NEUROPSYCHOLOGICAL ASSESSMENT

Fifth Edition

Muriel Deutsch Lezak
Diane B. Howieson
Erin D. Bigler
Daniel Tranel

OXFORD
UNIVERSITY PRESS

6 The Neuropsychological Examination: Interpretation

THE NATURE OF NEUROPSYCHOLOGICAL EXAMINATION DATA

The basic data of psychological examinations, like any other psychological data, are behavioral observations. In order to get a broad and meaningful sample of the patient's behavior from which to draw diagnostic inferences or conclusions relevant to patient care and planning, the psychological examiner needs to have made or obtained reports of many different kinds of observations, including historical and demographic information.

Different Kinds of Examination Data

Background data

Background data are essential for providing the context in which current observations can be best understood. In most instances, accurate interpretation of the patient's examination behavior and test responses requires at least some knowledge of the developmental and medical history, family background, educational and occupational accomplishments (or failures), and the patient's current living situation and level of social functioning. The examiner must take into account a number of patient variables when evaluating test performances, including sensory and motor status, alertness cycles and fatigability, medication regimen, and the likelihood of drug or alcohol dependency. An appreciation of the patient's current medical and neurological status can guide the examiner's search for a pattern of neuropsychological deficits.

The importance of background information in interpreting examination observations is obvious when evaluating a test score on school-related skills such as arithmetic and spelling or in the light of a vocational history that implies a particular performance level (e.g., a journeyman millwright must be of at least *average* ability but is more likely to achieve *high average* or even better scores on many tests; to succeed as an executive chef requires at least *high average* ability but, again, many would perform at a *superior* level on cognitive tests). However, motivation to reach a goal is also important: professionals can be of *average* ability while an individual with exceptional ability might be a shoe clerk. The contributions of such background variables as age or education to test performance have not always been appreciated in the interpretation of many different kinds of tests, including those purporting to measure neuropsychological integrity (e.g., not PsychCorp, 2008a; nor Reitan and Wolfson, 1995b; nor Wechsler, 1997a; 1997b provide education data for computed scores or score conversions on any tests).

Behavioral observations

Naturalistic observations can provide very useful information about how the patient functions outside the formalized, usually highly structured, and possibly intimidating examination setting. Psychological examiners rarely study patients in their everyday setting yet reports from nursing personnel or family members may help set the stage for evaluating examination data or at least raise questions about what the examiner observes or should look for.

The value of naturalistic observations may be most evident when formal examination findings alone would lead to conclusions that patients are more or less capable than they actually are (Capitani, 1997; Newcombe, 1987). Such an error is most likely to occur when the examiner confounds observed *performance* with *ability*. For example, many people who survive even quite severe head trauma in moving vehicle accidents ultimately achieve scores that are within or close to the *average* ability range on most tests of cognitive function (Crosson, Greene, Roth, et al., 1990; H.S. Levin, Grossman, Rose, and Teasdale, 1979; Ruttan et al., 2008). Yet, by some accounts, as few as one-third of them hold jobs in the competitive market as so many are troubled by problems of attention, temperament, and self-control (Bowman, 1996; Cohadon et al., 2002; Hoofien, Vakil, Cohen, and Sheleff, 1990; Lezak and O'Brien, 1990). The behavioral characteristics that compromise their adequate and sometimes even excellent cognitive skills are not elicited in the usual neuropsychiatric or neuropsychological examination. Mesulam (1986) reviewed several cases of patients with frontal lobe damage who exhibited no cognitive deficits on formal neuropsychological examination (see follow-up by Burgess, Alderman, and colleagues, 2009). However, these deficits become painfully apparent to anyone who is with these patients as they go about their usual activities—or, in many cases, inactivities. In contrast, there is the shy, anxious, or suspicious patient who responds only minimally to a white-coated examiner but whose everyday behavior is far superior to anything the examiner sees; and also patients whose coping strategies enable them to function well *despite* significant cognitive deficits (B.A. Wilson, 2000; R.L. Wood, Williams, and Kalyani, 2009).

How patients conduct themselves in the course of the examination is another source of useful information. Their comportment needs to be documented and evaluated as attitudes toward the examination, conversation or silence, the appropriateness of their demeanor and social responses, can tell a lot about their neuropsychological status as well as enrich the context in which their responses to the examination proper will be evaluated.

Test data

> In a very real sense there is virtually no such thing as a neuropsychological test. Only the method of drawing inferences about the tests is neuropsychological.
>
> *K.W. Walsh, 1992*

Testing differs from these other forms of psychological data gathering in that it elicits behavior samples in a standardized, replicable, and more or less artificial and restrictive situation (S.M. Turner et al., 2001; Urbina, 2004). Its strengths lie in the approximate sameness of the test situation for each subject, for it is the sameness that enables the examiner to compare behavior samples between individuals, over time, or with expected performance levels. Its weaknesses too lie in the sameness, in that psychological test observations are limited to the behaviors prompted by the test situation.

To apply examination findings to the problems that trouble the patient, the psychological examiner extrapolates from a limited set of observations to the patient's behavior in real-life situations. Extrapolation from the data is a common feature of other kinds of psychological data handling as well, since it is rarely possible to observe a human subject in every problem area. Extrapolations are likely to be as accurate as the observations on which they are based are pertinent, precise, and comprehensive, as the situations are similar, and as the generalizations are apt.

A 48-year-old advertising manager with originally *superior* cognitive abilities sustained a right hemisphere stroke with minimal sensory or motor deficits. He was examined at the request of his company when he wanted to return to work. His verbal skills in general were *high average* to *superior*, but he was unable to construct two-dimensional geometric designs with colored blocks, put together cut-up picture puzzles, or draw a house or person with proper proportions (see Fig. 6.1). The neuropsychologist did not observe the patient on the job but, generalizing from these samples, she concluded that the visuoperceptual distortions and misjudgments demonstrated on the test would be of a similar kind and would occur to a similar extent with layout and design material. The patient was advised against retaining responsibility for the work of the display section of his department. Later conferences with the patient's employers confirmed that he was no longer able to evaluate or supervise the display operations.

In most instances examiners rely on their commonsense judgments and practical experiences in making test-based predictions about their patients' real-life functioning. Studies of the *predictive validity* and *ecological validity* of neuropsychological tests show that many of them have a good predictive relationship with a variety of disease characteristics (e.g., pp. 125–126) and practical issues (see p. 126).

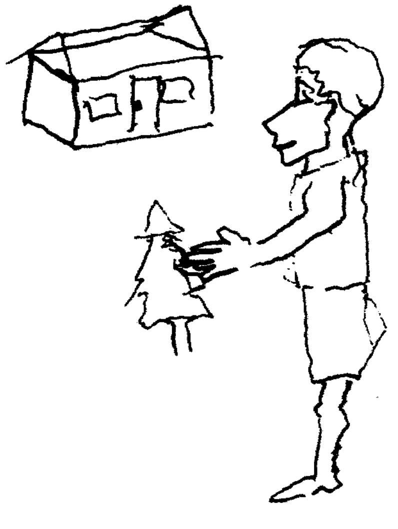


FIGURE 6.1 House-Tree-Person drawings of the 48-year-old advertising manager described in the text (size reduced to one-third of original).

Quantitative and Qualitative Data

Every psychological observation can be expressed either numerically as quantitative data or descriptively as qualitative data. Each of these classes of data can constitute a self-sufficient data base as demonstrated by two different approaches to neuropsychological assessment. An actuarial system (Reitan, 1966; Reitan and Wolfson, 1993)—elaborated by others (e.g., Heaton, Grant, and Matthews, 1991; J.A. Moses, Jr., Pritchard, and Adams, 1996, 1999)—exemplifies the quantitative method. It relies on scores, derived indices, and score relationships for diagnostic predictions. Practitioners using this method may have a technician examine the patient so that, except for an introductory or closing interview, their data base is in numerical, often computer-processed, form. At the other extreme is a clinical approach built upon richly described observations without objective standardization (A.-L. Christensen, 1979; Luria, 1966). These clinicians documented their observations in careful detail, much as neurologists or psychiatrists describe what they observe.

Both approaches have contributed significantly to the development of contemporary neuropsychology (Barr, 2008). Together they provide the observational frames of reference and techniques for taking into account, documenting, and communicating the complexity, variability, and subtleties of patient behavior. Although some studies suggest that reliance on actuarial evaluation of scores alone provides the best approach to clinical diagnosis (R.M. Dawes, Faust, and Meehl, 1989), this position has not been consistently supported in neuropsychology (Cimino, 1994; Heaton, Grant, Anthony, and Lehman, 1981; Ogden-Epker and Cullum, 2001). Nor is it appropriate for many—perhaps most—assessment questions in neuropsychology, as only simple diagnostic decision making satisfies the conditions necessary for actuarial predictions to be more accurate than clinical ones: (1) that there be only a small number of probable outcomes (e.g., left cortical lesion, right cortical lesion, diffuse damage, no impairment); (2) that the prediction variables be known (which limits the amount of information that can be processed by an actuarial formula to the information on which the formula was based); and (3) that the data from which the formula was derived be relevant to the questions asked (American Academy of Clinical Neuropsychology, 2007; Pankratz and Taplin, 1982).

Proponents of purely actuarial evaluations overlook the realities of neuropsychological practice in an era of advanced neuroimaging and medical technology: most assessments are not undertaken for diagnostic purposes but to describe the patient's neuropsychological status. Even in those instances in which the examination is undertaken for diagnostic purposes the issue is more likely to concern diagnostic discrimination requiring consideration of a broad range of disorders—including the possibility of more than one pathological condition being operative—than making a decision between three or four discrete alternatives. Moreover, not infrequently diagnosis involves variables that are unique to the individual case and not necessarily obvious to a naive observer or revealed by questionnaires, variables for which no actuarial formulas have been developed or are ever likely to be developed (Barth, Ryan, and Hawk, 1992).

It is also important to note that the comparisons in most studies purporting to evaluate the efficacy of clinical versus actuarial judgments are not presenting the examiners with real patients with whom the examiner has a live interaction, but rather with the scores generated in the examination—and just the scores, without even descriptions of the qualitative aspects of the performance (e.g., Faust, Hart, and Guilmette, 1988a; Faust, Hart, Guilmette, and Arkes, 1988b; see also this page). This debate has extended into one concerning "fixed" versus "flexible" approaches (Larrabee, Millis, and Meyers, 2008). Practical judgment and clinical experience supports the use of a "flexible" selection of tests to address the referral question(s) and problems/issues raised in neuropsychological consultation (American Academy of Clinical Neuropsychology, 2007).

Quantitative data

> The number is not the reality, it is only an abstract symbol of some part or aspect of the reality measured. The number is a reduction of many events into a single symbol. The reality was the complex dynamic performance.
>
> *Lloyd Cripe, 1996a, p. 191*

Scores are summary statements about observed behavior. Scores may be obtained for any set of behavior samples that can be categorized according to some principle. The scorer evaluates each behavior sample to see how well it fits a predetermined category and then gives it a place on a numerical scale (Urbina, 2004).

A commonly used scale for individual test items has two points, one for "good" or "pass" and the other for "poor" or "fail." Three-point scales, which add a middle grade of "fair" or "barely pass," are often used for grading ability test items. Few item scales contain more than five to seven scoring levels because the gradations become so fine as to be confusing to the scorer and meaningless for interpretation. Scored tests with more than one item produce a summary score that is usually the simple sum of the scores for all the individual items. Occasionally, test-makers incorporate a correction for guessing into their scoring systems so that the final score is not just a simple summation.

