BRAIN AND COGNITION **31,** 114–132 (1996)
ARTICLE NO. 0037

Issues in the Assessment of Cognitive Function in Dementia

WILLIAM MILBERG

Geriatric Research, Education and Clinical Center, VA Medical Center, and Department of Psychiatry, Harvard Medical School

The increasing prevalence of elderly patients presenting with disorders of cognitive functioning suggestive of dementia, coupled with limits in resources available to address these problems, is going to necessitate the development of new technologies to be used for assessment. These measures should be efficient, designed to reflect both the current knowledge of brain function and the developmental characteristics of older patients. Though few current measures of cognitive function meet these goals, neuropsychological ''microbatteries'' such as the DRS and NCSE may serve as models for a new generation of cognitively specific assessment instruments.

With the explosion of neuropsychology as a clinical science many assessment instruments have become available to help determine if a patient suffers from dementia or other acquired disorders of cognition. In some cases these instruments are lengthy, consisting of groups of psychological tests chosen for their sensitivity to the presence of cortical damage (e.g., Halstead-Reitan Neuropsychological Battery (Halstead, 1947; Reitan & Davison, 1974)). In other cases the instruments are brief but specifically designed to assess functions that define the clinical entity of dementia (e.g., Mattis Dementia Rating Scale (Mattis, 1976)). What properties should guide one's choice of currently available instruments sensitive to the changes in cognitive function? What properties should guide the development of future measures of this type? The answers to these questions, I argue, are not straightforward and go beyond an intuitive appeal to the traditional concepts of reliability and validity. Though

The work was supported in part by VA Merit Review 097-44-3765-001 to William Milberg at the West Roxbury VA Medical Center. I thank Laura Grande, Regina McGlinchey-Berroth, and Nancy Hebben for their help in the preparation of the manuscript. I also thank Larry Leach and Brian Levine for their astute suggestions for revisions of an earlier version of the manuscript. Address reprint requests to the author at GRECC, 1400 VFW Parkway, VA Medical Center, West Roxbury, MA 02132.

0278-2626/96 $18.00

great advances have been made in the field of neuropsychology in general, currently available instruments have advantages and disadvantages that usually represent a compromise between the various demands of psychometry and practicality. In this paper I discuss how such issues as efficiency, domain specificity, item difficulty, and reliability affect the assessment of older patients suspected of suffering from dementia. This discussion leads to a preliminary set of seven criteria to be considered in the choice of cognitive assessment tools to be used with older adults. These criteria will then be used to review some currently available test instruments. Though some will sound familiar and be applicable to the development of neuropsychological instruments in general, it is argued that the assessment of older adults with dementia presents a number of unique challenges that require the rethinking of traditional strategies of test construction. Table 1 summarizes these criteria and the assessment instruments that are discussed later in this paper.

TEST EFFICIENCY AND THE DEMOGRAPHIC AND ECONOMIC CONTEXT OF COGNITIVE ASSESSMENT

The technology of cognitive and neuropsychological assessment is approaching a crossroads created by evolving scientific and social forces. The past 20 years has witnessed an explosion in data and theory concerned with the representation and processing of complex information in the central nervous system. As I argue later, the state of the art of clinical assessment of neurally based cognitive disorders does not reflect the proliferation of empirical and theoretical work that has occurred in cognitive neuroscience, neuropsychology, and behavioral neurology, although these areas of basic science directly address many of the issues ultimately of concern to clinicians. It is ironic that rather than being shaped by relevant advances in basic science, clinical assessment technology may ultimately be shaped by social and economic trends affecting the number and kinds of patients serviced by that technology.

Advances in medicine and public health policy are slowly but surely increasing the expected life span of people living in the United States and other industrialized nations (Spencer, 1989). Quite independently, demographic forces are increasing the number of individuals with the potential to reach that life span (Services, 1992). Consequently, the incidence and prevalence of late life conditions that impact cognitive functions have increased dramatically in the past 20 years and are expected to continue to increase into the early 21st century (Spencer, 1989). Today approximately 11% of the United States population of 260 million is 65 and older. The percentage of elderly adults is expected to increase to 21% by 2030 given current estimates of life expectancy (Spencer, 1989). Alzheimer's disease

TABLE 1
An Evaluation of Current Measures of Cognitive Functions in Dementia

Measure	Criterion[a]						
	1. Neuropathological Sensitivity/Specificity	2. Cognitive Domain Specificity	3. Construct/ Process Specificity	4. Functional Specificity	5. Context Appropriate Reliability	6. Age Appropriate Item Difficulty	7. Efficiency
Full battery							
Halstead-Reitan	+	−	−	+	?	?	−
Boston Process	+	−/+	−/+	+	?	?	−
Luria-Nebraska	+	−/+	?	+	?	?	−
UCSD Battery	+	+	−/+	?	?	?	−
Microbattery							
Dementia Rating Scale	+/−	+/−	−	?	?	+	+
Neurobehavioral Cognitive Status Examination	+/−	+/−	−	?	?	+	+
Geriatric Evaluation of Mental Status	?	+	+	?	+	+	+
Screen							
Mini-Mental State	+/−	−	−	−	+	+	+

[a] Measures are rated according to the seven criteria described in the text. +, support for the criteria; −, no support; +/−, equivocal or inconsistent support; ?, no data available.

alone appears to affect approximately 10% of adults over age 65 (Evans et al., 1989), with its prevalence increasing to nearly 30% in adults over age 85. Add to these prevalence figures patients suffering from other causes of dementia, delirium and other varieties of disordered cognition, and it becomes apparent that the social and economic impact of cognitive disorders will grow as a major concern for the development of health and social policy. At the same time that these demographic forces are at work, however, the resources available for health care in general, and such ancillary health care activities as cognitive and neuropsychological assessment, in particular, are reaching a limit or are shrinking.

