., & Benedict, R. H. B. (2001). *Hopkins verbal learning test-revised*. Odessa, FL: Psychological Assessment Resources, Inc.
- Brown, F. C., Roth, R. M., Saykin, A. J., & Beverly-Gibson, G. (2007). A new measure of visual location learning and memory: Development and psychometric properties for the Brown Location Test (BLT). *The Clinical Neuropsychologist*, 21, 811–825.
- Brown, F. C., Tuttle, E., Westerveld, M., Ferraro, F. R., Chmielewicz, T., Vandemore, M., et al. (2010). Visual memory in patients after anterior right temporal lobectomy and adult normative data for the Brown Location Test. *Epilepsy & Behavior*, 17, 215–220.
- Brown, L. B., Stern, R. A., Cahn-Weiner, D. A., Rogers, B., Messer, M. A., Lannon, M. C., et al. (2005). Driving Scenes Test of the Neuropsychological Assessment Battery (NAB) and on-road driving performance in aging and very mild dementia. *Archives of Clinical Neuropsychology*, 20, 209–215.
- Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior*, 12, 543–550.
- Bush, S. S., Ruff, R. M., Troster, A., Barth, J., Koffler, S. P., Pliskin, N. H., et al. (2005). NAN position paper: Symptom validity assessment: Practice issues and medical necessity. *Archives of Clinical Neuropsychology*, 20, 419–426.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University Press.
- Chafetz, M. (2008). Malingering on the Social Security consultative exam: Predictors and base rates. *The Clinical Neuropsychologist*, 22, 529–546.
- Chelune, G. J. (2010). Evidence-based research and practice in clinical neuropsychology. *The Clinical Neuropsychologist*, 24, 454–467.
- Christensen, H., Griffiths, K., Mackinnon, A., & Jacomb, P. (1997). A quantitative review of cognitive deficits in depression and Alzheimer-type dementia. *Journal of the International Neuropsychological Society*, 3, 631–651.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Crook, T. H., & Larrabee, G. J. (1988). Interrelationships among everyday memory tests: Stability of factor structure with age. *Neuropsychology*, 2, 1–12.
- Davis, J. J., Millis, S. R., & Axelrod, B. N. (2012). Derivation of an embedded Rey Auditory Verbal Learning Test performance validity indicator. *The Clinical Neuropsychologist*, 26, 1397–1408.
- Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M., & Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment. *Journal of the International Neuropsychological Society*, 9, 936–946.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California verbal learning test (CVLT): Adult version* (Research ed.). San Antonio: Psychological Corporation.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California verbal learning test, II*. San Antonio, TX: Psychological Corporation.
- Dikmen, S. S., Machamer, J. E., Winn, H. R., & Temkin, N. R. (1995). Neuropsychological outcome at 1-year post head injury. *Neuropsychology*, 9, 80–90.
- Dikmen, S. S., Temkin, N. R., Machamer, J. E., Holubkov, A. L., Fraser, R. T., & Winn, H. R. (1994). Employment following traumatic head injuries. *Archives of Neurology*, 51, 177–186.
- Donders, J., & Axelrod, B. N. (2002). Two-subtest estimations of WAIS-III factor index scores. *Psychological Assessment*, 14, 360–364.
- Donders, J., Tulskey, D. S., & Zhu, J. (2001). Criterion validity of new WAIS-III subtest scores after traumatic brain injury. *Journal of the International Neuropsychological Society*, 7, 892–898.
- Earnst, K. S., Wadley, V. G., Aldridge, T. M., Steenwyk, A. B., Hammond, A. E., Harrell, L. E., et al. (2001). Loss of financial capacity in Alzheimer's disease: The role of working memory. *Aging, Neuropsychology, and Cognition*, 8, 109–111.
- Evans, J. D., Bond, G. R., Meyer, P. S., Kim, H. W., Lysaker, P. H., Gibson, P. J., et al. (2004). Cognitive and clinical predictors of success in vocational rehabilitation in schizophrenia. *Schizophrenia Research*, 70, 331–342.
- Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for data mining researchers*. Palo Alto: HP Laboratories.
- Frazier, T. W., Youngstrom, E. A., Chelune, G. J., Naugle, R. I., & Lineweaver, T. T. (2004). Increasing the reliability of ipsative interpretations in neuropsychology: A comparison of reliable components analysis and other factor analytic methods. *Journal of the International Neuropsychological Society*, 10, 578–579.
- Frencham, K. A. R., Fox, A. M., & Maybery, M. T. (2005). Neuropsychological studies of mild traumatic brain injury: A meta-analytic review of research since 1995. *Journal of Clinical and Experimental Neuropsychology*, 27, 334–351.
- Gervais, R. O., Ben-Porath, Y. S., Wygant, D. B., & Green, P. (2008). Differential sensitivity of the Response Bias Scale (RBS) and MMPI-2 validity scales to memory complaints. *Clinical Neuropsychologist*, 22, 1061–1079.
- Golden, C. J. (1978). *Stroop Color and Word Test*. Chicago: Stoelting.
- Golden, C. J., Purisch, A. D., & Hammeke, T. A. (1985). *A manual for the Luria-Nebraska Neuropsychological Battery, Forms I and II*. Los Angeles: Western Psychological Services.
- Green, P. (2007). The pervasive influence of effort on neuropsychological tests. *Physical Medicine and Rehabilitation Clinics of North America*, 18, 43–68.
- Green, P., Rohling, M. L., Iverson, G. L., & Gervais, R. O. (2003). Relationships between olfactory discrimination and head injury severity. *Brain Injury*, 17, 479–496.
- Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, 6, 218–224.
- Greiffenstein, M. F., Baker, W. J., & Gola, T. (1996). Motor dysfunction profiles in traumatic brain injury and postconcussion syndrome. *Journal of the International Neuropsychological Society*, 2, 477–485.
- Greve, K. W., Bianchini, K. J., Mathias, C. W., Houston, R. J., & Crouch, J. A. (2002). Detecting malingered performance with the Wisconsin Card Sorting Test: A preliminary investigation in traumatic brain injury. *The Clinical Neuropsychologist*, 16, 179–191.
- Griffith, H. R., Netson, K. L., Harrell, L. E., Zamrini, E. Y., Brockington, J. C., & Marson, D. C. (2006). Amnesic mild cognitive impairment: Diagnostic outcomes and clinical prediction over a two-year time period. *Journal of the International Neuropsychological Society*, 12, 166–175.
- Gronwall, D. M. A. (1977). Paced Auditory Serial Addition Task: A measure of recovery from concussion. *Perceptual and Motor Skills*, 44, 367–373.
- Hannay, H. J., Levin, H. S., & Grossman, R. G. (1979). Impaired recognition memory after head injury. *Cortex*, 15, 269–283.
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *The Wisconsin Card Sorting Test manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources, Inc.
- Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted norms for African American and Caucasian adults (HRB)*. Lutz, FL: Psychological Assessment Resources.
- Heilbrunner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, 23, 1093–1129.
- Heilman, K. M., Watson, R. T., & Valenstein, E. (2012). Neglect and related disorders. In K. M. Heilman, & E. Valenstein (Eds.), *Clinical neuropsychology* (5th ed., pp. 296–348). New York: Oxford University Press.
- Hill, B. D., Boettcher, A. C., Cary, B., Kline, J. S., Womble, M. N., & Rohling, M. L. (2013, February). Much Ado about Norming: A Comparison of the Heaton Demographically Adjusted Norms and the Mitrushina Meta-Norms. Poster presented at the 41st annual conference of the International Neuropsychological Society in Waikoloa, HI.
- Hill, B. D., Rohling, M. L., Boettcher, A. C., & Meyers, J. E. (2013). Cognitive intra-individual variability has a positive association with traumatic brain injury severity and suboptimal effort. *Archives of Clinical Neuropsychology*, 28, 640–648.