Thus, a final test score may misrepresent the behavior under examination on at least two counts: It is based on only one narrowly defined aspect of a set of behavior samples, and it is two or more steps removed from the original behavior. "Global," "aggregate," or "full-scale" scores calculated by summing or averaging a set of test scores are three to four steps removed from the behavior they represent.

Summary index scores based on item scores that have had their normal range restricted to just two points representing either pass or fail, or "within normal limits" or "brain damaged," are also many steps removed from the original observations. Thus "index scores," which are based on various combinations of scores on two or more—more or less similar—tests suffer the same problems as any other summed score in that they too obscure the data. One might wonder why index scores should exist at all: if the tests entering into an index score are so similar that they can be treated as though they examined the same aspects of cognitive functioning, then two tests would seem unnecessary. On the other hand, if each of two tests produces a different score pattern or normative distribution or sensitivity to particular kinds of brain dysfunction, then the two are different and should be treated individually so that the differences in patient performances on these tests can be evident and available for sensitive test interpretation.

The inclusion of test scores in the psychological data base satisfies the need for objective, readily replicable data cast in a form that permits reliable interpretation and meaningful comparisons. Standard scoring systems provide the means for reducing a vast array of different behaviors to a single numerical system (see pp. 165–167). This standardization enables the examiner to compare the score of any one test performance of a patient with all other scores of that patient, or with any group or performance criteria.

Completely different behaviors, such as writing skills and visual reaction time, can be compared on a single numerical scale: one person might receive a high score for elegant penmanship but a low one on speed of response to a visual signal; another might be high on both kinds of tasks or low on both. Considering one behavior at a time, a scoring system permits direct comparisons between the handwriting of a 60-year-old stroke patient and that of school children at various grade levels, or between the patient's visual reaction time and that of other stroke patients of the same age.

Problems in the evaluation of quantitative data

> To reason—or do research—only in terms of scores and score-patterns is to do violence to the nature of the raw material.
>
> *Roy Schafer, 1948*

When interpreting test scores it is important to keep in mind their artificial and abstract nature. Some examiners come to equate a score with the behavior it is supposed to represent. Others prize standardized, replicable test scores as "harder," more "scientific" data at the expense of unquantified observations. Reification of test scores can lead the examiner to overlook or discount direct observations. A test-score approach to psychological assessment that minimizes the importance of qualitative data can result in serious distortions in the interpretations, conclusions, and recommendations drawn from such a one-sided data base.

To be neuropsychologically meaningful, a test score should represent as few kinds of behavior or dimensions of cognitive functions as possible. The simpler the test task, the clearer the meaning of scored evaluations of the behavior elicited by that task. Correspondingly, *it is often difficult to know just what functions* contribute to a score obtained on a complex, multidimensional test task without appropriate evaluation based on a search for commonalities in the patient's performances on different tests, hypotheses generated from observations of the qualitative features of the patient's behavior, and the examiner's knowledge of brain–behavior relationships and how they are affected by neuropathological conditions (Cipolotti and Warrington, 1995; Darby and Walsh, 2005; Milberg, Hebben, and Kaplan, 1996).

If a score is overinclusive, as in the case of summed or averaged test battery scores, it becomes virtually impossible to know just what behavioral or cognitive characteristic it represents. Its usefulness for highlighting differences in ability and skill levels is nullified, for the patient's behavior is hidden behind a hodgepodge of cognitive functions and statistical manipulations (J.M. Butler et al., 1963; A. Smith, 1966). N. Butters (1984b) illustrated this problem in reporting that the "memory quotient" (MQ) obtained by summing and averaging scores on the Wechsler Memory Scale (WMS) was the same for two groups of patients, each with very different kinds of memory disorders based on very different neuropathological processes. His conclusion that "reliance on a single quantitative measure of memory ... for the assessment of amnesic symptoms may have as many limitations as does the utilization of an isolated score ... for the full description of aphasia" (p. 33) applies to every other kind of neuropsychological dysfunction as well. The same principle of multideterminants holds for single test scores too as similar errors lowering scores in similar ways can occur for different reasons (e.g., attentional deficits, language limitations, motor slowing, sensory deficits, slowed processing, etc.).

Further, the range of observations an examiner can make is restricted by the test. This is particularly the case with multiple-choice paper-and-pencil tests and those that restrict the patient's responses to button pushing or another mechanized activity that limits opportunities for self-expression. A busy examiner may not stay to observe the cooperative, comprehending, or docile patient manipulating buttons or levers or taking a paper-and-pencil test. Multiple-choice and automated tests offer no behavior alternatives beyond the prescribed set of responses. Qualitative differences in these test performances are recorded only when there are frank aberrations in test-taking behavior, such as qualifying statements written on the answer sheet of a personality test or more than one alternative marked on a single-answer multiple-choice test. For most paper-and-pencil or automated tests, *how* the patient solves the problem or goes about answering the question remains unknown or is, at best, a matter of conjecture based on such relatively insubstantial information as heaviness or neatness of pencil marks, test-taking errors, patterns of nonresponse, erasures, and the occasional pencil-sketched spelling tryouts or arithmetic computations in the margin.

In addition, the fine-grained scaling provided by the most sophisticated instruments for measuring cognitive competence is not suited to the assessment of many of the behavioral symptoms of cerebral neuropathology. Defects in behaviors that have what can be considered "species-wide" norms, i.e., that occur at a developmentally early stage and are performed effectively by all but the most severely impaired school-aged children, such as speech and dressing, are usually readily apparent. Quantitative norms generally do not enhance the observer's sensitivity to these problems nor do any test norms pegged at adult ability levels when applied to persons with severe defects in the tested ability area. Using a finely scaled vocabulary test to examine an aphasic patient, for example, is like trying to discover the shape of a flower with a microscope: the examiner will simply miss the point. Moreover, behavioral aberrations due to brain dysfunction can be so highly individualized and specific to the associated lesion that their distribution in the population at large, or even in the brain impaired population, does not lend itself to actuarial prediction techniques (W.G. Willis, 1984).

The evaluation of test scores in the context of direct observations is essential when doing neuropsychological assessment. For many brain impaired patients, test scores alone give relatively little information about the patient's functioning. The meat of the matter is often *how* a patient solves a problem or approaches a task rather than what the score is. "There are many reasons for failing and there are many ways you can go about it. And if you don't know in fact which way the patient was going about it, failure doesn't tell you very much" (Darby and Walsh, 2005). There can also be more than one way to pass a test.

> A 54-year-old sales manager sustained a right frontal lobe injury when he fell as a result of a heart attack with several moments of cardiac arrest. On the Hooper Visual Organization Test, he achieved a score of 26 out of a possible 30, well within the normal range. However, not only did his errors reflect perceptual fragmentation (e.g., he called a cut-up broom a "long candle in holder"), but his correct responses were also fragmented (e.g., "wrist and hand and fingers" instead of the usual response, "hand"; "ball stitched and cut" instead of "baseball").
>
> Another patient, a 40-year-old computer designer with a seven-year history of multiple sclerosis, made only 13 errors on the Category Test (CT), a number considerably lower than the 27 error mean reported for persons at his very high level of mental ability (Mitrushina, Boone, et al., 2005). (His scores on the Gates-MacGinitie Vocabulary and Comprehension tests were at the 99th percentile; WAIS-R Information and Arithmetic age-graded scaled scores were in the *very superior* and *superior* ranges, respectively.) On two of the more difficult CT subtests he figured out the response principle within the first five trials, yet on one subtest he made 4 errors after a run of 14 correct answers and on the other he gave 2 incorrect responses after 15 correct answers. This error pattern suggested difficulty keeping in mind solutions that he had figured out easily enough but lost track of while performing the task. Nine repetitions on the first five trials of the Auditory Verbal Learning Test and two serial subtraction errors unremarked by him, one on subtracting "7s" when he went from "16" to "19," the other on the easier task of subtracting 3s when he said "23, 21," further supported the impression that this graduate engineer "has difficulty in monitoring his mental activity ... and [it] is probably difficult for him to do more than one thing at a time." (K. Wild, personal communication, 1991).

This latter case also illustrates the relevance of education and occupation in evaluating test performances since, by themselves, all of these scores are well *within normal limits*, none suggestive of cognitive dysfunction.

Moreover, "Different individuals may obtain the same test score on a particular test for very different reasons" (C. Ryan and Butters, 1980b). Consider two patients who achieve the same score on the WIS-A Arithmetic test but may have very different problems and abilities with respect to arithmetic. One patient performs the easy, single operation problems quickly and correctly but fails the more difficult items requiring two operations or more for solution because of an inability to retain and juggle so much at once in his immediate memory. The other patient has no difficulty remembering item content. She answers many of the simpler items correctly but very slowly, counting aloud on her fingers. She is unable to conceptualize or perform the operations on the more difficult items. The numerical score masks the disparate performances of these patients. As this test exemplifies, what a test actually is measuring may not be what its name suggests or what the test maker has claimed for it: while it is a test of arithmetic ability for some persons with limited education or native learning ability, the WIS-A Arithmetic's oral format makes it a test of attention and short-term memory for most adults, a feature that is now recognized by the test maker (PsychCorp, 2008a; Wechsler, 1997a; see also p. 657). Walsh (1992) called this long-standing misinterpretation of what Arithmetic was measuring, "The Pitfall of Face Validity."

The potential for error when relying on test scores alone is illustrated in two well-publicized studies on the clinical interpretation of test scores.

> Almost all of the participating psychologists drew erroneous conclusions from test scores faked by three preadolescents and three adolescents, respectively (Faust et al., 1988a; 1988b). Although the investigators used these data to question the ability of neuropsychological examiners to detect malingering, their findings are open to two quite different interpretations: (1) *Valid interpretations of neuropsychological status cannot be accomplished by reliance on scores alone.* Neuropsychological assessment requires knowledge and understanding of how the subject performed the tests, of the circumstances of the examination—why, where, when, what for—and of the subject's appreciation of and attitudes about these circumstances. The psychologist/subjects of these studies did not have access to this information and apparently did not realize the need for it.
> (2) *Training, experience, and knowledge are prerequisites for neuropsychological competence.* Of 226 mailings containing the children's protocols that were properly addressed, only seventy-seven (34%) "usable ones" were returned; of the adolescent study, again only about one-third of potential judges completed the evaluation task. The authors made much of the 8+ years of practice in neuropsychology claimed by these respondent-judges, but they noted that in the child study only "about 17%" had completed formal postdoctoral training in neuropsychology, and in the adolescent study this number dropped to 12.5%. They did not report how many diplomates of the American Board of Professional Psychology in Neuropsychology participated in each study. (Bigler [1990b] found that only one of 77 respondents to the child study had achieved diplomate status!); nor did they explain that any psychologist can claim to be a neuropsychologist with little training and no supervision. An untrained person can be as neuropsychologically naive in the 8th or even the 16th year of practice as in the first. Those psychologists who were willing to draw clinical conclusions from this kind of neuropsychological numerology may well have been less well-trained or knowledgeable than the greater number of psychologists who actively declined or simply did not send in the requested judgments. (I was one who actively declined [mdl].)