Though these points may be familiar to many readers, it is nevertheless important to consider them in the current context. By necessity, the properties and features of assessment techniques will be increasingly shaped by the forces of demographics and economics. Quite simply, more patients will require the assessment of cognitive functions, while the time and resources to provide this service are likely to become increasingly limited. Apart from the other issues described below, these circumstances will demand that assessment techniques be designed to be as efficient as possible. Efficiency in this context means that the ratio of useful information obtained to time (and cost) must be high. This certainly sounds like the most uncontroversial and obvious recommendation that was ever proposed, yet ''efficiency'' per se is not usually considered a factor in the course of test construction (APA, 1985). Consider the current options for the assessment of cognitive functioning in dementia: neuropsychological batteries (e.g., Halstead-Reitan (Halstead, 1947; Reitan & Davison, 1974), Luria-Nebraska (Golden, Hammeke, & Purisch, 1980), Boston Process Approach (Milberg, Hebben, & Kaplan, 1986), etc.) and dementia screening instruments (e.g., Mini-Mental State Exam (Folstein, Folstein, & McHugh, 1975), Dementia Rating Scale (Mattis, 1976)). Formal neuropsychological batteries (both fixed and flexible) are certainly capable of providing useful and detailed information about the mental abilities of a suspected dementia patient, but usually take 2 to 8 hr to administer. Screening measures may be brief but limited in sensitivity and specificity and rarely provide sufficient information to solidify a diagnosis. The issue of what constitutes ''useful information'' will require further discussion, but it is safe to say that in the current state of the art, the efficiency of test instruments has not been held as a high priority. The efficiency ''ratio'' of most available cognitive test instruments suffers from either high levels of cost (in time and ultimately in required reimbursement) or low levels of yielded information. Apart from the traditional concerns of reliability and validity, efficiency and cost-effectiveness should be given much higher priority than is currently the case when cognitive tests are chosen and developed. In addition to being cost-effective, brief, assessment instruments have the potential to be less unpleasant, and intrusive, for older patients than longer test batteries.

CRITERION VALIDITY AND CRONBACH'S LAMENT: THE SEPARATION OF BASIC AND CLINICAL SCIENCE INFORMATION IN COGNITIVE/ NEUROPSYCHOLOGICAL ASSESSMENT

In his presidential address to the American Psychological Association Lee Cronbach (1964) lamented that psychology had separated into two streams: one concerned with the development of theory and the manipulation of experimental independent measures and the other concerned with individual differences and the use of correlational methods. These two disciplines map closely on the distinction drawn here between clinical and experimental neuropsychology. It was Cronbach's hope that eventually these two disciplines would unite so that individual differences would be a focus of experimental investigation and accounted for by general psychological theory and correlational methods would be informed by the variables and theory derived from experimenal psychology. Until very recently this kind of separation has characterized the development of neuropsychological measures. The developers of clinical neuropsychological measures have mostly been concerned with criterion validity (prediction of brain damage or presence of brain disease), often dealing with the interpretative or theoretical psychological issues after the fact. In other cases the development of the measures reflects theoretical standpoints that are simply outdated (e.g., antilocalizationism).

Clinical measures of cognitive function often come into clinical use through empirical association with brain lesions or etiologic factors[1] and often do not reflect the insights into the specification and description of brain function that have been gained through more basic laboratory research. Very often the measures or tasks that become part of ''standardized test batteries'' originated from clinicians' intuitions or were borrowed from other fields. Clinical observation and intuition may be an excellent source for hypotheses about the nature of cognitive function. However, tasks based purely on clinicians' intuitions, though sometimes correctly associated with the clinical entity in question, will often not have benefited from the refinements that experimental analysis and a theoretical context can provide. Tests designed for vocational selection or to evaluate scholastic aptitude are not necessarily optimal for evaluating brain function! For example, the Halstead-Reitan Battery includes the Sequin-Goddard Form Board and Seashore Rhythm Test and the Wechsler Adult Intelligence Scale, none of which was designed to

[1] This of course begs the question of the validity of the independent clinical and radiological evidence of neuropathology. Radiological measures are themselves subject to validation and are not necessarily the ultimate criterion for the evaluation of the predictive or concurrent validity of behavioral measures. For the sake of the current discussion, however, we assume that there is evidence of a predictive association between the behavioral measure and a currently accepted independent radiological measure of the structural integrity of relevant cortical and subcortical structures.

be a test of brain function per se. Other measures in the Halstead-Reitan Battery, like the Category Test, had their origin in a theory of brain function (i.e., Halstead's biological intelligence, a holistic theory of brain function) but have persisted in use mainly because of their validity as measures sensitive to brain disease and not because they specify a cognitive function or are easily interpretable. These tests, for example, do not reflect the considerable, and clinically relevant, refinements in our basic understanding of such areas as language (e.g., Goodglass, 1993), memory (Parkin & Leng, 1993), and executive functions (Shallice & Burgess, 1991).