- Holdnack, J. A., Millis, S. R., Larrabee, G. J., & Iverson, G. L. (2013). Assessing performance validity with the ACS. In J. A. Holdnack, L. W. Drozdick, L. G. Weiss, & G. L. Iverson (Eds.), *WAIS-IV, WMS-IV, and ACS* (pp. 331–365). San Diego: Academic Press.
- Holdnack, J. A., Zhou, X., Larrabee, G. J., Millis, S. R., & Salthouse, T. A. (2011). Confirmatory factor analysis of the WAIS-IV/WMS-IV. *Assessment*, 18, 178–191.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley Interscience.
- Hsaio, J. K., Bartko, J. J., & Potter, W. Z. (1989). Diagnosing diagnoses: Receiver operating characteristics methods and psychiatry. *Archives of General Psychiatry*, 46, 664–667.

- Hughes, C. P., Berg, L., Danziger, W. L., Coben, L. A., & Martin, R. L. (1982). A new clinical scale for the staging of dementia. *British Journal of Psychiatry*, 140, 566–572.
- Iverson, G. L. (2001). Can malingering be identified with the judgment of line orientation test? *Applied Neuropsychology*, 8 (3), 167.
- Jacobs, M. L., & Donders, J. (2007). Criterion validity of the California Verbal Learning Test-Second Edition (CVLT-II) after traumatic brain injury. *Archives of Clinical Neuropsychology*, 22, 143–149.
- Jasinski, L. J., Berry, D. T. R., Shandera, A. L., & Clark, J. A. (2011). Use of the Wechsler Adult Intelligence Scale Digit Span subtest for malingering detection: A meta-analytic review. *Journal of Clinical and Experimental Neuropsychology*, 33, 300–314.
- Jastak, J., & Wilkinson, G. S. (1984). *The wide range achievement test-revised*. Wilmington, DE: Jastak Associates.
- Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test* (2nd ed.). Philadelphia: Lea and Febiger.
- Keifer, E., & Tranel, D. (2013). A neuropsychological investigation of the Delis-Kaplan Executive Function System. *Journal of Clinical and Experimental Neuropsychology*, 35, 1048–1059.
- Killgore, W. D. S., & DellaPietra, L. (2000). Using the WMS-III to detect malingering: Empirical validation of the Rarely Missed Index (RMI). *Journal of Clinical and Experimental Neuropsychology*, 22, 761–771.
- Langeluddecke, P. M., & Lucas, S. K. (2003). Quantitative measures of memory malingering on the Wechsler Memory Scale-Third Edition in mild head injury litigants. *Archives of Clinical Neuropsychology*, 18, 181–197.
- Larrabee, G. J. (1986). Another look at VIQ-PIQ scores and unilateral brain damage. *International Journal of Neuroscience*, 29, 141–148.
- Larrabee, G. J. (2000). Association between IQ and neuropsychological test performance: Commentary on Tremont, Hoffman, Scott, and Adams (1998). *The Clinical Neuropsychologist*, 14, 139–145.
- Larrabee, G. J. (2003a). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, 17, 410–425.
- Larrabee, G. J. (2003b). Exaggerated pain report in litigants with malingered neurocognitive dysfunction. *The Clinical Neuropsychologist*, 17, 395–401.
- Larrabee, G. J. (2003c). Detection of symptom exaggeration with the MMPI-2 in litigants with malingered neurocognitive dysfunction. *The Clinical Neuropsychologist*, 17, 54–68.
- Larrabee, G. J. (2003d). Lessons on measuring construct validity: A commentary on Delis, Jacobson, Bondi, Hamilton, and Salmon. *Journal of the International Neuropsychological Society*, 9, 947–953.
- Larrabee, G. J. (Ed.). (2007). *Assessment of malingered neuropsychological deficits*. New York: Oxford University Press.
- Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist*, 22, 666–679.
- Larrabee, G. J. (2009). Malingering scales for the Continuous Recognition Memory Test and the Continuous Visual Memory Test. *The Clinical Neuropsychologist*, 23, 167–180.
- Larrabee, G. J. (2012a). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, 18, 625–630.