Qualitative data

Qualitative data are direct observations. In the formal neuropsychological examination these include observations of the patient's test-taking behavior as well as test behavior per se. Observations of patients' appearance, verbalizations, gestures, tone of voice, mood and affect, personal concerns, habits, and idiosyncrasies can provide a great deal of information about their life situation and overall adjustment, as well as attitudes toward the examination and the condition that brings them to it. More specific to the test situation are observations of patients' reactions to the examination itself, their approach to different kinds of test problems, and their expressions of feelings and opinions about how they are performing. Observations of the manner in which they handle test material, the wording of test responses, the nature and consistency of errors and successes, fluctuations in attention and perseverance, emotional state, and the quality of performance from moment to moment as they interact with the examiner and with the different kinds of test material are the qualitative data of the test performance itself (Milberg, Hebben, and Kaplan, 2009).

Limitations of qualitative data

Distortion or misinterpretation of information obtained by direct observation results from different kinds of methodological and examination problems. All of the standardization, reliability, and validity problems inherent in the collection and evaluation of data by a single observer are ever-present threats to objectivity (Spreen and Risser, 2003, p. 46). In neuropsychological assessment, the vagaries of neurological impairment compound these problems. When the patient's communication skills are questionable, examiners can never be certain that they have understood their transactions with the patient—or that the patient has understood them. Worse yet, the communication disability may be so subtle and well masked by the patient that the examiner is not aware of communication slips. There is also the likelihood that the patient's actions will be idiosyncratic and therefore unfamiliar and subject to misunderstanding. Some patients may be entirely or variably uncooperative, many times quite unintentionally.

Moreover, when the neurological insult does not produce specific defects but rather reduces efficiency in the performance of behaviors that tend to be normally distributed among adults, such as response rate, recall of words or designs, and ability to abstract and generalize, examiners benefit from scaled tests with standardized norms. The early behavioral evidence of a deteriorating disease and much of the behavioral expression of traumatic brain injury or little strokes can occur as a quantifiable diminution in the efficiency of the affected system(s) rather than as a qualitative distortion of the normal response. A pattern of generalized diminished function can follow conditions of rapid onset, such as trauma, stroke, or certain infections, once the acute stages have passed and the first vivid and highly specific symptoms have dissipated. In such cases it is often difficult if not impossible to appreciate the nature or extent of cognitive impairment without recourse to quantifiable examination techniques that permit a relatively objective comparison between different functions.

By and large, as clinicians gain experience with many patients from different backgrounds, representing a wide range of abilities, and suffering from a variety of cerebral insults, they are increasingly able to estimate or at least anticipate the subtle deficits that show up as lowered scores on tests. This sharpening of observational talents reflects the development of internalized norms based on clinical experience accumulated over the years.

Blurring the line between quantitative and qualitative evaluations

Efforts to systematize and even enhance observation of how subjects go about failing—or succeeding—on tests have produced a potentially clinically valuable hybrid: quantification of the qualitative aspects of test responses (Poreh, 2000). Glozman (1999) showed how the examination procedures considered to be most qualitative (i.e., some of Luria's examination techniques) can be quantified and thus adaptable for retest comparisons and research. She developed a 6-point scale ranging from 0 (no symptoms) to 3 (total failure), with half-steps between 0 and 1 and 2 to document relatively subtle differences in performance levels.

Other neuropsychologists have developed systems for scoring qualitative features. Joy, Fein and colleagues (2001) demonstrated this hybrid technique in their analysis of Block Design (WIS-A) performances into specific components that distinguish good from poor solutions. Based on their observations, they devised a numerical rating scheme and normed it on a large sample of healthy older (50 to 90 years of age) subjects, thus providing criteria for normal ranges of error types for this age group. Joy and his colleagues emphasized that the purely quantitative "pass–fail" scoring system does not do justice to older subjects who may copy most but not quite all of a design correctly. Similarly, Hubbard and colleagues (2008) used a mixture of quantitative and qualitative measures to assess performance of clock drawing performance in cognitively normal elderly persons (55 to 98 years of age). These measures provide a comparison for evaluating a number of neuropsychological functions including visuoconstructive and visuospatial as well as language skills and hemiattention. This type of scoring for qualitative features allows the clinician to make judgments based on the qualitative aspects of a patient's performance while supporting clinical judgment with quantitative data.

Quantified qualitative errors provide information about lateralized deficits that summary scores alone cannot give. For example, quantifying broken configuration errors on Block Design discriminated seizure patients with left hemisphere foci from those with foci on the right as the latter made more such errors ($p = .008$) although the raw score means for these two groups were virtually identical (left, 26.6 ± 12.4; right, 26.4 ± 12.8) (Zipf-Williams et al., 2000). Perceptual fragmentation (naming a part rather than the whole pictured puzzle) on the Hooper Visual Organization Test was a problem for more right than left hemisphere stroke patients, while the reverse was true for failures in providing the correct name of the picture (Merten, Volkel, and Dornberg, 2007; Nadler, Grace, et al., 1996, see p. 400).

Methods for evaluating strategy and the kinds of error made in copying the Complex Figure have been available for decades (see pp. 582–584). Their score distributions, relationships to recall scores, interindividual variability, and executive function correlates were evaluated by Troyer and Wishart (1997) who recommended that, although not all had satisfactory statistical properties, examiners "may wish to select a system appropriate for their needs."

Integrated data

The integrated use of qualitative and quantitative examination data treats these two different kinds of information as different parts of the whole data base. Test scores that have been interpreted without reference to the context of the examination in which they were obtained may be objective but meaningless in their individual applications. Clinical observations unsupported by standardized and quantifiable testing, although full of import for the individual, lack the comparability necessary for many diagnostic and planning decisions. Descriptive observations flesh out the skeletal structure of numerical test scores. Each is incomplete without the other. The value of taking into account all aspects of a test performance was exemplified in a study comparing the accuracy of purely score-based predictors of lateralization with accuracy based on score profiles plus qualitative aspects of the patient's performance (Ogden-Epker and Cullum, 2001). Accuracy was greatest when qualitative features entered into performance interpretation.

Neuropsychology is rapidly moving into an era where unprecedented clinical information will be available on every patient including genetic, neuroimaging, and other neurodiagnostics studies that ultimately needs to be integrated with the neuropsychological consultation and test findings. Indeed, the era of neuroinformatics contributing to neuropsychological decision making is upon us (Jagaroo, 2010). These kinds of data call for full integration.

Common Interpretation Errors

1. If this, then that: the problem of overgeneralizing

Kevin Walsh (1985) described a not uncommon kind of interpretation error made when examiners overgeneralize their findings. He gave the example of two diagnostically different groups (patients with right hemisphere damage and those with chronic alcoholism) generating the similar cluster of scores, a parallel that led some investigators to conclude that chronic alcoholism somehow shriveled the right but not the left hemisphere (see p. 306). At the individual case level, dementia patients as well as chronic alcoholics can earn depressed scores on the same WIS tests that are particularly sensitive to right hemisphere damage. If all that the examiner attends to is this cluster of low scores, then diagnostic confusion can result. The logic of this kind of thinking "is the same as arguing that because a horse meets the test of being a large animal with four legs [then] any newly encountered large animal with four legs must be a horse" (E. Miller, 1983).

2. Failure to demonstrate a reduced performance: the problem of false negatives

The absence of low scores or other evidence of impaired performance is expected in intact persons but will also occur when brain damaged patients have not been given an appropriate examination (Teuber, 1969). If a function or skill is not examined, its status will remain unknown. And again, the typical neuropsychological examination situation is no substitute for reality in that the examination is undertaken in a controlled environment usually minimizing all extraneous stimuli with assessment being done on a one-to-one basis. This does not replicate the real world circumstances that may be particularly challenging for the neurologically impaired individual.

3. Confirmatory bias

This is the common tendency to "seek and value supportive evidence at the expense of contrary evidence" when the outcome is [presumably] known (Wedding and Faust, 1989).

A neuropsychologist who specializes in blind analysis of Halstead-Reitan data reviewed the case of a highly educated middle-aged woman who claimed neuropsychological deficits as a result of being stunned when her car was struck from the rear some 21 months before she took the examination in question. In the report based on his analysis of the test scores alone the neuropsychologist stated that, "The test results would be compatible with some type of traumatic injury (such as a blow to the head), but they could possibly have been due to some other kind of condition, such as viral or bacterial infection of the brain." After reviewing the history he concluded that although he had suspected an infectious disorder as an alternative diagnostic possibility, the case history that he later reviewed provided no evidence of encephalitis or meningitis, deemed by him to be the most likely types of infection. He thus concluded that the injury sustained in the motor vehicle accident caused the neuropsychological deficits indicated by the test data. Interestingly, the patient's medical history showed that complaints of sensory alterations and motor weakness dating back almost two decades were considered to be suggestive of multiple sclerosis; a recent MRI scan added support to this diagnostic possibility.

4. Misuse of salient data: over- and underinterpretation

Wedding and Faust (1989) made the important point that a single dramatic finding (which could simply be a normal mistake; see Roy, 1982) may be given much greater weight than a not very interesting history that extends over years (such as steady employment) or base rate data. On the other hand, a cluster of a few abnormal examination findings that correspond with the patient's complaints and condition may provide important evidence of a cerebral disorder, even when most scores reflect intact functioning. Gronwall (1991) illustrated this problem using mild head trauma as an example, as many of these patients perform at or near premorbid levels except on tests sensitive to attentional disturbances. If only one or two such tests are given, then a single abnormal finding could seem to be due to chance when it is not.

5. Underutilization or misutilization of base rates

Base rates are particularly relevant when evaluating "diagnostic" signs or symptoms (D. Duncan and Snow, 1987).When a sign occurs more frequently than the condition it indicates (e.g., more people have mild verbal retrieval problems than have early Alzheimer's disease) relying on that sign as a diagnostic indicator "will always produce more errors than would the practice of completely disregarding the sign(s)" (B.W. Palmer, Boone, Lesser, and Wohl, 1998; Wedding and Faust, 1989). Another way of viewing this issue is to regard any sign that can occur with more than one condition as possibly suggestive but never pathognomonic. Such signs can lead to potentially fruitful hypotheses but not to conclusions. Thus, slurred speech rarely occurs in the intact adult population and so is usually indicative of some problem; but whether that problem is multiple sclerosis, a relatively recent right hemisphere infarct, or acute alcoholism—all conditions in which speech slurring can occur—must be determined by some other means. A major limitation in contemporary neuropsychology is that base rate data for neurobehavioral and neurocognitive symptoms/problems is often lacking for a particular disorder or available information is based on inadequate sampling. Proper base rate studies need to be large scale, prospective, done independently with several types of clinical disorders examined within a population. Such in-depth investigations of a neuropsychological variable are rare but necessary.

Compounding the base rate problem is use of inappropriate base rate data which can be as distorting than using no base rate data. For example, G.E. Smith, Ivnik, and Lucas (2008) note the differences in the ratios for identifying probable Alzheimer patients on the basis of a verbal fluency score depending on whether base rate was developed on patients coming to a memory clinic or persons in the general population (see also B.L. Brooks, Iverson, and White, 2007, for base rate variations and ability levels).

6. Effort effects

Both the *American Academy of Clinical Neuropsychology* and the *National Academy of Neuropsychology* have produced position papers supporting the use of *effort testing* in neuropsychological assessment as a means to address the validity of an assessment (S.S. Bush, Ruff, et al., 2005; Heilbronner, Sweet, et al., 2009). Underperformance on neuropsychological measures because of insufficient effort results in a patient's performance appearing impaired when it is not (see Chapter 20).