The Mini-Mental State Exam (Folstein et al., 1975), one of the most popular screening instruments for the presence of dementia, was constructed of traditional components of the neuropsychiatric bedside examination that were passed down to the authors through several generations of clinicians (Folstein, 1990). It has become popular in part because it does indeed identify clinically judged cases of dementia and undoubtedly also because it is brief. Though the establishment of the sensitivity and specificity of a cognitive measure to the presence of a brain lesion is a critical part of the validity of neuropsychological tests, the presence of such criterion validity does not mean the tests are constructed in an optimally interpretable or useful manner or that the obtained levels of sensitivity and specificity have been maximized or obtained at an acceptable cost.

If there is any general proposition that can be drawn from the past 25 years of research on brain function, it is that the human cerebral cortex is organized along function specific lines. Converging evidence from studies on the effect of focal lesions on cognition, and more recent studies using new radiological techniques sensitive to brief changes in neural metabolism (e.g., PET, SPECT, and fast MRI) in normal subjects, strongly suggests that the brain is at least in part organized as a series of localizable information processing devices that appear to make independent contributions to human mental activity. Small changes in the coordinates of a lesion will often make large differences in the clinical symptomatology displayed by the patient (Damasio & Damasio, 1989; Kertesz, 1994; Milberg & Albert, 1991). Small changes in task demands can make large differences in the points of peak measured metabolic activity as measured by such techniques as PET (Damasio & Damasio, 1989; Kertesz, 1994; Milberg & Albert, 1991). Though the boundaries of these functional neural maps have not been fully defined, the principle of localization of function remains a powerful and useful construct.[2]

[2] This is not to say that all of the functional activities of the brain are focally organized. There is evidence that some forms of information are represented in a widely distributed manner and that the populations of neurons concerned with different functions may overlap. These exceptions do not challenge the overwhelming value of a localizationist view of cortical function.

Although the idea that the central nervous system can be described as focally organized, the precise definition of the processing activities within these focal areas is far from fully settled. Nevertheless, a second general proposition derived both from neuropsychological and cognitive psychological research is that broad functions such as reading, memory, language comprehension and production, and attention can be broken down into more basic processing elements or subtasks (McCarthy & Warrington, 1990). Again, the precise definition of the principal components or defining processes characterizing various functions is not settled, but the principle that cognitive functions may be decomposed into more basic processing elements, and that these basic processing elements are candidate descriptions for neural function, is a central axiom of cognitive neuropsychology and much of neuroscience.

Even without a definitive consensus on the boundaries or contents of a neural map of cognitive functions, there is enough consensus within the scientific literature to derive useful clinical measures. Even a basic division into such broad functional domains as language, memory, visual processing, attention/arousal, and executive functions would be a useful starting point. Within each domain tasks may be constructed that are sensitive to the variables known to be of biological relevance and that reflect the current understanding of underlying psychological processes. Hence tasks should not only be domain specific, they should assess processes or constructs specific to those domains. For example, memory measures should allow some separation of encoding and retrieval operations. Language assessment should separate comprehension from production and lexical/semantics from syntactic operations. Cognitive assessment tools that are designed even with such specificity in mind can potentially be relatively brief and thereby more efficient than assessment tools that are in use only because of their empirical association with markers of brain pathology which are likely to be cognitively heterogeneous, or composed of elements derived from cognitive multiple systems. Measures that are specifically targeted to be functionally homogeneous can potentially be more directly relevant to the focal organization of the brain than most currently available measures. The use of functionally homogeneous cognitive measures is particularly important because of the increasing consensus that dementia is not a uniform disorder within or across etiologies. Patients with similar underlying pathologies may vary widely in the cortical distribution of the causative lesion and in the behavioral manifestations of those lesions (e.g., Jorm, 1985; Martin et al., 1985, 1986). Global measures of impairment do not capture this biological and cognitive variability.

An additional by-product of this feature of cognitive domain specificity is the possibility that the measure of cognitive function will provide information relevant to the patient's ability to function in day-to-day life. Part of this may be purchased simply through greater face validity. It appears self-

evident that tests that are cognitively homogeneous or specific are more likely to be clearly interpretable and easy to describe than tests that are cognitively heterogeneous. The more specific and obvious the measure, the more likely it will be to determine if it is a component of other tasks or if it is relevant to a patient's day-to-day functioning. There is currently little data concerning the relationship between measures of cognitive functioning and activities of daily living. The emerging evidence, however, suggests that domain specific homogeneous measures may be better predictors of activities of daily living than global measures of cognitive severity (Henderson, Mack, & Williams, 1989; Weintraub, Baratz & Mesulum, 1982; Vitaliano et al., 1984a).

RELIABILITY: POTENTIAL PROBLEMS AND A PARADOX

Classic reliability theory (Spearman, 1910) rests on the assumption that an observed test score is composed of two elements: the ''true score'' that represents the actual quantity that is to be measured and an ''error score'' that represents ''noise'' that randomly varies around the true score with each attempt at measurement. When estimated across a number of individuals or items so that there is a distribution of observed scores, the reliability of a measure is often defined as the ratio of variance (i.e., the square of the standard deviation of the distribution) attributable to the true score to the variance of the total or observed score (Magnusson, 1966). Hence, the greater the proportion of the observed score variance that is attributable to the ''true score,'' the greater the reliability of the measure. Since a given true score is assumed to be stable across time and measurement context and ''error scores'' are assumed to be random and uncorrelated, it follows that reliability will increase with repeated measurements of that ''true score.'' According to this classic view ''true score'' variance may increase as the square of the number of times test length is doubled! The reliability of a measure is assumed to represent the ''ceiling'' for the criterion validity of a measure. Specifically, the index of reliability (or the square root of the reliability correlation coefficient) represents the maximum correlation value that may be obtained in measuring the relationship between a test score and a criterion score (Magnusson, 1966).