- Larrabee, G. J. (2012b). Assessment of malingering. In G. J. Larrabee (Ed.), *Forensic neuropsychology. A scientific approach* (pp. 116–159). New York: Oxford.
- Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology*, 29, 364–373.
- Larrabee, G. J., & Curtiss, G. (1992). Factor structure of an ability-focused neuropsychological battery (abstract). *Journal of Clinical and Experimental Neuropsychology*, 14, 65.
- Larrabee, G. J., & Curtiss, G. (1995). Construct validity of various verbal and visual memory tests. *Journal of Clinical and Experimental Neuropsychology*, 17, 536–547.
- Larrabee, G. J., Kane, R. L., Schuck, J. R., & Francis, D. J. (1985). The construct validity of various memory testing procedures. *Journal of Clinical and Experimental Neuropsychology*, 7, 239–250.
- Larrabee, G. J., Lagen, J. W., & Levin, H. S. (1985). Sensitivity of age-decline resistant (“Hold”) WAIS subtests to Alzheimer’s disease. *Journal of Clinical and Experimental Neuropsychology*, 7, 497–504.
- Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2008). Sensitivity to brain dysfunction of the Halstead Reitan vs. an ability-focused neuropsychological battery. *The Clinical Neuropsychologist*, 22, 813–825.
- Larrabee, G. J., Trahan, D. E., Curtiss, G., & Levin, H. S. (1988). Normative data for the Verbal Selective Reminding Test. *Neuropsychology*, 2, 173–182.
- Leonberger, F. T., Nicks, S. D., Larrabee, G. J., & Goldfader, P. R. (1992). Factor structure and construct validity of a comprehensive neuropsychological battery. *Neuropsychology*, 6, 239–249.
- Levin, H. S., Benton, A. L., & Grossman, R. G. (1982). *Neurobehavioral consequences of closed head injury*. New York: Oxford.
- Lezak, M. D., Howieson, D. E., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.
- Loewenstein, D. A., & Mogosky, B. (1999). Functional assessment in the older adult patient. In P. Lichtenberg (Ed.), *Handbook of assessment in clinical gerontology* (pp. 529–554). New York: Wiley.
- Loring, D. W., & Larrabee, G. J. (2006). Sensitivity of the Halstead and Wechsler test batteries to brain damage: Evidence from Reitan’s original validation sample. *The Clinical Neuropsychologist*, 20, 221–229.
- Loring, D. W., Strauss, E., Hermann, B. P., Barr, W. B., Perrine, K., Trenerry, M. R., et al. (2008). Differential neuropsychological test sensitivity to left temporal lobe epilepsy. *Journal of the International Neuropsychological Society*, 14, 394–400.
- Lu, P. H., Boone, K. G., Cozolino, L., & Mitchell, C. (2003). Effectiveness of the Rey-Osterreith Complex Figure Test and the Meyers and Meyers Recognition Trial in the detection of suspect effort. *The Clinical Neuropsychologist*, 17, 426–440.
- Machamer, J., Temkin, N., Fraser, R., Doctor, J. N., & Dikmen, S. S. (2005). Stability of employment after traumatic brain injury. *Journal of the International Neuropsychological Society*, 11, 807–816.
- Malec, J. F., Kragness, M., Evans, R. W., Finlay, K. L., Kent, A., & Lezak, M. D. (2003). Further psychometric evaluation and revision of the Mayo-Portland Adaptability Inventory in a national sample. *Journal of Head Trauma Rehabilitation*, 18, 479–492.
- Marson, D. C., Ingram, K. K., Cody, H. A., & Harrell, L. E. (1995). Assessing the competency of patients with Alzheimer’s disease under different legal standards: A prototype instrument. *Archives of Neurology*, 52, 949–954.

- McGrew, K. S. (2009). CHC Theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1–10.
- McGurk, S. R., Mueser, K. T., Harvey, P. D., LaPuglia, R., & Marder, J. (2003). Cognitive and symptom predictors of work outcome for clients with schizophrenia in supported employment. *Psychiatric Services*, 54, 1129–1135.