EVALUATION OF NEUROPSYCHOLOGICAL EXAMINATION DATA

Qualitative Aspects of Examination Behavior

Two kinds of behavior are of special interest to the neuropsychological examiner when evaluating the qualitative aspects of a patient's behavior during the examination. One, of course, is behavior that differs from normal expectations or customary activity for the circumstances. Responding to Block Design instructions by matter-of-factly setting the blocks on the stimulus cards is obviously an aberrant response that deserves more attention than a score of zero alone would indicate. Satisfaction with a blatantly distorted response or tears and agitation when finding some test items difficult also should elicit the examiner's interest, as should statements of displeasure with a mistake unaccompanied by any attempt to correct it. Each of these behavioral aberrations may arise for any number of reasons. However, each is most likely to occur in association with certain neurological conditions and thus can also alert the examiner to look for other evidence of the suspected condition.

Regardless of their possible diagnostic usefulness, these aberrant responses also afford the examiner samples of behavior that, if characteristic, tell a lot about how patients think and how they perceive themselves, the world, and its expectations. The patient who sets blocks on the card not only has not comprehended the instructions but also is not aware of this failure when proceeding—unselfconsciously?—with this display of very concrete, structure-dependent behavior. Patients who express pleasure over an incorrect response are also unaware of their failures but, along with a distorted perception of the task, the product, or both, they demonstrate self-awareness and some sense of a scheme of things or a state of self-expectations that this performance satisfied.

The second kind of qualitatively interesting behaviors deserves special attention whether or not they are aberrant. Gratuitous responses are the comments patients make about their test performance or while they are taking the test, or the elaborations beyond the necessary requirements of a task that may enrich or distort their drawings, stories, or problem solutions, and usually individualize them. The value of gratuitous responses is well recognized in the interpretation of projective test material, for it is the gratuitously added adjectives, adverbs, or action verbs, flights of fancy whether verbal or graphic, spontaneously introduced characters, objects, or situations, that reflect the patient's mood and betray his or her preoccupations. Gratuitous responses are of similar value in neuropsychological assessment. The unnecessarily detailed spokes and gears of a bike with no pedals (see Fig. 6.2) tell of the patient's involvement with details at the expense of practical considerations. Expressions of self-doubt or self-criticism repeatedly voiced during a mental examination may reflect perplexity or depression and raise the possibility that the patient is not performing up to capacity (Lezak, 1978b).

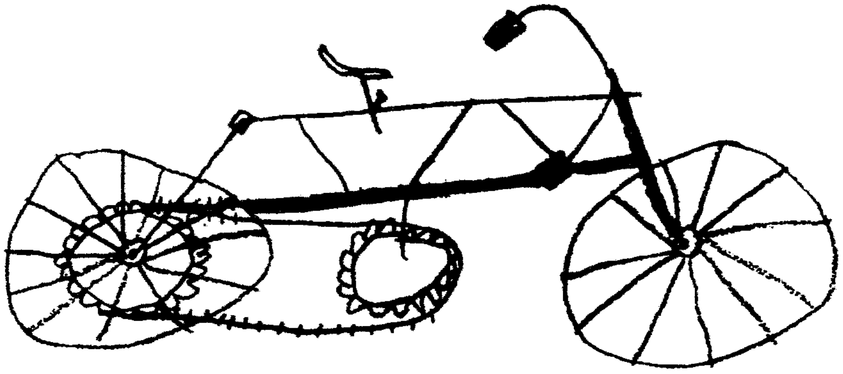


FIGURE 6.2 This bicycle was drawn by a 61-year-old retired millwright with a high school education. Two years prior to the neuropsychological examination he had suffered a stroke involving the right parietal lobe. He displayed no obvious sensory or motor deficits, and was alert, articulate, and cheerful but so garrulous that his talking could be interrupted only with difficulty. His highest WAIS scores, Picture Completion and Picture Arrangement, were in the *high average* ability range.

In addition, patient responses gained by testing the limits or using the standard test material in an innovative manner to explore one or another working hypothesis have to be evaluated qualitatively. For example, in asking a patient to recall a set of designs ordinarily presented as a copy task (e.g., Wepman's variations of the Bender-Gestalt Test, see p. 571) the examiner will look for systematically occurring distortions—in size, angulation, simplifications, perseverations—that, if they did not occur on the copy trial, may shed some light on the patient's visual memory problems. In looking for systematic deviations in these and other drawing characteristics that may reflect dysfunction of one or more behavioral systems, the examiner also analyzes the patient's self-reports, stories, and comments for such qualities as disjunctive thinking, appropriateness of vocabulary, simplicity or complexity of grammatical constructions, richness or paucity of descriptions, etc.

Test Scores

Test scores can be expressed in a variety of forms. Rarely does a test-maker use a *raw* score—the simple sum of correct answers or correct answers minus a portion of the incorrect ones—for in itself a raw score communicates nothing about its relative value. Instead, test-makers generally report scores as values of a scale based on the raw scores made by a *standardization population* (the group of individuals tested for the purpose of obtaining normative data on the test). Each score then becomes a statement of its value relative to all other scores on that scale. Different kinds of scales provide more or less readily comprehended and statistically well-defined standards for comparing any one score with the scores of the standardization population.

B.L. Brooks, Strauss, and their colleagues (2009) review four themes underlying the interpretation and reporting of test scores and neuropsychological findings: (1) the adequacy of the normative data for the test administered; (2) inherent measurement error of any neuropsychological test instrument including ceiling and floor effects; (3) what represents normal variability; and (4) what represents a significant change over time with sequential testing. To make clinical sense out of test data is the focus of neuropsychological assessment and is dependent on the fundamental assumptions discussed below.

Standard scores

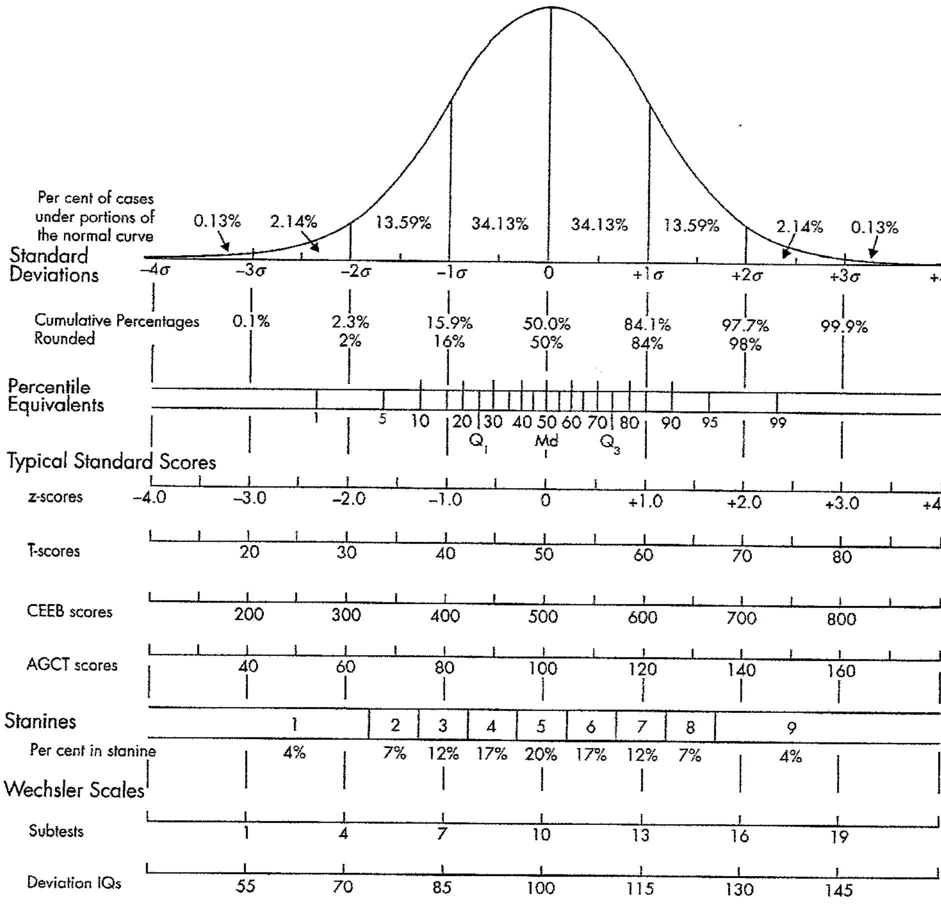
The usefulness of standard scores. The treatment of test scores in neuropsychological assessment is often a more complex task than in other kinds of cognitive evaluations because test scores can come from many different sources. In the usual cognitive examination, generally conducted for purposes of academic evaluation or career counseling, the bulk of the testing is done with one test battery, such as one of the WIS-A batteries or the Woodcock-Johnson Tests of Cognitive Ability. Within these batteries the scores for each of the individual tests are on the same scale and standardized on the same population so that test scores can be compared directly.

On the other hand, no single test battery provides all the information needed for adequate assessment of most patients presenting neuropsychological questions. Techniques employed in the assessment of different aspects of cognitive functioning have been developed at different times, in different places, on different populations, for different ability and maturity levels, with different scoring and classification systems, and for different purposes. Taken together, they are an unsystematized aggregate of more or less standardized tests, experimental techniques, and observational aids that have proven useful in demonstrating deficits or disturbances in some cognitive function or activity. These scores are not directly comparable with one another.

To make the comparisons necessary for evaluating impairment, the many disparate test scores must be convertible into one scale with identical units. Such a scale can serve as a kind of test users' *lingua franca*, permitting direct comparison between many different kinds of measurements. The scale that is most meaningful statistically and that probably serves the intermediary function between different tests best is one derived from the normal probability curve and based on the standard deviation unit (*SD*) (Urbina, 2004) (see Fig. 6.3). Thus the most widely used scale is based on the *standard score.*

The value of basing a common scale on the standard deviation unit lies primarily in the statistical nature of the standard deviation as a measure of the spread or dispersion of a set of scores (X_1, X_2, X_{-3}, etc.) around their mean (*M*). Standard deviation units describe known proportions of the normal probability curve (note on Fig. 6.3, "Percent of cases under portions of the normal curve"). This has very practical applications for comparing and evaluating psychological data in that the position of any test score on a standard deviation unit scale, in itself, defines the proportion of people taking the test who will obtain scores above and below the given score. Virtually all scaled psychological test data can be converted to standard deviation units for intertest comparisons. Furthermore, a score based on the standard deviation, a *standard score*, can generally be estimated from a *percentile*, which is the most commonly used nonstandard score in adult testing (Crawford and Garthwaite, 2009).

The likelihood that two numerically different scores are significantly different can also be estimated from their relative positions on a standard deviation unit scale. This use of the standard deviation unit scale is



NOTE: This chart cannot be used to equate scores on one test to scores on another test. For example, both 600 on the CEEB and 120 on the AGCT are one standard deviation above their respective means, but they do not respresent "equal" standings because the scores were obtained from different groups.

FIGURE 6.3 The relationship of some commonly used test scores to the normal curve and to one another. AGCT, Army General Classification Test; CEEB, College Entrance Examination Board. (Reprinted from the *Test Service Bulletin* of The Psychological Corporation, 1955).

of particular importance in neuropsychological testing, for evaluation of test scores depends upon the significance of their distance from one another or from the comparison standard. Since direct statistical evaluations of the difference between scores obtained on different kinds of tests are rarely possible, the examiner must use estimates of the ranges of significance levels based on score comparisons. In general, differences of two standard deviations or more may be considered significant, whereas differences of one to two standard deviations suggest a trend; although M.J. Taylor and Heaton (2001) accept scores falling at –1 *SD* as indicating deficit.