Though it would seem that reliability should serve as the most incontrovertible goal in the course of test development, there are nevertheless practical difficulties in the determination of test reliability especially as these measures are applied to older patients with dementia. These limits in the current methods to determine the reliability of a measure may in some cases undermine its utility, especially in patients with dementia. First consider the basic assumption that the reliability of a measure necessarily sets the upper limits of the validity of a measure.

There are circumstances where a test may appear to have modest reliabil-

ity, but extremely high classificatory power. This is especially true in cases involving dichotomous classification (i.e., predicting the presence or absence of an underlying state). For example the statistic kappa (Cohen & Zuckerman, 1960) is often used to determine interrater or retest reliability in cases of dichotomous classifications. Kappa is calculated as follows:

$$\kappa = \frac{p_o - p_c}{1 - p_c},$$

where p_o is the proportions of agreement that were observed and p_c is the proportion of agreement expected by chance. So if two raters agree all the time and chance is 0.5, kappa will equal 1.0. Carey and Gottesman (1978) point out that it is possible to be presented with a situation where classification rates are high, but kappa is low. This phenomenon is particularly dramatic when the actual base rate of the disorder being classified has a low population base rate (less than 20%). Faraone and Tsuang (1994) provide examples of this effect when the prevalence of clinical phenomenon is 0.01, kappa = 0.14 indicating very low interrater agreement, while sensitivity (i.e., correct identification of individuals who have the diagnostic entity) and specificity (correct identification of individuals who **do not** have the diagnostic entity) are a remarkable 0.95! This issue has been confronted within the literature on psychiatric diagnosis but has not yet received wide discussion in the neuropsychological literature.

As indicated the recognition that a test may be ''valid'' in the face of modest reliability is particularly important when the tests in question concern the assessment of cognitive disorders in patients with dementia. Consider the theoretical relationship between test length and reliability mentioned earlier. Critical to the ''rule'' that reliability increases as a function of test length are the assumptions that ''error scores'' will vary randomly with each measurement or test item, that the standard deviation of this error score distribution is constant, and that error scores across test items are uncorrelated. In practice, however, these assumptions may be questionable, particularly when the cognitive symptoms presented by the patient produce a positive relationship between test length and error of measurement. For example, if a patient suffers from increased fatigue, distractibility, or a memory disorder that increasingly affects compliance with test instructions over time, test item $n + 1$ may show increased error variance compared to test item n. If the rate of error variance increment remains less than the rate of true score variance increment, then test length may lead to only modest gains in reliability. If the rate of error variance increment is greater than the rate of true score variance increment, then increases in test length may actually lead to decreases in reliability. This scenario undoubtedly sounds familiar to clinicians who have attempted to administer long batteries of cognitive tests to elderly patients with dementia. It is a common experience that patients will perform

more poorly on tests given later in a battery than in the beginning of the test session. Not only may longer tests produce under some circumstances less reliable results than short tests, but the reliability of a given measure may be affected by the fact that it is embedded in a large series of tests that may affect the reliability of all the items in that measure.

Since efficiency is usually not a major concern to test developers, and test reliability is usually assessed for each measure individually, there is simply no available data to address the issue of the optimal length of a test as it is given to its target population or in the actual context of its administration.

In practice reliability is assessed in three ways: through the examination of a measure's internal consistency (interim correlations, split half reliability, Cronbach's a, etc.), temporal stability (i.e., test–retest), or through the scorer agreement (i.e., interrater). Each of these methods of assessing reliability may potentially be affected by factors that affect the stability of the ''true score'' itself. In the case of the patient with dementia the internal consistency of a measure may be affected by the factors mentioned earlier (fatigue, fluctuations in attention, distractibility, memory, etc.). The measurement of temporal stability may be limited due to diurnal fluctuations in symptoms and because the quality and severity of the symptoms may evolve over time. Interrater agreement may not be affected by these factors but is of limited usefulness as a method of assessing the reliability of objective standardized measures.

In the future, the reliability of cognitive measures to be used in patients with dementia (and other disorders of cognition) must be determined within the clinical context in which the measure is being used. The goal should be to optimize reliability as it empirically affects test validity and efficiency. In some cases brief measures will turn out to be more reliable than long measures. Measures that individually may appear to be modestly reliable may nevertheless retain sensitivity to the diagnostic entity of interest.

DEVELOPMENTAL APPROPRIATENESS

In addition to the strategy of targeting specific functional domains another strategy that may be used to enhance the reliability and efficiency of cognitive tasks is to employ items representing a range and difficulty level appropriate for the population in question. Many of the tasks used in general neuropsychological test batteries, for example, were designed with items representing a sufficient range of difficulty to retain sensitivity across the entire range of ages and premorbid ability levels. In some cases these tests contain items that are difficult even for healthy elderly adults. Whether these changes in apparent item difficulty are due to actual age based declines in function or to continually evolving culturally based cohort effects (Schaie et al., 1973) it is likely that task sensitivity would be maximized using items designed around the normative capabilities of healthy older adults. Finally,

using an age appropriate set of items would not only permit the overall length of test instruments to be shortened considerably but is likely to make the experience of being tested less frustrating for the patients in question. Although more difficult to implement, a similar strategy could be used to customize the range of test items around the estimated premorbid abilities of the patient. Premorbid intellectual abilities can be estimated only coarsely using such demographic data as education and occupation (Barona, Reynolds, & Chastain, 1984; Sweet, Moberg, & Tovian, 1990) and such IQ related tasks as irregular word reading (Neslon & O'Connell, 1978) that are relatively resilient to the effects of brain dysfunction. However, even if estimates based on such information were used to grossly classify patients into two or three broad categories based on premorbid ability, the effect on the focus and efficiency of tests items could be significant. This strategy of selectively ''bracketing'' test items to optimize sensitivity and efficiency of test items has not been yet been empirically tested.