- Meyers, J. E., Galinsky, A. M., & Volbrecht, M. (1999). Malingering and mild brain injury: How low is too low? *Applied Neuropsychology*, 6 (4), 208.
- Meyers, J. E., & Meyers, K. (1995). *The Meyers Scoring System for the Rey complex figure and the recognition trial: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Meyers, J. E., Miller, R. M., Thompson, L. M., Scalese, A. M., Allred, B. C., Rupp, Z. W., et al. (2014). Using likelihood ratios to detect invalid performance with performance validity measures. *Archives of Clinical Neuropsychology*, 29, 224–235.
- Meyers, J. E., Miller, R. M., & Tuita, A. R. R. (2013). Using pattern analysis matching to differentiate TBI and PTSD in a military sample. *Applied Neuropsychology: Adult*, 21, 60–68.
- Meyers, J. E., & Rohling, M. L. (2004). Validation of the Meyers Short Battery on mild TBI patients. *Archives of Clinical Neuropsychology*, 19, 637–651.
- Meyers, J. E., & Volbrecht, M. E. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology*, 18, 61–276.
- Miller, J. B., Fichtenberg, N. L., & Millis, S. R. (2010). Diagnostic efficiency of an ability focused battery. *The Clinical Neuropsychologist*, 24, 678–688.
- Millis, S. R., Putnam, S. H., Adams, K. M., & Ricker, J. H. (1995). The California Verbal Learning Test in the detection of incomplete effort in neuropsychological evaluation. *Psychological Assessment*, 7, 463–471.
- Millis, S. R., & Volinsky, C. T. (2001). Assessment of response bias in mild head injury: Beyond malingering tests. *Journal of Clinical and Experimental Neuropsychology*, 23, 809–828.
- Mitrushina, M., Boone, K., Razani, J., & D'Elia, L. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.
- Mittenberg, W., Azrin, R., Millsaps, C., & Heilbrunner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale-Revised. *Psychological Assessment*, 5, 34–40.
- Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology*, 24, 1094–1102.
- Mittenberg, W., Theroux-Fichera, S., Zielinski, R. E., & Heilbrunner, R. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale-Revised. *Professional Psychology: Research and Practice*, 26, 491–498.
- Morgan, J. E., & Sweet, J. J. (2009). *Neuropsychology of malingering casebook*. Hove, East Sussex, UK: Psychology Press.
- Novack, T. A., Banos, J. H., Alderson, A. L., Schneider, J. J., Weed, W., Blankenship, J., et al. (2006). UFOV performance and driving ability following traumatic brain injury. *Brain Injury*, 20, 455–461.
- Osterrieth, P. A. (1944). Le test de copie d'une figure complexe: Contribution à l'étude de la perception et de la mémoire. *Archives de Psychologie*, 30, 286–356.
- Pearson. (2009). *Advanced clinical solutions for use with WAIS-IV and WMS-IV*. San Antonio: Pearson Education.
- Powell, J. B., Cripe, L. I., & Dodrill, C. B. (1991). Assessment of brain impairment with the Rey Auditory Verbal Learning Test: A comparison with other neuropsychological measures. *Archives of Clinical Neuropsychology*, 6, 241–249.
- Purves, D., Brannon, E. M., Cabeza, R., Huettel, S. A., LaBar, K. S., Platt, M. L., et al. (2008). *Principles of cognitive neuroscience*. Sunderland, MA: Sinauer Associates.
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, 20, 33–65.
- Rappaport, M., Hall, K., Hopkins, K., Belleza, T., & Cope, D. (1982). Disability Rating Scale for severe head trauma: Coma to community. *Archives of Physical Medicine and Rehabilitation*, 63, 118–123.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological test battery. theory and clinical interpretation* (3rd ed.). Tucson, AZ: Neuropsychology Press.
- Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*, 28, 286–340.
- Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law and Human Behavior*, 29, 615–620.
- Rohling, M. L., Axelrod, B., & Wall, J. (2008, June). What impact does co-norming have on Neuropsychological Test Scores: Myth vs. reality? Poster presented at the 6th annual conference of the American Academy of Clinical Neuropsychology in Boston, MA.