Kinds of standard scores. Standard scores come in different forms but are all translations of the same scale, based on the mean and the standard deviation The *z-score* is the basic, unelaborated standard score from which all others can be derived. The *z*-score represents, in standard deviation units, the amount a score deviates from the mean of the population from which it is drawn.

$$z = \frac{X - M}{s}$$

The mean of the normal curve is set at zero and the standard deviation unit has a value of one. Scores are stated in terms of their distance from the mean as measured in standard deviation units. Scores above the mean have a positive value; those below the mean are negative. Elaborations of the *z*-score are called *derived scores*. Derived scores provide the same information as do *z*-scores, but the score value is expressed in scale units that are more familiar to most test users than *z*-scores. Test-makers can assign any value they wish to the standard deviation and mean of their distribution of test scores. Usually, they follow convention and choose commonly used values. (Note the different means and standard deviations for tests listed in Fig. 6.3.) When the standardization populations are similar, all of the different kinds of standard scores are directly comparable with one another, the standard deviation and its relationship to the normal curve serving as the key to translation.

Estimating standard scores from nonstandard scores. Since most published standardized tests today use a standard score format for handling the numerical test data, their scores present little or no problem to the examiner wishing to make intertest comparisons. However, a few test makers still report their standardization data in percentile or IQ score equivalents. In these cases, standard score approximations can be estimated. Unless there is reason to believe that the standardization population is not normally distributed, a standard score equivalent for a percentile score can be estimated from a table of normal curve functions. Table 6.1 gives *z*-score approximations, taken from a normal curve table, for 21 percentiles ranging from 1 to 99 in five-point steps. The *z*-score that best approximates a given percentile is the one that corresponds to the percentile closest to the percentile in question.

TABLE 6.1 Standard Score Equivalents for 21 Percentile Scores Ranging from 1 to 99

Percentile Score	*z-Score*	*Percentile Score*	*z-Score*	*Percentile Score*	*z-Score*
99	+2.33	65	+0.39	30	–0.52
95	+1.65	60	+0.25	25	–0.68
90	+1.28	55	+0.13	20	–0.84
85	+1.04	50	0	15	–1.04
80	+0.84	45	–0.13	10	–1.28
75	+0.68	40	–0.25	5	–1.65
70	0.52	35	–0.39	1	–2.33

Exceptions to the use of standard scores

Standardization population differences. In evaluating a patient's performance on a variety of tests, the examiner can only compare scores from different tests when the standardization populations of each of the tests are identical or at least reasonably similar, with respect to both demographic characteristics and score distribution (Axelrod and Goldman, 1996; Mitrushina, Boone, et al., 2005; Urbina, 2004; see Chapter 2). Otherwise, even though their scales and units are statistically identical, the operational meanings of the different values are as different as the populations from which they are drawn. This restriction becomes obvious should an examiner attempt to compare a vocabulary score obtained on a WIS-A test, which was standardized on cross-sections of the general adult population, with a score on the Graduate Record Examination (GRE), standardized on college graduates. A person who receives an *average* score on the GRE would probably achieve scores of one to two standard deviations above the mean on WIS-A tests, since the average college graduate typically scores one to two standard deviations above the general population mean on tests of this type (Anastasi, 1965). Although each of these mean scores has the same *z*-score value, the performance levels they represent are very different.

Test-makers usually describe their standardization populations in terms of sex, race, age, and/or education. Intraindividual comparability of scores may differ between the sexes in that women tend to do less well on advanced arithmetic problems and visuospatial items and men are more likely to display a verbal skill disadvantage (see pp. 362–364). Education, too, affects level of performance on different kinds of tests differentially, making its greatest contribution to tasks involving verbal skills, stored information, and other school-related activities, but affects test performances in all areas (see pp. 360).

Age can be a very significant variable when evaluating test scores of older patients (see pp. 356–360 and Chapters 9–16, *passim*). In patients over 50, the normal changes with age may obscure subtle cognitive changes that could herald an early, correctable stage of a tumor or vascular disease. The use of age-graded scores puts the aging patient's scoring pattern into sharper focus. Age-graded scores are important aids to differential diagnosis in patients over 50 and are essential to the clinical evaluation of test performances of patients over 65. Although not all tests an examiner may wish to use have age-graded norms or age corrections, enough are available to determine the extent to which a patient might be exceeding the performance decrements expected at a given age. An important exception is in the use of age-graded scores for evaluating older persons' performances on tasks which require a minimum level of competence, such as driving (Barrash, Stillman, et al., 2010). This research team found that non-age-graded scores predicted driving impairment better than age-graded ones.

A major debate continues in neuropsychology as to whether significant differences in neuropsychological performance relates to race (Gasquoine, 2009; Manly, 2005). Significant differences between major racial groups have not been consistently demonstrated in the score patterns of tests of various cognitive abilities or in neuropsychological functioning (A.S. Kaufman, McLean, and Reynolds, 1988; Manly, Jacobs, Touradji, et al., 2002; P.E. Vernon, 1979). Nevertheless, there are racial differences in expression of various neurological disorders (Brickman, Schupf, et al., 2008). Race norms have been developed for some standardized neuropsychological measures (Lucas, Ivnik, Smith, et al., 2005), but there are limitations as to how they should be used (Gasquoine, 2009; Manly, 2005).

Vocational and regional differences between standardization populations may also contribute to differences between test norms. Clinicians should always keep in mind that vocational differences generally correlate highly with educational differences, and regional differences tend to be relatively insignificant compared with age and variables that are highly correlated with income level, such as education or vocation.

Children's tests. Some children's tests are applicable to the examination of patients with severe cognitive impairment or profound disability. Additionally, many good tests of academic abilities such as arithmetic, reading, and spelling have been standardized for child or adolescent populations. The best of these invariably have standard score norms that, by and large, cannot be applied to an adult population because of the significant effect of age and education on performance differences between adults and children.

Senior high school norms are the one exception to this rule. On tests of mental ability that provide adult norms extending into the late teens, the population of 18-year-olds does not perform much differently than the adult population at large (e.g., PsychCorp, 2008; Wechsler, 1997a), and four years of high school is a reasonable approximation of the adult educational level. This exception makes a great number of very well-standardized and easily administered paper-and-pencil academic skill tests available for the examination of adults, and no scoring changes are necessary.

All other children's tests are best scored and reported in terms of mental age (MA), which is psychologically the most meaningful score derived from these tests. Most children's tests provide mental age norms or grade level norms (which readily convert into mental age). Mental age scores allow the examiner to estimate the extent of impairment, or to compare performance on different tests or between two or more tests administered over time, just as is done with test performances in terms of standard scores. When test norms for children's tests are given in standard scores or percentiles for each age or set of ages the examiner can convert the score to a mental age score by finding the age at which the obtained score is closest to a score at the 50th percentile or the standard score mean. Mental age scores can be useful for planning educational or retraining programs.

Small standardization populations. A number of interesting and potentially useful tests of specific skills and abilities have been devised for studies of particular neuropsychological problems in which the standardization groups are relatively small (often under 20) (Dacre et al., 2009; McCarthy and Warrington, 1990, *passim*). Standard score conversions are inappropriate if not impossible in such cases. When there is a clear relationship between the condition under study and a particular kind of performance on a given test, there is frequently a fairly clear-cut separation between patient and control group scores. Any given patient's score can be evaluated in terms of how closely it compares with the score ranges of either the patient or the control group reported in the study.

Nonparametric distributions

It is not uncommon for score distributions generated by a neuropsychologically useful test to be markedly skewed—often due to ceiling (e.g., digit span) or floor (e.g., Trail Making Test) effects inherent in the nature of the test and human cognitive capability (Retzlaff and Gibertini, 1994). For these tests, including many used in neuropsychological assessments, standard scores—which theoretically imply a distribution base that reasonably approximates the parametric ideal of a bell-shaped curve—are of questionable value as skewing greatly exaggerates the weight of scores at the far end of a distribution. These distorted distributions produce overblown standard deviations (Lezak and Gray, 1984a [1991]). When this occurs, standard deviations can be so large that even performances that seemingly should fall into the abnormal range appear to be *within normal limits*. The Trail Making Test provides an instructive example of this statistical phenomenon (see Mitrushina, Boone, et al., 2005).

R.K. Heaton, Grant, and Matthews (1986) thoughtfully provided score ranges and median scores along with means and standard deviations of a normative population. Their 20–29 age group's average score on Trails B was 86 ± 39 sec but the range of 47" to 245" with a median score of 76 indicates that many more subjects performed below than above the mean and that the large standard deviation—swollen by a few very slow responders—brings subjects under the umbrella of *within normal limits* who—taking as much as 124" (i.e., < –1 SD) to complete Trails B—do not belong there.

Benton showed the way to resolve the problem of skewed distributions by identifying the score at the 5th percentile as the boundary for abnormality—i.e., *defective* performance (see Benton, Sivan, Hamsher, et al., 1994). Benton and his coworkers used percentiles to define degrees of competency on nonparametric test performances which also avoids the pitfalls of trying to fit nonparametric data into a Procrustean parametric bed.

Evaluation Issues

Norms

Most tests of a single cognitive function, ability, or skill do not have separate norms for age, sex, education, etc. A few widely used tests of general mental abilities take into account the geographic distribution of their standardization population; the rest are usually standardized on local people. Tests developed in Minnesota will have Minnesota norms; New York test makers use a big city population; and British tests are standardized on British populations. Although this situation results in less than perfect comparability between the different tests, in most cases the examiner has no choice but to use norms of tests standardized on an undefined mixed or nonrandom adult sample. Experience quickly demonstrates that this is usually not a serious hardship, for these "mixed-bag" norms generally serve their purpose. "*I sometimes determine SD units for a patient's score* on several norms to see if they produce a different category of performance. Most of the time it doesn't make a significant difference. [If] it does then [one has] to use judgment [H.J. Hannay, 2004, personal communication]." Certainly, one important normative fault of many single-purpose tests is that they lack discriminating norms at the population extremes. Different norms, derived on different samples in different places, and sometimes for different reasons, can produce quite different evaluations for some subjects resulting in false positives or false negatives, depending on the subject's score, condition, and the norm against which the score is compared (Kalechstein et al., 1998; Lezak, 2002).

Thus, finding appropriate norms applicable for each patient is still a challenge for clinicians. Many neuropsychologists collect a variety of norms over the years from the research literature. The situation has improved to some degree in recent years with the publication of collections of norms for many but not all of the most favored tests (Mitrushina, Boone, et al., 2005; E. Strauss, Sherman, and Spreen, 2006). However, there are times when none of these norms really applies to a particular person's performance on a specific test. In such cases, the procedure involves checking against several norm samples to see if a reasonable degree of consistency across norms can be found. When the data from other tests involving a different normative sample but measuring essentially the same cognitive or motor abilities are not in agreement, this should alert the clinician about a problem with the norms for that test as applied to this individual. This problem with norms is very important in forensic cases when the choice of norms can introduce interpretation bias (van Gorp and McMullen, 1997). The final decision concerning the selection of norms requires clinical judgment (S.S. Bush, 2010).