Simple age based test norms may be insensitive to premorbid variation in ability and cohort effects. The use of age and premorbid level appropriate items would very likely result in a reduction in ''false positive'' classifications of cognitive changes that may be age appropriate for the individual being assessed.

A PRELIMINARY SET OF CRITERIA FOR THE EVALUATION AND DEVELOPMENT OF COGNITIVE MEASURES OF DEMENTIA

The discussion thus far can be used to suggest a set of idealized characteristics for clinical measures of cognitive functioning to be used in patients with suspected dementia. Although listed separately, the seven criteria listed below, as has been argued, cannot under most circumstances be achieved independently of one another. For example, the choice of cognitive domain as suggested in Criterion 2 should reflect the neuropatholgical constraints outlined in Criteria 1. In some cases desirable properties may be incompatible with each other so that test construction may involve a compromise to optimize the overall quality of the instrument. For example, the desire for reliability may countervail the desire for efficiency. In other cases, however, it may be desirable to optimize the battery for a specific purpose even though other criteria may be only minimally met. For example, one could construct measures that are highly predictive of specific daily activities, with high ecological validity, though biological and cognitive specificity may be sacrificed. Though most of these recommendations fit into the traditional notions of test validity and reliability, they are presented in the context of the specific demands of cognitive assessment with dementia described above:

1. Neuropathological Sensitivity/Specificity: Ability to distinguish patients who show independent radiological or postmortem evidence of a neuropathological entity. Included under this criterion should be differential

sensitivity to localizable changes in cortical and subcortical function. More controversially this criterion might include the ability to differentiate etiological agents (e.g., Alzheimer's disease versus vascular dementia).

2. Cognitive Domain Specificity: Measures should be relatively homogeneous and designed to assess different cognitive domains. The choice of these domains should be justified on both biological and psychological grounds.
3. Construct or Process Specificity: Tasks should assess empirically and/or theoretically justified variables that reflect the information processing demands specific to each domains. The tasks should reflect our most recent understanding of psychological processes.
4. Functional Specificity: Tasks should, at least in principal, be relevant to a patient's daily functional abilities.
5. Contextually Appropriate Reliability: Reliability for a given measure should be estimated for the clinical population and within the assessment context the task is likely to appear. Task length may be ''titrated'' to optimize the reliability of a measure as it will actually be used.
6. Age Appropriate Item Difficulty: The tasks should be appropriate in form and difficulty level for healthy adults in the age range of the patients being assessed.
7. Efficiency: The test should be as brief as possible given the constraints outlined above.

A BRIEF REVIEW OF SOME CURRENT MEASURES OF COGNITIVE FUNCTION

Instruments for the assessment of cognitive function that are currently available may be classified into three categories: full neuropsychological batteries, screening instruments, and an emergent intermediate category that I refer to as ''microbatteries.'' The evaluations described below are summarized in Table 1. The ratings are merely suggestive and are not based on quantitative or formal metaanalytic methods.

Examples of commonly used neuropsychological batteries include the Halstead-Reitan Battery (Halstead, 1947; Reitan & Davison, 1974) and its derivatives, the Luria-Nebraska Neuropsychological Battery (Golden et al., 1980), and though less strictly a fixed battery, the Boston Process Approach (Milberg et al., 1986). These batteries have in common that they were developed by neuropsychologists to be generally sensitive to the cognitive sequelae of brain damage, regardless of etiology. For the most part the tests used in the Halstead-Reitan Battery and in the Boston Process approach were developed as psychometric instruments for purposes other than the assessment of the effects of brain damage. These tests usually were included in the battery because of empirical association with the presence of brain pa-

thology or for the possibility of extracting information that could be attributed to brain pathology. Measures did not typically become part of these batteries because independent theory or data suggested a priori that the functions represented by those test should be assessed. Interestingly, the Luria-Nebraska (Golden et al., 1980) battery is composed of a series of tasks that were in fact designed to assess the specific effects of brain damage in an interlocking decision tree-like format. Ironically, in the context of the Luria-Nebraska Battery, however, these measures are collapsed into scales and validated in a manner that arguably does not capture their original intent or design (Golden et al., 1980). Typically neuropsychological batteries take 2–8 hr to administer. Administration time is likely to be longer in older patients with dementia, particularly if testing must be broken up into brief sessions that fall within the attentional capabilities of the patient. Specific contextual reliability figures for most of these measures is not available, though the individual measures generally have high reliability when administered to young adult subjects without brain damage (Golden et al., 1980). Because these batteries were not designed with efficiency in mind, there are no data available on the optimal length of the constituent tests. It is safe to say, however, that most of the tests used in currently popular batteries are not sufficiently domain specific and even fewer were developed to reflect the current research literature defining the critical behavioral variables making up those domains. In addition the tests in most of these batteries cover a wide range of ability levels and were typically not designed around the ability levels of older adults. For these reasons it is likely that most neuropsychological batteries are not efficient implements for gathering specific information about cognitive functions. It is likely that most measures used as part of neuropsychological batteries could be redesigned to be more specific and brief.