- Rohling, M. L., Binder, L. M., Demakis, G. J., Larrabee, G. J., Ploetz, D. M., & Langhinrichsen-Rohling, J. (2011). A meta-analysis of neuropsychological outcome after mild traumatic brain injury: Re-analyses and reconsiderations of Binder et al. (1997), Frencham et al. (2005), and Pertab et al. (2009). *The Clinical Neuropsychologist*, 25, 608–623.
- Rohling, M. L., Larrabee, G. J., Greiffenstein, M. F., Ben Porath, Y. S., Lees-Haley, P., Green, P., et al. (2011). A misleading review of response bias: Commentary on McGrath, Mitchell, Kim, and Hough (2010). *Psychological Bulletin*, 137, 708–712.
- Rohling, M. L., Meyers, J. E., & Millis, S. R. (2003). Neuropsychological impairment following traumatic brain injury: A dose-response analysis. *The Clinical Neuropsychologist*, 17, 289–302.
- Rohling, M. L., Miller, L. S., & Langhinrichsen-Rohling, J. (2004). Rohling's interpretive method for neuropsychological case data: A response to critics. *Neuropsychology Review*, 14, 155–169.
- Russell, E. W., & Starkey, R. I. (1993). *Halstead Russell Neuropsychological Evaluation System (manual and computer program)*. Los Angeles, CA: Western Psychological Services.
- Salmon, D. P., Heindel, W. C., & Lange, K. L. (1999). Differential decline in word generation from phonemic and semantic categories during the course of Alzheimer's disease: Implications for the integrity of semantic memory. *Journal of the International Neuropsychological Society*, 5, 692–703.
- Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive function. *Neuropsychology*, 19, 532–545.
- Salthouse, T. A., & Saklofske, D. H. (2010). Do the WAIS-IV tests measure the same aspects of cognitive functioning in adults under and over 65? In L. G. Weiss, D. H. Saklofske, D. Coalson, & S. E. Raiford (Eds.), *WAIS-IV. Clinical use and interpretation* (pp. 217–235). San Diego: Academic Press.
- Schmidt, M. (1996). *Rey Auditory and Verbal Learning Test*. Los Angeles: Western Psychological Services.

- Schretlen, D. J., & Shapiro, A. M. (2003). A quantitative review of the effects of traumatic brain injury on cognitive functioning. *International Review of Psychiatry*, 15, 341–349.
- Sherod, M. G., Griffith, H. R., Copeland, J., Belue, K., Kryzwanski, S., Zamrini, E. Y., et al. (2009). Neurocognitive predictors of financial capacity across the dementia spectrum: Normal aging, MCI, and Alzheimer's disease. *Journal of the International Neuropsychological Society*, 15, 258–267.
- Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, 13, 545–561.
- Smith, A. (1983). *The Symbol Digit Modalities Test (SDMT). Manual (revised)*. Los Angeles: Western Psychological Services.
- Stern, R., & White, T. R. (2003). *Neuropsychological assessment battery*. Manual. Lutz: Psychological Assessment Resources.
- Strauss, E., Spreen, O., & Sherman, E. M. S. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.
- Stuss, D. T., Stethem, L. L., Hugenholtz, H., & Richard, M. T. (1989). Traumatic brain injury: A comparison of three clinical tests and analysis of recovery. *The Clinical Neuropsychologist*, 3, 145–156.
- Suhr, J. A., & Boyer, D. (1999). Use of the Wisconsin Card Sorting Test in the detection of malingering in student simulator and patient samples. *Journal of Clinical and Experimental Neuropsychology*, 21, 701–708.
- Sweet, J. J., Meyer, D. G., Nelson, N. W., & Moberg, P. J. (2011). The TCN/AACN 2010 "Salary Survey": Professional practices, beliefs, and incomes of U. S. Neuropsychologists. *The Clinical Neuropsychologist*, 25, 12–61.