A large body of evidence clearly indicates that demographic variables—especially age and education (and sex and race on some tests)—are related to performance (see data presented in Chapters 9–16, *passim*). Yet some have argued against the use of demographically based norms and suggest that test score adjustment may invalidate the raw test scores (Reitan and Wolfson, 1995b). This argument is based on findings that test performance was significantly related to age and education for normal subjects but not to age and barely for education in a brain damaged group. However, a reduction in the association between demographics and performance is to be expected on a statistical basis for brain damaged individuals.

Suppose that variable X is significantly related to variable Y in the normal population. If a group of individuals is randomly selected from the population, the relationship between variables X and Y will continue to be present in this group. Add random error to one of the variables, for instance Y, and the relationship between X and (Y + random error) will be reduced. Now apply this reasoning to an example bearing on the argument against use of demographic score adjustments. Age is related to performance on a memory test in the normal population. Some individuals, a random sample from the normal population, have a brain disorder and are asked to take the memory test. The effects of their brain dysfunction on memory performance introduces random error, given that brain dysfunction varies in the cause, location, severity, and effects on each person's current physiology, psychiatric status, circumstances, motivation, etc. As a result, the statistical association between age and memory test performance is likely to be reduced.

If aspects of the brain damage itself had been held constant in the Reitan and Wolfson (1995b) study that prompted questioning about use of demographic variables, perhaps the associations would have been quite significant in the brain damaged group, too (Vanderploeg, Axelrod, Sherer, et al., 1997). If younger individuals had more severe brain damage than older ones or more educated individuals had greater brain damage than less educated ones, the age–education relationships could be small or insignificant. In short, changes in these relationships do not invalidate the use of demographically based norms. Since premorbid neuropsychological test data are rare, demographically based norms aid test interpretation. Without demographically appropriate norms, the false positive rate for older or poorly educated normal individuals tends to increase (Bornstein, 1986a; R.K. Heaton, Ryan, and Grant, 2009; also see pp. 374–375). Some false negative findings can be expected (J.E. Morgan and Caccappolo-van Vliet, 2001). Yet, should a test consistently produce many false negatives or false positives with particular demographic combinations, this problem requires reevaluation of norms or demographic scoring adjustments.

Another major demographic issue in contemporary clinical neuropsychology is the use of tests across cultures and different languages and their standardization and normative base (K.B. Boone, Victor, et al., 2007; Gasquoine, 2009; K. Robertson et al., 2009). Neuropsychology had a Western European and North American origin with most standardized tests coming from these countries and languages. Eastern European, Asian, and African countries are just beginning this process and therefore additional demographic factors and normative data will likely become available.

At this time, relatively few normative samples include all of the demographic variable combinations that may be pertinent to measurement data on a particular ability or behavior. Those few samples in which all relevant demographic variables have been taken into account typically have too few subjects for dependable interpretation in the individual case. Major efforts are underway to correct this limitation for certain neuropsychological measures (Cherner et al., 2007; Gavett et al., 2009; Iverson, Williamson, et al., 2007; Pena-Casanova et al., 2009).

Possibly the most ambitious undertaking along these lines is sponsored by the National Institutes of Health. (NIH): the *NIH Toolbox* (Gershon et al., 2010). When the NIH Toolbox is complete it will provide the clinician with a well-standardized and normed brief assessment battery from which the appropriate cognitive measure can be selected to assess motor, sensory, emotional, and cognitive functioning for clinical or research purposes. The cognitive module includes assessment of the following domains: executive, episodic memory, working memory, processing speed, language, and attention. All measures will be standardized and normed in both English and Spanish on individuals 3 to 85 years of age.

Impairment criteria

Neuropsychologists generally use a criterion for identifying when performance on a particular test may represent impairment, but it is not necessarily explicitly stated and is unlikely to appear in reports. Once test data have been reliably scored and appropriate norms have been chosen to convert scores to standard scores and percentiles, the clinician needs to determine if performance on individual tests is impaired or not, and whether the pattern of performance is consistent with the patient's background and relevant neurologic, psychiatric, and/or other medical disorders. Sometimes, when poor performance does not represent an acquired impairment, simple questions about a person's abilities may elicit information that confirms lifelong difficulty in these areas of cognitive or motor ability. A poor performance may also indicate that the person was not motivated to do well or was anxious, depressed, or hostile to the test-taking endeavor rather than impaired.

Estimates of premorbid level of a patient's functioning become important in determining whether a given test performance represents impairment (see pp. 553–555, 561–563). In some cases such estimates are relatively easy to make because prior test data are available from school, military, medical, psychological, or neuropsychological records. At other times, the current test data are the primary source of the estimate. A change from this estimate, perhaps 1, 1.5, or 2 *SD*s lower than the premorbid estimate, may be used as the criterion for determining the likelihood that a particular test performance is impaired. A test score that appears to represent a 1 *SD* change from premorbid functioning may not be a statistically significant change but may indicate an impairment to some examiners and only suggest impaired performance to others. A 2 *SD* score depression is clear evidence of impairment.

Since approximately 15% of intact individuals obtain scores greater than 1 *SD* below test means, there is concern that too many individuals who are intact with respect to particular functions will be judged impaired when using −1 *SD* as an impairment criterion. When the criterion is less stringent (e.g., −1 *SD* rather than −2), more intact performance will be called impaired (i.e., *false positive*) and more "*hits*" (i.e., impaired performance correctly identified) are to be expected. On the other hand, when criteria become overly strict (e.g., > −2) the possible increase in *misses* occurs such that a truly impaired performance is judged normal (i.e., *false negative*). These errors can be costly to patients with a developing, treatable disease such as some types of brain tumors which will grow and do much mischief if not identified as soon as possible. Should this be a false alarm, the patient is no worse off in the long run but may have paid in unnecessary worry and expensive medical tests. In the case of a possible dementia, this would not be so costly an error since there is no successful treatment at the moment and the disorder will progress and have to be managed until the individual dies. However, neuropsychological conclusions must not rest on a single aberrant score. Regardless of the criterion used, it is the resulting *pattern* of change in performance that should make diagnostic sense.

Some neuropsychologists interpret as "probably impaired" any test score 1 or more *SD* lower than the mean of a normative sample that may or may not take into account appropriate demographics (e.g., Golden, Purisch, and Hammeke, 1991; R.K. Heaton, Grant, and Matthews, 1991). This latter group converted scores from the Halstead-Reitan battery plus other tests into *T*-scores based on age, education, and sex corrections. In this system a *T*-score below 40 (> −1 *SD* below the mean) is considered likely to represent impaired performance. The *pattern* of test scores is also important and must make sense in terms of the patient's history and suspected disorder or disease process (R.K. Heaton, Ryan, and Grant, 2009). In evaluating test performances, it must be kept in mind that intact individuals are likely to vary in their performance on any battery of cognitive tests and it is not unusual for them to score in the impaired range on one or two tests (Jarvis and Barth, 1994; M.J. Taylor and Heaton, 2001).

It is important to note that using a criterion for decision making that represents a deviation from the mean of the normative sample rather than change from premorbid level of functioning is likely to miss significant changes in very high functioning individuals while suggesting that low functioning individuals have acquired impairments that they do not have.

For instance, a concert pianist might begin to develop slight difficulties in hand functioning in the early stages of Parkinson's disease that were noticeable to him but not to an examiner who uses a criterion for impairment linked to the mean of the distribution of scores for males of his age, education, and sex. In that case another musician might pick up the difference by comparing recordings of an earlier performance with a current performance. Contrast this example with one of several painters who claimed to be brain-damaged after inhaling epoxy paint fumes in a poorly ventilated college locker room. On the basis of his age and education he would be expected to perform at an *average* level. Linking poor performance on many tests to toxic exposure by one psychologist seemed appropriate. However, once his grade school through high school records were obtained, it was found that he had always been functioning at a *borderline* to *impaired* level on group mental ability and achievement tests.

When such evidence of premorbid functioning is available—and often it is not—it far outweighs normative expectations. "If I had reason to believe that the person was not representative of what appears to be the appropriate normative sample, I would compare the individual with a more appropriate sample [e.g., compare an academically skilled high-school dropout to a higher educational normative sample] and be prepared to defend this decision" (R.K. Heaton, personal communication, 2003). This is how competent clinicians tend to decide in the individual case whether to use impairment criteria based on large sample norms or smaller, more demographically suitable norms.

Sensitivity/specificity and diagnostic accuracy

It has become the custom of some investigators in clinical neuropsychology to judge the "goodness" of a test or measure and its efficiency in terms of its diagnostic accuracy, i.e., the percentage of cases it correctly identifies as belonging to either a clinical population or a control group or to either of two clinical populations. This practice is predicated on questionable assumptions, one of which is that the accuracy with which a test makes diagnostic classifications is a major consideration in evaluating its clinical worth. Most tests are not used for this purpose most of the time but rather to provide a description of the individual's strengths and weaknesses, to monitor the status of a disorder or disease, or for treatment and planning. The criterion of diagnostic accuracy becomes important when evaluating screening tests for particular kinds of deficits (e.g., an aphasia screening test), single tests purporting to be sensitive to brain dysfunction, and sometimes other tests and test batteries as well.

The accuracy of diagnostic classification depends to some degree on its sensitivity and specificity (see p. 127). The percentage of cases classified accurately by any given test, however, will depend on the base rate of the condition(s) for which the test is sensitive in the population(s) used to evaluate its goodness. It will also depend on the demographics of the population, for instance, level of education (Ostrosky-Solis, Lopez-Arango, and Ardila, 2000). With judicious selection of populations, an investigator can virtually predetermine the outcome. If high diagnostic accuracy rates are desired, then the brain damaged population should consist of subjects who are known to suffer the condition(s) measured by the test(s) under consideration (e.g., patients with left hemisphere lesions suffering communication disorders tested with an aphasia screening test); members of the comparison population (e.g., normal control subjects, neurotic patients) should be chosen on the basis that they are unlikely to have the condition(s) measured by the test. Using a population in which the frequency of the condition measured by the test(s) under consideration is much lower (e.g., patients who have had only one stroke, regardless of site) will necessarily lower the sensitivity rate. However, this lower hit rate should not reflect upon the value of the test.

The extent to which sensitivity/specificity rates will differ is shown by the large differences reported in studies using the same test(s) with different kinds of clinical (and control) populations (Bornstein, 1986a; Mitrushina, Boone, et al., 2005). Moreover, it will usually be inappropriate to apply sensitivity/specificity data collected on a population with one kind of neurological disorder to patients suspected of having a different condition. Since the "sensitivity/specificity diagnostic accuracy rate" standard can be manipulated by the choice of populations studied and the discrimination rate found for one set of populations or one disorder may not apply to others, it is per se virtually meaningless as a measure of a test's effectiveness in identifying brain impaired or intact subjects except under similar conditions with similar populations. A particular test's sensitivity to a specific disorder is, of course, always of interest.

The decision-making procedure (or combination of procedures) that best accomplishes the goal of accurate diagnosis has yet to be agreed upon; and there may be none that will be best in all cases. In the end, decisions are made about individuals. Regardless of how clinicians reach their conclusions, they must always be sensitive to those elements involved in each patient's case that may be unique as well as those similar to cases seen before: qualitative and quantitative data from test performance, behavioral observation, interviews with family members and others as possible, and the history. Disagreements among clinicians are most likely to occur when the symptoms are vague and/or mild; the developmental, academic, medical, psychiatric, psychosocial, and/or occupational histories are complex or not fully available; and the pattern of test performance is not clearly associated with a specific diagnostic entity.