Recently there have been attempts to construct batteries of tests around neurally relevant domains of cognitive function with specific relevance to patients with dementia. One such battery was developed by the neuropsychology group in the Alzheimer's Disease Research Center at UCSD. This battery consists of detailed measures of attention, memory, abstraction, language, visuospatial abilities, and general intellectual functioning. Although well designed from this standpoint, the battery is nevertheless lengthy (ca. 4 hr) containing possible functional redundancies (e.g., Rey Auditory Verbal Learning Test (Rey, 1964), Buschke-Fuld Selective Reminding Test (Buschke, 1973), Mini-Mental State Exam (Folstein et al., 1975), and Mattis Dementia Rating Scale (Mattis, 1976)). Though appropriate for collecting data for research purposes, and capable of providing detailed information regarding a demented patient's cognitive status, this battery also is not likely to be maximally efficient. Information on contextual reliability is not available.

In addition to being brief, the second category consisting of screening

measures has in common that they are designed around the clinical phenomenology of dementia. Most of these measures are based on techniques popular in the psychiatric or neurological mental status exam and include such measures as the Mini-Mental State Exam (Folstein et al., 1975), the Blessed Dementia Scale (Blessed, Tomlinson, & Roth, 1968), and a number of other similar measures.

The Mini-Mental State Exam is the most widely used of these measures, at least in the United States, and appears to produce stable results across examiners and across moderate retest intervals. Although the Mini-Mental State Exam is designed to produce a global score or rating of mental status change it contains some tasks that might be classified as being ''domain specific.'' However, these domains are not assessed completely or systematically nor do they reflect conceptual details of what is known about the domains sampled. For example, the tasks assessing memory do not adequately assess factors affecting encoding and retrieval, the measures of naming do not reflect word frequency effects or allow for an assessment of nonlinguistic factors that may affect confrontation naming, etc. These tasks therefore do not easily lend themselves to being assessed as independent functions (although there have been some attempts to use it this way (see Brandt, Folstein, Folstein, 1988). There is also some question about the sensitivity of the test, particularly to early stages of dementia and to changes in mental status in patients who functioned at an above average level premorbidly. These limitations are not surprising since the Mini-Mental State Exam is for the most part the formalization of traditional techniques employed in the psychic or neurological mental status examination. Most of these techniques or tasks, whose origins are mainly obscure, were passed down through generations of clinicians usually without being influenced by developments in nonclinical neuroscience and the study of cognition. These measures do have the advantage of being within the capabilities of older adults yielding a very low ''false positive'' rate of classification.

The final category of measures is also relatively brief (i.e., requiring administration times of 30 min or less) and is sometimes considered to be screening measures. The difference, however, between these measures and the brief screening measures described above is that they were designed specifically to allow a independent examination of separate cognitive domains. For this reason these measures will be called ''microbatteries.'' As a group these measures use items appropriate to the age of the patients likely to suffer from dementia. The most widely used of these mcirobatteries is the Dementia Rating Scale (DRS) developed by Mattis (1976) which has been available for nearly 20 years.

The DRS is divided into five subscales sampling the domains of attention, initiation/perseveration, construction, conceptualization, and memory. These scales in part map onto what might be considered cognitive domains in the manner under the current framework (e.g., attention, memory). However,

the scales of initiation/perseveration, conceptualization, and construction represent specific task or symptom designations rather than broad cognitive domains that are biologically specified. There is little specific attention paid to language function but memory is examined in greater detail than most screening batteries. Unfortunately, though these subscales are available, the primary validation of the DRS has been as a global measure. Additional research relating the subscales to specific performance deficits, as measured by other standard cognitive tasks, and to radiological data would be desirable. The global score of the DRS has been used widely in research on dementia. It is a relatively sensitive indicator of the presence of cognitive deficits though it may not have sufficient sensitivity to assess mild dementia in high functioning adults (Vitaliano et al., 1984b). Its internal consistency and reliability are excellent (Vitaliano et al., 1984b). It is not clear how the DRS compares in sensitivity to full neuropsychological batteries though this data may be available soon (Mattis, personal communication). Some of its subscales have been shown to be useful in the prediction of activities of daily living (e.g., Nadler et al., 1993). The potential therefore, for the DRS to ultimately be demonstrated to be an efficient clinical tool, providing at least some specification of cognitive functions in patients with dementia, is good.

Another promising ''microbattery'' is the Neurobehavioral Cognitive Status Examination (NCSE) (Kiernan, Mueller, Langston, & Van Dyke, 1987). This consists of a number of subscales including a clinical rating of patients level of consciousness, orientation, attention, language, constructional ability, memory, calculations, and reasoning. Although these subscales are not all ''domains,'' the sampling of behaviors is quite broad and relevant to the range of cognitive phenomena that is important in the evaluation of patients with dementia. An interesting feature of the NCSE is its use of screening questions designed to determine the presence or absence of a deficit and follow up ''metric'' questions to scale that deficit within each domain. The NCSE is more sensitive to the presence of dementia'' than the Mini-Mental State Exam and is acceptably reliable. Like the DRS the NCSE should be validated against specific markers of neuropathology and in relationship to other specific measures of cognitive functioning. Its ability to detect the presence of neuropathology relative to full neuropsychological batteries is also unknown. It too has excellent potential to be an efficient, specific measure of cognitive functioning in dementia.