- Swets, J. J. (1973). The relative operating characteristic in psychology. *Science*, 182, 990–1000.
- Tabachnick, B. G., & Fidell, L. (2005). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.
- Tellegen, A., & Ben-Porath, Y. S. (2008/2011). *MMPI-2-RF* (Minnesota Multiphasic Personality Inventory-2 Restructured Form) technical manual. Minneapolis: University of Minnesota Press.
- Tenhula, W. N., & Sweet, J. J. (1996). Double cross-validation of the Booklet Category Test in detecting malingered traumatic brain injury. *The Clinical Neuropsychologist*, 10, 104–116.
- Trahan, D. E., & Larrabee, G. J. (1988). *The Continuous Visual Memory Test*. Lutz, FL: Psychological Assessment Resources.
- Trahan, D. E., Larrabee, G. J., & Quintana, J. W. (1990). Visual recognition memory in normal adults and patients with unilateral vascular lesions. *Journal of Clinical and Experimental Neuropsychology*, 12, 857–872.
- Trahan, D. E., Larrabee, G. J., Quintana, J. W., Goethe, K. E., & Willingham, A. C. (1989). Development and clinical validation of an expanded paired associates test with delayed recall. *The Clinical Neuropsychologist*, 3, 169–183.
- Trenerry, M. R., Crosson, B., DeBoe, J., & Leber, W. R. (1989). *The Stroop Neuropsychological Screening Test*. Lutz, FL: Psychological Assessment Resources, Inc.
- Tulsky, D. S., & Price, L. R. (2003). The joint WAIS-III and WMS-III factor structure: Development and cross-validation of a six factor model of cognitive functioning. *Psychological Assessment*, 15, 149–162.
- Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, M. A., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*, 23, 297–313.
- Vollbrecht, M. E., Meyers, J. E., & Kaster-Bundgaard, J. (2000). Neuropsychological outcome of head injury using a short battery. *Archives of Clinical Neuropsychology*, 15, 251–264.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale, revised. Manual*. New York: The Psychological Corporation.
- Wechsler, D. (1987). *Wechsler Memory Scale-Revised manual*. San Antonio: The Psychological Corporation.
- Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale, III, manual*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler Memory Scale III, manual*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1997c). *WAIS-III/WMS-III Technical manual*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2008). *WAIS-IV. Technical and interpretive manual*. San Antonio: Pearson.
- Wechsler, D. (2009). *WMS-IV. Technical and interpretive manual*. San Antonio: Pearson.
- Welsh, K. A., Butters, N., Mohs, R. C., Beekly, D., Edland, S., Fillenbaum, G., et al. (1994). The consortium to establish a registry for Alzheimer's disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology*, 44, 609–614.
- Whelihan, W. M., DiCarlo, M. A., & Paul, R. H. (2005). The relationship of neuropsychological functioning to driving competence in older persons with early cognitive decline. *Archives of Clinical Neuropsychology*, 20, 217–228.
- Williams, M. W., Rapport, L. J., Hanks, R. A., Millis, S. R., & Greene, H. A. (2013). Incremental validity of neuropsychological evaluations to computed tomography in predicting long-term outcomes after traumatic brain injury. *The Clinical Neuropsychologist*, 27, 356–375.
- Wilkinson, G. S. (1993). *WRAT-3: The Wide Range Achievement Test administration manual* (3rd ed.). Wilmington, DE: Wide Range.
- Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test-4 (WRAT-4)*. Lutz, FL: Psychological Assessment Resources, Inc.
- Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., & Sweet, J. J. (2010). Effort indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist*, 24, 153–168.
- Wygant, D. B., Sellbom, M., Ben-Porath, Y. S., Stafford, K. P., Freeman, D. B., & Heilbronner, R. H. (2007). The relation between symptom validity testing and MMPI-2 scores as a function of forensic evaluation context. *Archives of Clinical Neuropsychology*, 22, 489–499.
- Zakzanis, K. K., Leach, L., & Kaplan, E. F. (1999). *Neuropsychological differential diagnosis*. New York, NY: Taylor & Francis.