Screening Techniques

Different screening techniques make use of different kinds of behavioral manifestations of brain damage. Some patients suffer only a single highly specific defect or a cluster of related disabilities while, for the most part, cognitive functioning remains intact. Others sustain widespread impairment involving changes in cognitive, self-regulating, and executive functions, in attention and alertness, and in their personality. Still others display aberrations characteristic of brain dysfunction (*signs*) with more or less subtle evidence of cognitive or emotional deficits. With such a variety of signs, symptoms, and behavioral alterations, it is no more reasonable to expect accurate detection of every instance of brain disorder with one or a few instruments or lists of signs and symptoms than to expect that a handful of laboratory tests would bring to light all gastrointestinal tract diseases. Yet many clinical and social service settings need some practical means for screening when the population under consideration—such as professional boxers, alcoholics seeking treatment, persons tested as HIV positive, or elderly depressed patients, to give just a few instances—is at more than ordinary risk of a brain disorder.

The accuracy of screening tests varies in a somewhat direct relationship to the narrowness of range or specificity of the behaviors assessed by them (Sox et al., 1988). Any specific cognitive defect associated with a neurological disorder affects a relatively small proportion of the brain-impaired population as a whole, and virtually no one whose higher brain functions are intact. For instance, *perseveration* (the continuation of a response after it is no longer appropriate, as in writing three or four "e's" in a word such as "deep" or "seen" or in copying a 12-dot line without regard for the number, stopping only when the edge of the page is reached) is so strongly associated with brain damage that the examiner should suspect it on the basis of this defect alone. However, since most patients with brain disorders do not give perseverative responses, it is not a practical criterion for screening purposes. Use of a highly specific sign or symptom such as perseveration as a screening criterion for brain damage results in virtually no one without brain damage being misidentified as brain damaged (*false positive errors*), but such a narrow test will let many persons who are brain damaged slip through the screen (*false negative errors*). In contrast, defects that affect cognitive functioning generally, such as distractibility, impaired immediate memory, and concrete thinking, are not only very common symptoms of brain damage but tend to accompany a number of emotional disorders as well. As a result, a sensitive screening test that relies on a defect impairing cognitive functioning generally will identify many brain damaged patients correctly with few false negative errors, but a large number of people without brain disorders will also be included as a result of false positive errors of identification.

Limitations in predictive accuracy do not invalidate either tests for specific signs or tests that are sensitive to conditions of general dysfunction. Each kind of test can be used effectively as a screening device as long as its limitations are known and the information it elicits is interpreted accordingly. When testing is primarily for screening purposes, a combination of tests, including some that are sensitive to specific impairment, some to general impairment, and others that tend to draw out diagnostic signs, will make the best diagnostic discriminations.

Signs

The reliance on signs for identifying persons with a brain disorder has a historical basis in neuropsychology and is based on the assumption that brain disorders have some distinctive behavioral manifestations. In part this assumption reflects early concepts of brain damage as a unitary kind of dysfunction (e.g., Hebb, 1942; Shure and Halstead, 1958) and in part it arises from observations of response characteristics that do distinguish the test performances of many patients with brain disease.

Most pathognomonic signs in neuropsychological assessment are specific aberrant test responses or modes of response. These signs may be either *positive*, indicating the presence of abnormal function, or *negative* in that the function is lost or significantly diminished. Some signs are isolated response deviations that, in themselves, may indicate the presence of an organic defect. Rotation in copying a block design or a geometric figure has been considered a sign of brain damage. Specific test failures or test score discrepancies have also been treated as signs of brain dysfunction, as for instance, marked difficulty on a serial subtraction task (Ruesch and Moore, 1943) or a wide spread between the number of digits recalled in the order given and the number recalled in reversed order (Wechsler, 1958). The manner in which the patient responds to the task may also be considered a sign indicating brain damage. M. Williams (1979) associated three response characteristics with brain damage: "stereotyping and perseveration"; "concreteness of behavior," defined by her as "response to all stimuli as if they existed only in the setting in which they are presented"; and "catastrophic reactions" of perplexity, acute anxiety, and despair when the patient is unable to perform the presented task.

Another common sign approach relies on not one but on the sum of different signs, i.e., the total number of different kinds of specific test response aberrations or differentiating test item selections made by the patient. This method is used in some mental status examinations to determine the likelihood of impairment (see p. 127). In practice, a number of behavior changes can serve as signs of brain dysfunction (see Table 6.2). None of them alone is pathognomonic of a specific brain disorder. When a patient presents with more than a few of these changes, the likelihood of a brain disorder runs high.

Cutting scores

The score that separates the "normal" or "not impaired" from the "abnormal" or "impaired" ends of a continuum of test scores is called a *cutting score*, which marks the *cut-off* point (Dwyer, 1996). The use of cutting scores is akin to the sign approach, for their purpose is to separate patients in terms of the presence or absence of the condition under study. A statistically derived cutting score is the score that differentiates brain impaired patients from others with the fewest instances of error on either side. A cutting score may also be derived by simple inspection, in which case it is usually the score just below the poorest score attained by any member of the "normal" comparison group or below the lowest score made by 95% of the "normal" comparison group (see Benton, Sivan, Hamsher, et al., 1994, for examples).

Cutting scores are a prominent feature of most screening tests. However, many of the cutting scores used for neuropsychological diagnosis may be less efficient than the claims made for them (Meehl and Rosen, 1967). This is most likely to be the case when the determination of a cutting score does not take into account the base rate at which the predicted condition occurs in the sample from which the cutting score was developed (Urbina, 2004; W.G. Willis, 1984).

Other problems also tend to vitiate the effectiveness of cutting scores. The criterion groups are often not large enough for optimal cutting scores to be determined (Soper, Cicchetti, et al., 1988). Further, cutting scores developed on one kind of population may not apply to another. R.L. Adams, Boake, and Crain (1982)

TABLE 6.2 Behavior Changes that Are Possible Indicators of a Pathological Brain Process

*Functional Class**	*Symptoms and Signs*	*Functional Class**	*Symptoms and Signs*
Speech and language	Dysarthria Dysfluency Marked change in amount of speech output Paraphasias Word finding problems	Visuospatial abilities	Diminished or distorted ability for manual skills (e.g., mechanical repairs, sewing) Spatial disorientation Impaired spatial judgment Right–left disorientation
Academic skills	Alterations in reading, writing, calculating, and number abilities; e.g., poor reading comprehension, frequent letter or number reversals in writing	Emotional	Diminished emotional control with temper outbursts, antisocial behavior Diminished empathy or interest in interpersonal relationships without depression
Thinking	Perseveration of speech or action components Simplified or confused mental tracking, reasoning, concept formation		Affective changes without known precipitating factors (e.g., lability, flattening, inappropriateness)
Motor	Lateralized weakness or clumsiness Problems with fine motor coordination		Personality changes without known precipitating factors Increased irritability without known precipitating factors
Perception	Diplopia or visual field alterations Inattention (usually left-sided, may be perceptual and/or in productions) Somatosensory alterations (particularly lateralized or confined to one limb)	Comportment†	Altered appetites and appetitive activities (eating, drinking, play, sex) Altered grooming habits (overly fastidious, careless) Hyper- or hypoactivity Social inappropriateness

*Many emotionally disturbed persons complain of memory deficits that typically reflect their self-preoccupations, distractibility, or anxiety rather than a dysfunctional brain. Thus memory complaints in themselves are not good indicators of neuropathology.

†These changes are most likely to have neuropsychological relevance in the absence of depression, but they can be mistaken for depression.

Adapted from Howieson and Lezak, 2002; © 2002, American Psychiatric Association Press.

pointed out the importance of adjusting cutting scores for "age, education, premorbid intelligence, and race–ethnicity" by demonstrating that the likelihood of false positive predictions of brain damage tends to increase for nonwhites and directly with age, and inversely with education and intelligence test scores. Bornstein (1986a) and Bornstein, Paniak, and O'Brien (1987) demonstrated how cutting scores, mostly developed on a small and relatively young normative sample, classified as "impaired" from 57.6% to 100% of normal control subjects in the 60–90 age range.

When the recommended cutting scores are used, these tests generally do identify impaired patients better than chance alone. They all also misdiagnose both intact persons (false positive cases) and persons with known brain impairment (false negative cases) to varying degrees. The nature of the errors of diagnosis depends on where the cut is set: if it is set to minimize misidentification of intact persons, then a greater number of brain impaired patients will be called "normal" by the screening. Conversely, if the test-maker's goal is to identify as many patients with brain damage as possible, more intact persons will be included in the brain damaged group. Only rarely does the cutting score provide a distinct separation between two populations, and then only for tests that are so simple that all ordinary intact adults would not fail. For example, the Token Test, which consists of simple verbal instructions involving basic concepts of size, color, and location, is unlikely to misidentify verbally intact persons as impaired.

Single tests for identifying brain disorders

The use of single tests for identifying brain damaged patients—a popular enterprise several decades ago—was based on the assumption that brain damage, like measles perhaps, can be treated as a single entity. Considering the heterogeneity of brain disorders, it is not surprising that single tests have high misclassification rates (G. Goldstein and Shelly, 1973; Spreen and Benton, 1965). Most single tests, including many that are not well standardized, can be rich sources of information about the functions, attitudes, and habits they elicit. Yet to look to any single test for decisive information about overall cognitive behavior is not merely foolish but can be dangerous as well, since the absence of positive findings does not rule out the presence of a pathological condition.

Usefulness of screening techniques

In the 1940s and 1950s, in the context of the simple "organic" versus "functional" distinction, brain damage was still thought by many to have some general manifestation that could be demonstrated by psychological tests, screening techniques were popular, particularly for identifying the brain impaired patients in a psychiatric population. As a result of better understanding of the multifaceted nature of brain pathology and of the accelerating development and refinement of other kinds of neurodiagnostic techniques, the usefulness of neuropsychological screening has become much more limited. Screening is unnecessary or inappropriate in most cases referred for neuropsychological evaluation: either the presence of neuropathology is obvious or otherwise documented, or diagnosis requires more than simple screening. Furthermore, the extent to which screening techniques produce false positives and false negatives compromises their reliability for making decisions about individual patients.

However, screening may still be useful with populations in which neurological disorders are more frequent than in the general population (e.g., community dwelling elderly people [Cahn, Salmon, et al., 1995]). The most obvious clinical situations in which neuropsychological screening may be called for are examinations of patients entering a psychiatric inpatient service or at-risk groups such as the elderly or alcoholics/substance abusers when they seek medical care. Screening tests are increasingly used in the U.S. and Canada to identify and monitor concussions in sports participants, especially soccer and football (Covassin et al., 2009; Van Kampen et al., 2007). Dichotomizing screening techniques are also useful in research for evaluating tests or treatments, or for comparing specific populations with respect to the presence or absence of impaired functions.

Once a patient has been identified by screening techniques as possibly having a brain disorder, the problem arises of what to do next, for simple screening at best operates only as an early warning system. These patients still need careful neurological and neuropsychological study to determine whether a brain disorder is present and, if so, to help develop treatment and planning for their care as needed.

Evaluating screening techniques

In neuropsychology as in medicine, limitations in predictive accuracy do not invalidate either tests for specific signs or disabilities or tests that are sensitive to conditions of general dysfunction. We have not thrown away thermometers because most sick people have normal temperatures, nor do we reject the electroencephalogram (EEG) just because many patients with brain disorders test normal by that method. Thus, in neuropsychology, each kind of test can be used effectively as a screening device as long as its limitations are known and the information it elicits is interpreted accordingly. For screening purposes, a combination of tests, including some that are sensitive to specific impairment, some to general impairment, and others that tend to draw out diagnostic signs, will make the best diagnostic discriminations.