There have been other recent attempts to create brief domain specific microbatteries that contain detailed domain specific measures with the potential for impressive levels of diagnostic sensitivity and specificity. These measures include the battery proposed by the Consortium To Establish a Registry for Alzheimer's Disease (CERAD) (Morris, Mohs, Rogers, Fillenbaum, & Heyman, 1988) and more recently The Geriatric Evaluation of Mental Status (The GEMS). The GEMS has been under development for a number of years through the collaborative efforts of neuropsychologists, neurologists, and

geriatricians (Hamann et al., 1994) at the Geriatric Research, Education and Clinical Center of the Brockton/West Roxbury VAMC. The GEMS consists of subscales designed to measure attention, verbal and nonverbal memory, construction, executive functions, and language. The assessment of memory functioning allows a detailed examination of recall, recognition, and encoding strategies. There are tasks designed to assess repetition, confrontation naming, and comprehension. The tests of executive functions include clinical adaptation of the delayed alternation procedure. The GEMs appears to be acceptably reliable and to correlate well with a number of full versions of comparable standard neuropsychological tasks. Early data suggest that it is more sensitive and specific than the Mini-Mental State Exam though it takes only 15 min to administer. The GEMS represents an attempt to integrate current thinking about the individual characteristics of separate cognitive domains and has the potential for providing clinically relevant information in an extremely efficient cost effective manner. Additional data will need to be collected before the utility of this instrument can be fully judged.

CONCLUSION: TOWARD A NEW GENERATION OF COGNITIVE ASSESSMENT INSTRUMENTS

The increasing prevalence of patients presenting with disorders of cognitive functioning, coupled with limits in resources available to address these problems, is going to necessitate the development of new technologies to be used for assessment. These instruments will have to provide specific, clinically relevant information in a reliable, efficient manner. Relatively brief assessment instruments are also desirable because the symptoms of dementia may limit the utility of standard techniques that are tedious or unpleasant to the patient and time consuming. Most currently available techniques for the assessment of dementia were designed without a premium placed on such issues as test length or cognitive specificity.

Though no currently available measures of cognitive functioning completely meet these goals, the field of clinical neuropsychology has advanced considerably in the past 20 years and has available the knowledge base to adapt to the practical demands of assessing patients with dementia. Neuropsychological microbatteries that are already available may represent the best models for a new generation of measures. Though there is a great deal of room for development of this format, these instruments contain subscales designed to assess different aspects of cognitive function and are generally designed using developmentally appropriate tasks.

In the next generation of cognitive measures, efficiency may be purchased by designing these measures around the goals of domain specificity and developmental appropriateness. These measures should reflect current research on brain organization and employ test items appropriate to the age and perhaps the premorbid abilities of the patients being assessed. Measures con-

structed in this manner would also have the advantage of describing differences between patients that reflect differences in etiology and structural impact of the causative lesion. These specific measures would also be more easily applied to the prediction of daily adjustment and to assess treatment.

REFERENCES

American Psychological Association AERA NCME. 1985. *Standards for evaluating and psychological testing.* Washington: APA.

Barona, A., Reynolds, C. R., & Chastain, R. 1984. A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology,* **52,** 885–887.

Blessed, G., Tomlinson, B. E., & Roth, M. 1968. The association between quantitative measures of dementia and senile change in cerebral gray matter of elderly subjects. *Journal of General Psychiatry,* **114,** 797–811.

Brandt, J., Folstein, S. E., & Folstein, M. F. 1988. Differential cognitive impairment in Alzheimer's disease and Huntington's disease. *Annals of Neurology,* **23,** 555–561.

Buschke, H. 1973. Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior,* **12,** 543–550.

Carey, G., & Gottesman, I. I. 1978. Reliability and validity in binary ratings. *Archives of General Psychiatry,* **35,** 1454–1459.

Cohen, J. R., & Zuckerman, H. 1960. A coefficient of agreement for nominal scales. In *Educational and psychological measurement.* Pp. 37–46. Damasio, H., & Damasio, A. R. 1989. *Lesion analysis in neuropsychology.* New York: Oxford University Press.

Cronbach, E. J. 1964. Cronbach's lament: Experimental vs. correlational psychology. In E. J. Cronbach (Eds.), *The two disciplines of scientific psychology.* New York: McGraw–Hill.

Evans, D. A., Funkenstein, H., Albert, M. S., Scherr, P. A., Cook, N. R., Chown, M. J., Hebert, L. E., Henekens, C. H., & Taylor, J. O. 1989. Prevalence of Alzheimer's Disease in a community population of older persons. *Journal of the American Medical Association,* **262,** 2551–2556.

Faraone, S. V., & Tsuang, M. T. 1994. Measuring diagnostic accuracy in the absence of a ''Gold Standard.'' *American Journal of Psychiatry,* **151**(5), 650–657.

Folstein, M. F. 1990. The birth of the MMS. *Current Contents,* **11,** 14.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. 1975. ''Mini-Mental State'': A practical method for grading the cognitive state of outpatients for the clinician. *The Journal of Psychiatric Research,* **12,** 189–198.

Golden, C. J., Hammeke, T. A., & Purisch, A. D. 1980. *The Luria-Nebraska Neuropsychological Battery: Manual.* Los Angeles: Western Psychological Services.

Goodglass, H. 1993. *Understanding aphasia.* Boston: Academic Press.

Halstead, W. C. 1947. *Brain and intelligence.* Chicago: University of Chicago Press.