When evaluating tests for screening, it is important to realize that, although neuropsychological testing has proven effective in identifying the presence of brain disorders, it cannot guarantee its absence, i.e., "rule out" brain dysfunction. Not only may cerebral disease occur without behavioral manifestations, but the examiner may also neglect to look for those neuropsychological abnormalities that are present. Inability to prove the negative case in neuropsychological assessment is shared with every other diagnostic tool in medicine and the behavioral sciences. When a neuropsychological examination produces no positive findings, the only tenable conclusion is that the person in question performed *within normal limits* on the tests taken at that time. While the performance may be adequate for the test conditions at that time of assessment, the neuropsychologist cannot give a "clean bill of health."

Pattern Analysis

Intraindividual variability

Discrepancy, or variability, in the pattern of successes and failures in a test performance is called *scatter*. Variability within a test is *intratest scatter*; variability between the scores of a set of tests is *intertest scatter* (Wechsler, 1958).

Intratest scatter. Scatter within a test is said to be present when there are marked deviations from the normal pass–fail pattern. On tests in which the items are presented in order of difficulty, it is usual for the subject to pass almost all items up to the most difficult passed item, with perhaps one or two failures on items close to the last passed item. Rarely do cognitively intact persons fail very simple items or fail many items of middling difficulty and pass several difficult ones. On tests in which all items are of similar difficulty level, most subjects tend to do all of them correctly, with perhaps one or two errors of carelessness, or they tend to flounder hopelessly with maybe one or two lucky "hits."

Variations from these two common patterns deserve the examiner's attention.

Certain brain disorders as well as some emotional disturbances may manifest themselves in intratest scatter patterns. Hovey and Kooi (1955) demonstrated that, when taking mental tests, patients with epilepsy who exhibit paroxysmal brain wave patterns (sudden bursts of activity) were significantly more likely to be randomly nonresponsive or forgetful than were psychiatric, brain damaged, or other epileptic patients. Some patients who have sustained severe head injuries respond to questions that draw on prior knowledge as if they had randomly lost chunks of stored information. For example, moderately to severely injured patients as a group displayed more intratest scatter than a comparable control group, although scatter alone did not reliably differentiate brain injured from control subjects on an individual basis (Mittenberg, Hammeke, and Rao, 1989). Variability, both intratest and over time, characterized responses of patients with frontal lobe dementia (Murtha et al., 2002). E. Strauss, MacDonald, and their colleagues (2002) found a relationship between inconsistency in physical performance and fluctuations on cognitive tests.

If scatter is present within test performances, the challenge for the examiner is to assess whether the observed scatter in a given patient is beyond what would occur for the relevant reference group. As few intratest scatter studies for specific diagnostic groups have been undertaken, the examiner can only rely on experience, personal judgment, and what is known about scatter patterns for particular tests (e.g., Crawford, Allan, McGeorge, and Kelly, 1997). Intratest scatter may also be influenced by cultural and language factors (Rivera Mindt et al., 2008).

Intertest scatter. Probably the most common approach to the psychological evaluation of brain disorders is through comparison of the test score levels obtained by the subject—in other words, through analysis of the intertest score scatter. By this means, the examiner attempts to relate variations between test scores to probable neurological events—or behavioral descriptions in those many cases in which a diagnosis is known. This technique clarifies a seeming confusion of signs and symptoms of behavioral disorder by giving the examiner a frame of reference for organizing and evaluating the data.

Making sense of intraindividual variability

A significant discrepancy between any two or more scores is the basic element of test score analysis (Silverstein, 1982). Any single discrepant score or response error can usually be disregarded as a chance deviation. A number of errors or test score deviations may form a pattern. Marked quantitative discrepancies in a person's performance—within responses to a test, between scores on different tests, and/or with respect to an expected level of performance—suggest that some abnormal condition is interfering with that person's overall ability to perform at their characteristic level of cognitive functioning. Brain dysfunction is suspected when a neurological condition best accounts for the patient's behavioral abnormalities.

In order to interpret the pattern of performance in a multivariate examination, the clinician must fully understand the nature of the tests administered, what the various tests have in common and how they differ in terms of input and output modalities, and what cognitive processes are required for successful completion. Appropriate interpretation of the data further requires a thoughtful integration of historical, demographic, and psychosocial data with the examination information.

A 32-year-old doctoral candidate in the biological sciences sustained a head injury with momentary loss of consciousness just weeks before she was to take her qualifying examinations. She was given a small set of neuropsychological tests two months after the accident to determine the nature of her memory complaints and how she might compensate for them. Besides a few tests of verbal, visuospatial, and conceptual functions, the relatively brief examination consisted mainly of tests of attention and memory as they are often most relevant to mild post traumatic conditions.

The patient had little problem with attentional or reasoning tests, whether verbal or visual, although some tendency to concrete thinking was observed. Both story recall and sentence repetition were excellent; she recalled all of nine symbol–digit pairs immediately after 3 min spent assigning digits to an associated symbol, and seven of the pairs a half hour later (Symbol Digit Modalities Test); and she recognized an almost normal number of words (12) from a list of 15 she had attempted to learn in five trials (Auditory-Verbal Learning Test). However, this very bright woman, whose speaking skills were consistent with her high academic achievement, could not retrieve several words without phonetic cueing (Boston Naming Test); and she gave impaired performances when attempting to learn a series of nine digits (Serial Digit Learning), on immediate and delayed recall of the 15-word list, and on visual recall on which she reproduced the configuration of the geometric design she had copied but not the details (Complex Figure Test). Thus she clearly demonstrated the ability for verbal learning at a normal level, and her visual recall indicated that she could at least learn the "big picture." Her successes occurred on all meaningful material and when she had cues; when meaning or cues—hooks she could use to aid retrieval—were absent, she performed at *defective* levels. Analysis of her successes and failures showed a consistent pattern implicating retrieval problems that compromised her otherwise adequate learning ability. This analysis allowed the examiner to reassure her regarding her learning capacity and to recommend techniques for prodding her sluggish retrieval processes.

Pattern analysis procedures

The question of neuroanatomical or neurophysiological likelihood underlies all analyses of test patterns undertaken for differential diagnosis. As in every other diagnostic effort, the most likely explanation for a behavioral disorder is the one that requires the least number of unlikely events to account for it. Once test data have been reliably scored and appropriate norms have been chosen to convert scores to standard scores or percentiles, the clinician determines whether the pattern of performance is typical of individuals with a particular diagnosis. The many differences in cognitive performance between diagnostic groups and between individuals within these groups can be best appreciated and put to clinical use when the evaluation is based on test score patterns and item analyses taken from tests of many different functions. If it fits a diagnostic pattern, the clinician then must consider what would be the behavioral ramifications of this individual's unique pattern, as even within a diagnostic category, few persons will have an identical presentation.

Now that neuroimaging and laboratory technology often provide the definitive neurological diagnosis, how a brain disorder or disease might play out in real life may be the most important issue in the neuropsychological examination. In planning the examination, the examiner will have in mind questions about the patient's real life functioning, such as potential for training or rehabilitation, return to work or requiring assisted living, quality of life and capacity for interpersonal relationships. These examinations require a fairly broad review of functions. Damage to cortical tissue in an area serving a specific function not only changes or abolishes the expression of that function but changes the character of all activities and functions in which the impaired function was originally involved, depending upon how much the function itself has changed and the extent to which it entered into the activity (see pp. 347–348). A minor or well-circumscribed cognitive deficit may show up on only one or a very few depressed test scores or may not become evident at all if the test battery samples a narrow range of behaviors.

Most of the functions that a neuropsychologist examines are complex. In analyzing test score patterns, the examiner looks for both commonality of dysfunction and evidence of impairment on tests involving functions or skills that are associated neuroanatomically, in their cognitive expression, and with well-described disease entities and neuropathological conditions. First, the examiner estimates a general level of premorbid functioning from the patient's history, qualitative aspects of performance, and test scores, using the examination or historical indicators that reasonably allow the highest estimate (see Chapter 4). This aids the examiner in identifying impaired test performances. The examiner then follows the procedures for dissociation of dysfunction by comparing test scores with one another to determine whether any factors are consistently associated with high or low scores, and if so, which ones (see p. 131). The functions which contribute consistently to impaired test performances are the possible behavioral correlates of brain dysfunction, and/or represent those areas of function in which the patient can be expected to have the most difficulty. When the pattern of impaired functions or lowered test scores does not appear to be consistently associated with a known or neurologically meaningful pattern of cognitive dysfunction, discrepant scores may well be attributable to psychogenic, developmental, or chance deviations (L.M. Binder, Iverson, and Brooks, 2009).

By and large, the use of pattern analysis has been confined to tests in the Wechsler batteries because of their obvious statistical comparability. However, by converting different kinds of test scores into comparable score units, the examiner can compare data from many different tests in a systematic manner, permitting the analysis of patterns formed by the scores of tests from many sources. For example, R.K. Heaton, Grant, and Matthews (1991) converted scores from a large number of tests to a single standard score system.

INTEGRATED INTERPRETATION

Pattern analysis is insufficient to deal with the numerous exceptions to characteristic patterns, with the many rare or idiosyncratically manifested neurological conditions, and with the effects on test performance of the complex interaction between patients' cognitive status, their emotional and social adjustment, and their appreciation of their altered functioning. For the examination to supply answers to many of the diagnostic questions and most of the treatment and planning questions requires integration of all the data—from tests, observations made in the course of the examination, and the history of the problem.

Some conditions do not lend themselves to pattern analysis beyond the use of large and consistent test score discrepancies to implicate brain damage. For example, malignant tumors are unlikely to follow a regular pattern of growth and spread (e.g., see Plates x and x). In order to determine which functions are involved and the extent of their involvement, it is usually necessary to evaluate the qualitative aspects of the patient's performance very carefully for evidence of complex or subtle aberrations that betray damage in some hitherto unsuspected area of the brain. Such painstaking scrutiny may not be as necessary when dealing with a patient whose disease generally follows a well-known and regular course.

Test scores alone do not provide much information about the emotional impact of brain damage on the individual patient's cognitive functioning or how fatigue may alter performance. However, behavior during the examination is likely to reveal a great deal about reactions to the disabilities and how these reactions in turn affect performance efficiency. Emotional reactions of brain damaged patients can affect their cognitive functioning adversely. The most prevalent and most profoundly handicapping of these are anxiety and depression. Euphoria and carelessness, while much less distressing to the patient, can also seriously interfere with expression of a patient's abilities.

Many brain impaired patients have other characteristic problems that generally do not depress test scores but must be taken into account in rehabilitation planning. These are motivational and control (executive function) problems that show up in a reduced ability to organize, to react spontaneously, to initiate goal-directed behavior, or to carry out a course of action independently. They are rarely reflected in test scores since almost all tests are well structured and administered by an examiner who plans, initiates, and conducts the examination (see Chapter 16 for tests that elicit these problems). Yet, no matter how well patients do on tests, if they cannot develop or carry out their own course of action, they are incompetent for all practical purposes. Such problems become apparent during careful examination, but they usually must be reported descriptively unless the examiner sets up a test situation that can provide a systematic and scorable means of assessing the patient's capacity for self-direction and planning.