Hamann, C., McGlichey-Berroth, R., Odenheimer, G. L., Weitzen, S., McDonald, R. A., Berger, M., Kilduff, P. T., Milberg, W. P., & Minaker, K. L. 1994. The next generation of geriatric cognitive screening: The Geriatric Evaluation of Mental Status (GEMS). In *International State of the Art Conference on Comprehensive Geriatric Assessment,* Florence, Italy.

Henderson, V. W., Mack, W., & Williams, B. W. 1989. Spatial disorientation in Alzheimer's disease. *Archives of Neurology,* **46,** 391–394.

Jorm, A. F. 1985. Subtypes of Alzheimer's dementia: A conceptual analysis and critical review. *Psychological Medicine,* **15,** 543–553.

Kertesz, A. 1994. Localization and function: Old issues revisited and new developments. In A. Kertesz (Ed.), *Localization and neuroimaging in neuropsychology.* San Diego: Academic Press.

Kiernan, R. J., Mueller, J., Langston, W., & VanDyke, C. 1987. The Neurobehavioral Cognitive Status Exam: A brief but differentiated approach to cognitive assessment. *Annals of Internal Medicine,* **107,** 481–485.

Magnusson, D. 1966. *Test theory.* Stolkholm: Almqvist & Wiksell/Gebers Forlag AB.

Martin, A., Brouwers, P., & Fedio, P. 1985. A note on the different patterns of impaired and preserved cognitive abilities and their relation to episodic memory deficits in Alzheimer's patients. *Brain and Language,* **25,** 323–341.

Martin, A., Brouwers, P., Lalonde, F., Cox, C., Teleska, P., Fedio, P., Foster, N. L., & Chase, T. N. 1986. Towards a behavioral typology of Alzheimer's patients. *Journal of Clinical and Experimental Neuropsychology,* **8,** 594–610.

Mattis, S. 1976. Mental status examination for organic mental syndrome in the elderly patient. In L. Bellak & T. B. Karasu (Eds.), *Geriatric psychiatry.* New York: Grune & Stratton.

McCarthy, R., & Warrington, E. 1990. *Cognitive neuropsychology: A clinical introduction.* San Diego: Academic Press.

Milberg, W., & Albert, M. 1991. The speed of constituent mental operations and its relationship to neuronal representation: An hypothesis. In R. G. Lister & H. J. Weingertner (Eds.), *Perspectives on cognitive neuroscience.* New York: Oxford University Press.

Milberg, W. P., Hebben, N. A., & Kaplan, E. 1986. The Boston Process Approach to neuropsychological assessment. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders.* New York: Oxford University Press.

Morris, J. C., Mohs, R. C., Rogers, H., Fillenbaum, G., & Heyman, A. 1988. Consortium to establish a registry for Alzheimer's disease (CERAD) clinical and neuropsychological assessment of Alzheimer's disease. *Psychopharmacology Bulletin,* **24**(4), 641–652.

Nadler, J. D., Richardson, E. D., Malloy, P. F., Marran, M. E., & Hostetler Brinson, M. E. 1993. The ability of the Dementia Rating Scale to predict everyday functioning. *Archives of Clinical Neuropsychology,* **8,** 449–460.

Neslon, H. E., & O'Connell, A. 1978. Dementia: The estimation of premorbid intelligence levels using the new adult reading test. *Cortex,* **14,** 234–244.

Parkin, A. J., & Leng, N. R. C. 1993. *Neuropsychology of the amnesic syndrome.* Hillsdale: Lawrence Erlbaum Associates.

Reitan, R., & Davison, L. A. 1974. *Clinical neuropsychology: Current status and applications.* New York: Hemisphere.

Rey, A. 1964. L'examen clinique en psychologie. *Archives de Psychologie,* **28,** 286–340.

Schaie, K. W., Labouvie, G. V., & Buech, B. U. 1973. Generational and cohort-specific differences in adult cognitive function: A fourteen-year study of independent samples. *Developmental Psychology,* **9,** 151–166.

Services, U. S. D. o. H. a. H. 1992. *Chartbook on health data on older americans: United States 1992: Series 3 analytic and epidemiological studies* No. National Center for Health Statistics.

Shallice, T., & Burgess, P. 1991. Higher-order cognitive impairments and frontal lobe lesions in man. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction.* New York: Oxford University Press.

Spearman, C. 1910. Correlation calculated from faulty data. *British Journal of Psychology,* **3,** 271–295.

Spencer, G. 1989. *Projections of the population of the United States by age, sex, and race 1988 to 2030.* US Department of Commerce.

Sweet, J. J., Moberg, P. J., & Tovian, S. M. 1990. Evaluation of Wechsler Adult Intelligence Scale-Revised premorbid IQ formulas in clinical populations. *Psychological Assessment,* **2,** 41–44.

Vitaliano, P. P., Breen, A. R., Albert, M. S., Russo, J., & Prinz, P. N. 1984a. Memory, attention and functional status in community-residing Alzheimer type dementia patients and optimally healthy aged individuals. *Journal of Gerontology,* **39**(1), 58–64.

Vitaliano, P. P., Breen, A. R., Russo, J., Albert, M., Vitiello, M., & Prinz, P. N. 1984b. The

clinical utility of the dementia rating scale for assessing Alzheimer patients. *Journal of Chronic Diseases,* **37**(9/10), 743–753.

Weintraub, S., Baratz, R., & Mesulum, M. M. 1982. Daily living activities in the assessment of dementia. In S. Corkin (Ed.), *Alzheimer's disease: A report of progress (Aging, Vol. 19).* New York: Raven Press. Pp. 189–192.